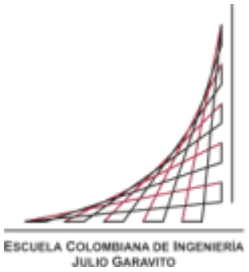
Minería de texto histórica - colaboración al proyecto ‘Revealing Cooperation and Conflict Project’

Diana María del Pilar Socha Díaz
Juan Sebastián Martínez Serna
Cristian Camilo Medina Mosquera



Proyecto de grado
Facultad Ingeniería de Sistemas
Escuela Colombiana de Ingeniería Julio Garavito

Director: PhD. Ignacio Pérez Vélez
Co-Director: PhD. Roger Louis Martínez-Dávila

31 de enero de 2017

Índice general

1 Revealing Cooperation and Conflict Project **6**

1.1. Descripción 6

1.2. Contexto Histórico 7

1.2.1. Las familias más importantes de Plasencia 9

1.2.2. Actas capitulares de la Catedral de Plasencia 10

1.3. El MOOC 11

1.3.1. Objetivo del MOOC 13

1.3.2. Datos generados por el MOOC 14

2 Justificación **16**

2.1. Aporte al proyecto Revealing Cooperation and Conflict Project . . . 18

2.2. Aporte desde la Ingeniería 19

3 Objetivos **22**

3.1. Objetivo general 22

3.2. Objetivos específicos 22

4 Minería de texto **24**

4.1. ¿Qué es? 24

4.2. Ciclo de trabajo 25

4.2.1. Corpus 26

4.2.2. Stop - Words 27

4.2.3. Lematización 28

5 Antecedentes de minería de texto histórica **30**

5.1. Expressing complex associations in medieval historical documents: the Henry III Fine Rolls Project 30

5.2. Visualization of relationships among historical persons from Japanese historical documents 36
5.3. Information Access to Historical Documents from the Early New High German Period 41
5.4. Information Retrieval from Historical Corpora 49
5.5. Topic Modeling on Historical Newspapers 52

6 Herramientas para minería de texto **57**
6.1. KNIME Analytics Platform 57
6.2. RapidMiner 58
6.3. R-Programming y R-Studio 59
6.4. Weka 60

7 Proceso del trabajo **63**
7.1. Metodología de trabajo 63
7.2. El contexto de los datos 66
7.2.1. Estructura de los datos 66
7.2.2. Limpieza de datos 69
7.2.3. Flujos de trabajo 72
7.3. Correlaciones 72

8 Conclusiones **75**

A Diccionarios **77**
A.1. Diccionario de Nombres 77
A.2. Diccionario de Lugares 92
A.3. Diccionario de Roles 100
A.4. Diccionario de Roles - general de la época 104

B Código Python de limpieza de las transcripciones **107**

C Código Python para las correlaciones **121**

Siglas **124**

Glosario **126**

Bibliografía **129**

Índice de cuadros

4.1. Relación entre los temas y las diferentes áreas de la minería de texto . . 26
4.2. Ejemplo stop-words 27
4.3. Ejemplo 1 lematización 28
4.4. Ejemplo 2 lematización 29

Índice de figuras

1.1. Imagen del proyecto RCCP 6
1.2. Confederación de la catedral 10
1.3. Fragmento Actas Capitulares de la Catedral de Plasencia 11

4.1. Las siete áreas practicas del análisis de textos 25
4.2. Diagrama del ciclo de trabajo 27

5.1. Ejemplo de asociaciones a partir de documento estructurado 32
5.2. Proceso para construir la Ontología 33
5.3. Ejemplo de anidaciones entre conceptos, TEI XML 35
5.4. Gráfica del relaciones entre personas y factores históricos 39
5.5. Tabla del relaciones entre personas y factores históricos 40
5.6. Gráfica del resultado del agrupamiento utilizando información por locación 41

5.7. Tabla del resultado del agrupamiento utilizando información por locación 41
5.8. Ejemplo Variantes en grafemas vocálicos 45
5.9. Resultados del sistema al encontrar relaciones 48
5.10. Variantes y sinónimos encontrados en documentos históricos y en documentos modernos 50
5.11. Resultados de la evaluación de desempeño 52

6.1. Pipelining modular de datos 57
6.2. KNIME Analytics Platform 58
6.3. RapidMiner Studio 60
6.4. RStudio 61
6.5. Weka Knowledge Flow 62

7.1. Diagrama CRIPS-DM 64
7.2. Carpetas con todas las actividades del curso 67
7.3. Carpetas de los estudiantes de cada actividad 67
7.4. La actividad del estudiante con los calificadores de la misma 68
7.5. Transcripciones digitalizadas “sucias” 68
7.6. Documento obtenido del proceso de la limpieza 72
7.7. Flujo en KNIME de Tag Cloud para limpieza y extracción 73
7.8. Tag Cloud para limpieza y extracción 74

Capítulo 1

Revealing Cooperation and Conflict Project

Figura 1.1: Imagen del proyecto RCCP

1.1. Descripción

Revealing Cooperation and Conflict Project (RCCP) es un proyecto de investigación histórica, el cual va desde **principios del siglo XIV hasta finales del siglo XVI**. El objetivo principal es *"reconstruir tanto los procesos de cooperación como de disputas que surgieron durante un período que alternaba tanto la integración intercultural como la violencia en España y en Europa"*[1], conociendo las relaciones de coexistencia entre judíos, cristianos y musulmanes; debido a que durante está época, en España se vivían fuertes episodios de violencia y además surgía una integración intercultural que alteraba aún más la situación, y fue por esto mismo que jugó un papel clave en la violencia que vivió Europa, por tal motivo, los estudios

para la investigación del proyecto RCCP se centrarán en la ciudad de Plasecencia, España, más específicamente en la Catedral de Plasencia, porque esta era la máxima autoridad en la ciudad y se encargaba del manejo de todos los documentos oficiales, *"sabemos que la Catedral de Plasencia transfería dinero regularmente a los banqueros del norte de Europa a principios del siglo XVI, lo que sugiere que la región estaba fuertemente ligada a los mercados y a los asuntos políticos europeos"* [1]

Vale aclarar que *"el proyecto se centra en asuntos interreligiosos porque la evidencia archivística apunta a que las relaciones entre judíos, católicos y musulmanes eran mucho más fluidas, tanto positiva como negativamente, de lo que los estudiosos contemporáneos sugieren y de lo que percibe la opinión pública."* [1]. Para el análisis y descubrimiento de aquellos acontecimientos relevantes para la investigación, que pueden dar luces acerca de lo que ocurría entonces y puntos en los cuáles se aleja de lo planteado en dichas otras investigaciones contemporáneas, se utiliza como principal fuente de información documentos denominados Actas Capitulares de la Catedral de Plasencia [1].

El proyecto RCCP es dirigido por el **Dr. Roger Louis Martínez-Dávila** , y además es un proyecto sin fines lucrativos, abierto para que todos participen al avance del mismo. *"El proyecto reúne expertos de geovisualización, historiadores, geógrafos, lingüistas e* ***informáticos*** *de EE.UU., Suiza y España, así como académicos y ciudadanos de* ***todo el mundo****"* [2]. Además se considera que presenta una innovación en la forma de realizar proyectos de alta complejidad, pues este proyecto *"se aprovecha también el conocimiento de ciudadanos expertos al permitir la colaboración abierta en la transcripción y la indexación de los documentos históricos. Se implementa un formato de base de datos más flexible para captar mejor las relaciones no lineales entre los protagonistas históricos y las distintas circunstancias que rodeaban sus entornos."* [1]

1.2. Contexto Histórico

En la época medieval la comunidad Judía se hallaba muy repartida, la ciudad donde más se encontraban judíos era en Toledo, en la comunidad de Andalucía se destacaba las ciudades de Sevilla y Córdoba y en Extremadura Caceres y Plasencia. Las relaciones entre Cristianos y Judíos entre los siglos XI a XIII fueron mayormente pacíficas, pero durante el siguiente siglo las dificultades económicas contribuyeron a hacer de los Judíos un chivo expiatorio de todos los males, a eso se suma la propagan-

da demagógica de Enrique de Trastámara quien atacando a los judíos, quiso ganarse el favor de la población castellana en su guerra contra Pedro I. En consecuencia, durante el siglo XIV se produjeron disturbios antijudíos en ciudades como Barcelona y Valencia, pero el punto mas critico de ese proceso, fueron los violentos ataques en la ciudad de Sevilla, que se fueron propagando por varias ciudades como Córdoba y Andújar, la onda de violencia se extendió rápidamente por el norte de España produciendo asesinatos en Madrid, Toledo, Segovia, Sepúlveda y en la corona de Aragón sufrieron matanzas de judíos en Barcelona, Valencia y Palma. El clima de persecución hizo que muchos judíos abandonaron la península, dirigiéndose al norte de África, cuando en 1492 se confirmó su expulsión definitiva [3].

La expulsión de los Judíos fue ordenada por los Reyes Católicos mediante el Edicto de Granada, la finalidad de éste, era impedir que siguieran influyendo a los cristianos nuevos para que éstos judaizaran. La decisión de expulsar a los judíos está relacionada con la instauración de la Inquisición catorce años antes en la Corona de Castilla y nueve en la Corona de Aragón, porque precisamente fue creada para perseguir a los judeoconversos que seguían practicando su antigua fe[4]. Durante el siglo XV en Plasencia, España, se creó una relación de coexistencia entre las diferentes religiones que había. Judíos, musulmanes y cristianos lograron convivir en comunidades lo que les permitió coexistir mientras en el resto del país la situación era marcada por la violencia. Plasencia tiene una participación importante en la historia de Europa, ya que existía una estrecha relación con los banqueros del norte de Europa, lo que sugiere que estaba estrechamente relacionada a los mercados y los temas políticos en Europa. Allí, se dieron relaciones inusuales que se iniciaron en la frontera de Plasencia, como por ejemplo, se destaca el hecho de que la cooperación y el conflicto hacían parte del día a día; la comunidad a pesar de estar formada por personas de múltiples religiones como judíos o musulmanes o en algunos casos familias convertidas a otras religiones por intereses comunes, formaron alianzas religiosas, políticas y económicas lo cual no sucedía en otros lugares de España. Otro comportamiento poco común era que debido a la gran variedad de recursos con los que contaba Plasencia y los pueblos más cercanos, se buscaba tener un control sobre los recursos como el arrendamiento de tierras y el intercambio de ganado, poniendo en contradicción la situación que se vivía en el resto de España, entre las diferentes religiones.[5]

1.2.1. Las familias más importantes de Plasencia

De acuerdo con los registros de impuestos archivísticos de la Diócesis de Plasencia, en el año 1400 sólo había 119 hombres adultos y sus familias, 40 cristianos (34 %), 50 judíos (42 %) y 29 musulmanes (24 %). Los historiadores especulan que la población total de la ciudad era de aproximadamente de 800 a 1000 habitantes, aunque la población de la ciudad no llegó a 1,000 personas hasta la década de 1570. Así, los judíos y los musulmanes eran un componente clave de la base de la población a lo largo del siglo XV y fueron increíblemente importantes para la economía local. [4]

El Dr. Martínez-Dávila considera de vital importancia el estudiar de cerca no solo relaciones económicas y políticas que se presentaron en la época sino también las relaciones familiares y ciertas familias que ha descubierto que tenían suficiente poder sobre la sociedad de aquel entonces en Plasencia. *"Estamos identificando situaciones en las que las familias rabínicas judías, como la de Loya, trabajaron estrechamente con los conversos (convertidos del judaísmo al cristianismo) y antiguas familias de caballeros cristianos, como la de Santa María y* ***Carvajal****, para contrarrestar y equilibrar el poder de los señores, como los Estúñiga, Condes de Béjar y Plasencia. De esta manera, estamos en presencia de alianzas de "extraños compañeros de cama"* [5]. Como ya se dieron luces anteriormente, durante éste periodo existieron tres familias que resaltan debido al poder que éstas ejercían sobre Plasencia, dichas familias, además de lo ya mencionado tenían roles "profesionales" diferentes y religiones, pero mantenían contacto cercano con la Catedral de Plasencia; éstas familias fueron:

- **La Familia Carvajal** eran una familia cuyo rol de "profesional" principal era de caballeros nobles, sus integrantes eran católicos.
- **La Familia Santa María** eran antiguos judíos Ha-levis y arrendaban propiedades de la iglesia a otras familias judías
- **La Familia Estúñiga** eran una familia perteneciente a la elite de la cuidad, eran Condes de Plasencia y Béjar

La familia Carvajal considerada principalmente una familia de caballeros nobles, a finales de 1300 y comienzos de 1500, empezaron a tener una importante participación en los puestos de mayor actividad económica y política de Plasencia junto a la familia Santa María, quienes se convirtieron al cristianismo y desarrollaron una especial conexión con la familia Carvajal. La familia Santa María durante su paso

por Plasencia adquieren varios apellidos como Gutierrez de la Calleja y Fernandez de Cabrero lo que indica que tuvieron que convertirse para continuar con el poder que adquirieron en la catedral, algunos cargos de mayor importancia fueron archidiácono de Plasencia, Bejar, Trujillo, Medellín y Coria, tesorero, cantor, vicario general, obispo, caballero, burócrata real, líder eclesiástico, concejal de la región y concejal de la ciudad. En la siguiente imagen se puede ver que hasta 1420 aparece Gonzalo Gutierrez de la Calleja como tesorero de la catedral, luego el puesto es cedido en 1436 a Gonzalo Garcia de Carvajal y luego a varios integrantes de la familia Santa Maria, lo que sugiere que mantenían una relación cercana las dos familias y lo que buscaban era construir una confederación duradera en la catedral.

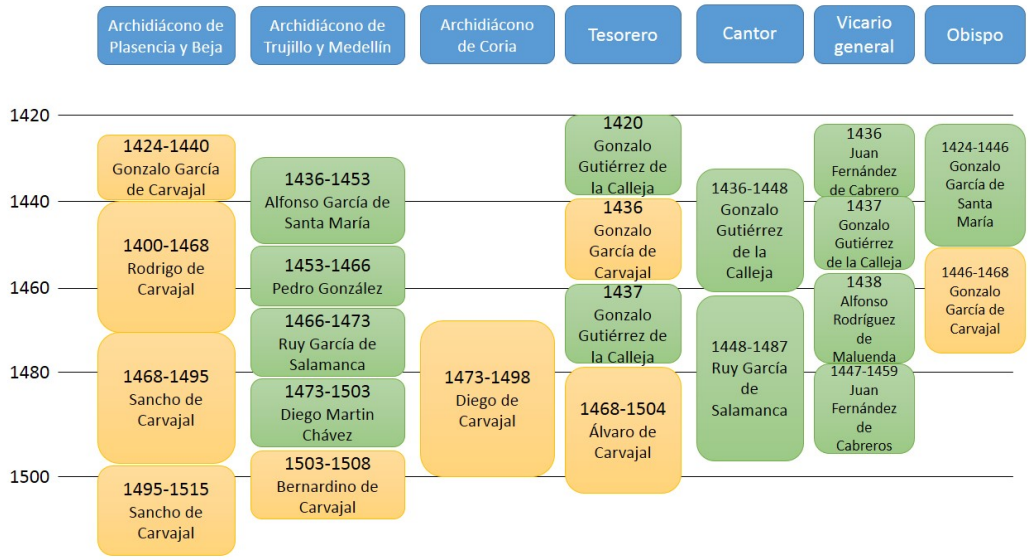


Figura 1.2: Confederación de la catedral

1.2.2. Actas capitulares de la Catedral de Plasencia

Las Actas Capitulares de la Catedral de Plasencia *"son los documentos más representativos de la gestión y administración de los concejos medievales castellanos"*[6] que son pertenecientes a los siglos desde el **siglo XIV hasta el siglo XVI** [7], en estos documentos se recogían todas las decisiones y actuaciones de quienes se encontraban a cargo de la política (principalmente concejiles), en estas se trataban mayormente temas que contienen información sobre la economía y gestión de la ciudad.

La periodicidad que presentaban las Actas Capitulares de la Catedral de Plasencia era semanal [6], aunque en caso de ser necesario de escribir una antes debido a algún evento extraordinario se hacia, también se llegaban a escribir varias actas en un mismo día. Adicionalmente estas presentaban *"una estructura uniforme con un encabezamiento que contiene, entre otros componentes, la data crónica y a veces la data tópica del lugar concreto donde se celebraba la sesión del concejo productora del acta. Seguidamente, se desplegaba el cuerpo del acta"* [6]

Estos documentos tienen el aspecto más crítico del proyecto dirigido por el Dr. Martínez-Dávila puesto que *"son una fuente inestimable para la reconstrucción de los diversos aspectos de la vida de la ciudad de Plasencia"*[6]. En la imagen de la figura 1.3 se evidencia que contiene un pequeño fragmento de un documento de éste tipo, pertenece a la **Unidad documental simple /004 - Acta capitular de 30 de mayo de 1522**, ha sido tomada del archivo municipal de Plasencia consignado digitalmente en la página oficial `http://archivo.plasencia.es/index.php`

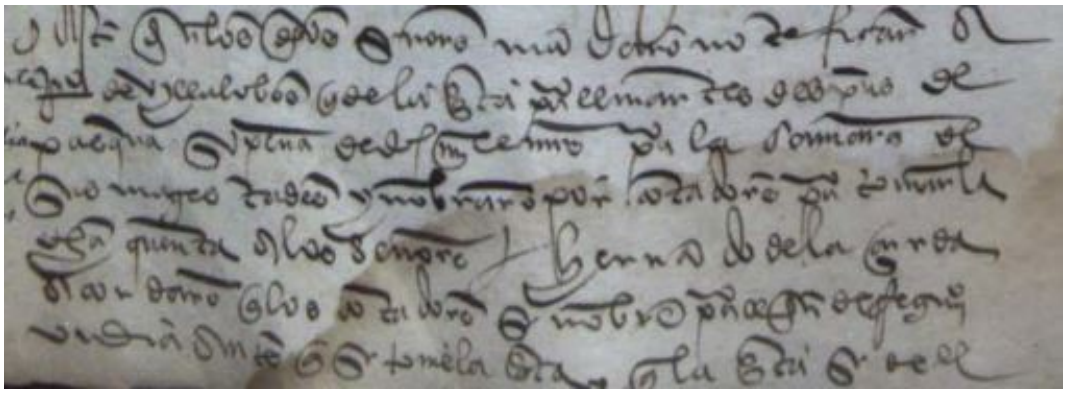


Figura 1.3: Fragmento Actas Capitulares de la Catedral de Plasencia

1.3. El MOOC

Deciphering Secrets: Cursos Abiertos Masivos en Línea, es un proyecto que involucra una serie de cursos tipo MOOC, que quiere decir que son abiertos y masivos, presentados en línea. Todos aquellos cursos presentados dentro del proyecto son dirigidos por el Dr. Roger Louis Martínez-Dávila con el apoyo tanto de la Universidad de Carlos III de Madrid como de University of Colorado-Colorado Springs. Todos los MOOCS se centran en entender las relaciones y coexistencia de personas en una sociedad marcada por interrelaciones religiosas delicadas, durante la Edad

Media en España. Todos los MOOCS que hacen parte del proyecto *Deciphering Secrets* hacen parte de una iniciativa global de educación, investigación e ingeniería denominada *"Global Citizen Scholars (GCS)"* que tiene como objetivo recuperar la curiosidad cultural sobre la humanidad, fomentar el uso de la tecnología para incrementar la participación y contribución de los estudiantes colaboradores y por último contribuir en la investigación académica que tiene que ver con la cooperación y el conflicto judío, cristiano y musulmán medieval generando así una nueva forma de investigación para el siglo XXI.

Deciphering Secrets comienza en el año 2014, como un curso en línea, masivo y abierto (MOOC), impartido por el Dr. Martínez-Dávila en la plataforma coursera.org (denominado ***Deciphering Secrets: Unlocking the Manuscripts of Medieval Spain***). Tanto en su momento como en el tiempo que estuvo disponible al público, el curso fue todo un éxito atrayendo a más de 10,000 estudiantes de 140 naciones y más de 2,500 estudiantes lo completaron con éxito. A destacar de éste curso que de aquellos estudiantes, más de 1,000 de ellos deciden continuar como lo que se denomina **"ciudadanos académicos"**, colaborando al desarrollo del proyecto Revealing Cooperation and Conflict Project (RCCP). [8]

Para este MOOC los estudiantes debían estudiar la historia antigua de los judíos, cristianos y los musulmanes de la España medieval, comprendiendo dicho periodo y parte del siglo XV. A grandes rasgos, el estudiante aprendería sobre aspectos tanto positivos como negativos de las relaciones entre las personas de aquella época, teniendo en cuenta la coexistencia tanto religiosa como cultural y social que había en Plasencia.

Además de lo anterior y, *"lo que es más importante, los estudiantes contribuyen a un esfuerzo académico internacional ayudando a transcribir manuscritos"*. [8]

Según manifiestan los diseñadores del curso y profesores, como lo son el Dr. Roger L. Martínez-Dávila y Dr. Kathryn Andrus *"Muchos de nosotros creemos que nuestro mundo existe a lo largo de una trayectoria del progreso humano. Este curso trata sobre la duda, el cuestionamiento y las complicaciones. Se abre con la proposición de que nuestro mundo actual no es ni excepcional ni especial, sino que es el beneficiario y el esclavo de la historia."* [9].

1.3.1. Objetivo del MOOC

En el curso, los estudiantes podrían aprender sobre aspectos tanto negativos como positivos de la coexistencia que se vivía en Plasencia, España, al mismo tiempo que colaboraban en una investigación académica internacional Revealing Cooperation and Conflict Project (RCCP). Por medio del MOOC era posible que se aprendiera también sobre los desafíos que tuvo que pasar la comunidad en una sociedad en donde debían convivir personas de diferentes religiones y culturas. Otro punto importante del que se hacía énfasis en el curso online era en las familias tanto judías nobles, comerciantes, clanes de caballeros y caballeros medievales.

Los inscritos al curso contribuirían al trabajo que se está llevando a cabo con respecto a la reconstrucción de cómo era la interacción, tanto cooperativa como conflictiva, en Plasencia. Todo se realizó por medio del estudio de la España medieval y el entendimiento de algunas temáticas nombradas en las Actas Capitulares de la Catedral de Plasencia (documentos base para el trabajo de transcripción).

Las nombradas Actas Capitulares de la Catedral de Plasencia son los documentos que se utilizaron para que los estudiantes trascribieran su contenido y así mismo se conociera mucho más sobre las personas históricas de aquel entonces, este punto es importante recalcarlo porque gracias a dicha información es posible reconstruir las relaciones interpersonales y entender parte de la importancia que tenía la familia Carvajal dentro la sociedad y así revelar condiciones políticas y religiosas que se crearon, logrando entonces alianzas o conflictos que es importante entender para el fin último del MOOC.

Se resaltan entonces aspectos importantes como, el estudio y evaluación a la vida de judíos, cristianos y musulmanes; las relaciones interpersonales entre dichas comunidades bajo ciertas creencias religiosas y bajo la presión social y cultural que implicó; la transcripción de las Actas Capitulares de la Catedral de Plasencia; el descubrimiento y estudio de personajes históricos.

La primera parte del curso se centró en el estudio de la historia de la Europa medieval general, para pasar a centrarse en el estudio específico de la España medieval y finalmente en la historia de Plasencia. La segunda parte del curso tuvo que ver con la transcripción de las Actas Capitulares de la Catedral de Plasencia por lo que el Dr. Martínez-Dávila incursiona en temas como la paleografía, da a conocer sus estudios y avances en el tema para pasar a permitir que los estudiantes hagan sus

propias transcripciones de fragmentos de los documentos a partir de lo aprendido.

Los objetivos específicos del curso, descritos por los profesores son [9]:

- *Los estudiantes estudian y entienden la historia de la España medieval y la comunidad de Plasencia.*
- *Los estudiantes exploran el mundo de los manuscritos y los textos medievales.*
- *Los estudiantes aprenden a leer documentos históricos.*
- *Los estudiantes transcriben y evalúan documentos.*

1.3.2. Datos generados por el MOOC

"La fuente principal que los estudiantes transcribirán en este curso es el Libro Cinco (1499-1513) de las Actas Capitulares de la Catedral de Plasencia, que es un documento censal de tipo contabilidad que detalla las actividades y transacciones comerciales de la catedral. La mayoría de los estudiantes trabajarán con la transcripción del siglo XIX del texto original del siglo XV. Los estudiantes que deseen desafiarse a sí mismos tendrán la oportunidad de trabajar con el texto original de finales del siglo XIV/XV." [9]

Respecto a lo anterior, es válido mencionar que se vuelve tedioso el proceso para transcribir y digitalizar las Actas Capitulares de la Catedral de Plasencia pues no es tan sencillo entender su contenido a simple vista (como se ve en la imagen (1.3)), por lo que los interesados de todo el mundo colaboran en la realización de las transcripciones por medio de ejercicios prácticos los cuales, son básicamente fragmentos de texto de las actas que se le entregan al estudiante y que éste debe transcribir siguiendo parámetros aplicables que como ya se mencionó, son enseñados a partir de documentos sobre paleografía que ya ha realizado el Dr. Roger L. Martínez-Dávila y que se enseñan durante el curso online.

Las respuestas a los ejercicios propuestos durante el MOOC, se almacenan en una base de datos que es gestionada por el Dr. Roger Louis Martínez-Dávila y que es compartida a éste proyecto con fines académicos; existe cierto nivel de confidencialidad de los datos allí almacenados por lo que se tratan y manipulan con discreción. La base de datos se encuentra en formato *.xlsx* y almacena directamente los resultados arrojados y almacenados por la plataforma coursera.org. Dichos resultados son pertenecientes a los de los estudiantes que participaron en el curso para el año 2014.

Se tienen entonces datos relevantes a:

- Código de identificación del estudiante relacionado directamente con una transcripción realizada.
- Información de confidencialidad de la información almacenada.
- Transcripciones identificadas por número de imagen, grupo del curso y relacionadas directamente al estudiante que la realizó.
- Identificación y extracción de años mencionados en los fragmentos de las Actas Capitulares de la Catedral de Plasencia.
- Identificación y extracción de lugares mencionados en los fragmentos de las Actas Capitulares de la Catedral de Plasencia.
- Identificación y extracción de personajes históricos y nombres, mencionados en los fragmentos de las Actas Capitulares de la Catedral de Plasencia.
- Ejercicios marcados como bonos dentro del curso.
- Nombre de la persona y nacionalidad de la persona a la que se atribuye la transcripción almacenada.

La información entregada en dicha fuente de datos o base de datos es el motor de arranque para iniciar con la contextualización y preparación de los datos.

Capítulo 2

Justificación

Considerando que desde diferentes ramas de la informática es posible aportar, desde su propio desarrollo, a ámbitos de estudio que no necesariamente se relacionan con la ciencia y además teniendo en cuenta que los resultados demuestran apoyo no solo para la investigación sino también para la academia, es totalmente provechoso poder colaborar desde la perspectiva del desarrollo de algoritmia y metodologías en proyectos humanísticos como es el caso de ***RCCP***. Es de suma importancia encontrar más información que no ha sido realmente evidente acerca de Plasencia y la convivencia entre sus habitantes, lugares y los roles que las personas podían llevar a cabo pues, aportar directamente a dicho proceso da como resultado directamente avances para Dr. Martínez-Dávila en su investigación. Lo anterior es posible trabajarlo con ayuda de las tecnologías y metodologías (adaptables) que la ciencia de la computación expone, para éste proyecto lo más conveniente es hacer uso la minería de texto en este caso histórica, como técnica de recuperación y organización de la información.

Desde los avances que ha tenido la investigación, se conoce que existen algunos interrogantes sobre el desarrollo de la vida de las personas pertenecientes a la ciudad e incluso se han podido descubrir nuevos hechos que hacen que sea necesario explorar más profundamente las Actas Capitulares de la Catedral de Plasencia. Un ejemplo de lo anterior es La Familia Carvajal y el interrogante que surge al rededor de la misma es que La Familia Carvajal vivía una especie de *transformación* pues no seguían las formas de vida tradicionales para la era medieval. Lo anterior hace referencia al hecho de que en aquella época el linaje de un integrante de la familia junto con su rol o actividad a desempeñar estaba prácticamente establecida, entonces, si un padre era sastre, con toda seguridad, sus hijos también lo serían, lo más seguro es que así continuaría generación tras generación.

Por medio de éste desarrollo y la creciente posibilidad de que haya información oculta esperando a ser encontrada, es posible mejorar la calidad de los descubrimientos actuales y también dar una posible respuesta a preguntas puntuales que hayan surgido a lo largo de la investigación. Una de las pregunta más interesantes que el Dr. Martínez-Dávila se plantea (relacionada con el ejemplo de La Familia Carvajal y su inusual actuar) y de la que queda decir que no es fácil de responder sin un análisis detallado, se encuentra formulada específicamente en el vídeo No. 13 del MOOC ya antes mencionado, con título ***Video #13 - Lecture: "Jews, Catholics, and Converts: Reassessing the Resilience of Convivencia in Fifteenth Century Plasencia, Spain"*** justo en el minuto 00:43:17, corresponde a la siguiente: ***"What changed for this carvajal family during this time period that would have facilitated or could have helped to created a new family?"*** (¿Qué ha cambiado para ésta familia Carvajal durante este período de tiempo que habría facilitado o podría haber ayudado a crear una nueva familia?); esto después de haber explicado un poco más a fondo lo que se menciona anteriormente, lo relacionado con el linaje familiar y las tradiciones familiares de aquella época.

Desarrollando metodologías que permitan a historiadores contestar a preguntas como la anterior, que para el mundo de la investigación histórica son realmente de mucho valor para reconstruir el pasado, hace mucho más fácil que los expertos puedan probar o refutar sus hipótesis planteadas. No solo es posible dar respuesta a preguntas puntuales sino que también se generan muchas otras junto con su posible explicación, no solo enriquecerán las diferentes investigaciones, también permitirá resultados más profundos y certeros.

Además de lo anterior, gracias a la lectura detenida y el análisis de las Actas Capitulares de la Catedral de Plasencia por parte del historiador, a él le ha sido posible identificar que la historia que estas presentan difiere en muchos aspectos con la historia que se presenta actualmente, claro está, centrado siempre el foco de la investigación. Aunque en la práctica es imposible encontrar la verdad absoluta sobre el pasado, pues depende de aspectos externos como también de la percepción de cada académico, sí es posible encontrar "respuestas" mucho más acertadas y certeras sobre lo que sucedía. El uso de herramientas tecnológicas e informáticas sobre la gran cantidad de información que se tiene actualmente para estudiar, sobre Plasencia y en sí los aspectos económicos, políticos y sociales de la época, puede ayudar de manera absoluta a tener un mejor conocimiento y dominio de los textos y datos en un tiempo

más razonable, por lo que es importante destacar que, con el afán de fomentar la participación de dichas herramientas tecnológicas e informáticas en diferentes casos en donde puedan aportar valor (en este caso, las ciencias humanas), se desea apoyar al desarrollo del proyecto RCCP con el fin de cumplir con los objetivos propuestos además de colaborar a la comunidad y dejar huella con un producto que permita mejorar el proceso que un historiador, en este caso Dr. Martínez-Dávila , debe llevar a cabo para analizar grandes volúmenes de textos e información obteniendo así mejor dominio y entendimiento de aquel corpus.

A partir de lo anterior, cabe resaltar que con éste proyecto, no solo se apoya al desarrollo de las diferentes tecnologías como también al desarrollo de la línea de la minería de texto histórica sino que también es posible apoyar al desarrollo del proyecto de carácter internacional y abierto para llegar así a refutar o aprobar la hipótesis propuesta.

2.1. Aporte al proyecto Revealing Cooperation and Conflict Project

Desde el punto de vista del proyecto ***RCCP*** es de suma importancia, no solo para el desarrollo del mismo, sino también para el Dr. Roger Louis Martínez-Dávila , encontrar una posible forma de analizar más rápidamente el corpus base con el que cuenta, con el fin de seguir construyendo y avanzando en su proyecto, junto con lo anterior también es interesante para él, poder ver plasmadas gráficamente aquellas relaciones “ocultas” ya mencionadas para recrear de manera más próxima el lugar. Desde el punto de vista de la investigación histórica y humanística, lo que ocurría en aquel lugar y en aquel periodo junto con el hecho de que las actuales hipótesis pueden desviarse un poco de la situación real que se vivía, es bastante interesante de responder y de analizar pues apalanca directamente el desarrollo de éste campo de trabajo.

Haciendo uso de la minería de texto sobre las Actas Capitulares de la Catedral de Plasencia que está analizando el Dr. Martínez-Dávila , es posible no solo agilizar, sino también hacer menos tedioso el proceso de trabajo que debe hacerse para responder a los diferentes interrogantes. El volumen de textos y de datos con los que cuenta el historiador es considerablemente grande y el esfuerzo que debe hacerse para gestionar tal masa es bastante. Teniendo en cuenta que ya se cuenta con una parte las Actas Capitulares de la Catedral de Plasencia digitalizadas junto

con las respectivas transcripciones, pueden adecuarse para fines tales como análisis y clasificación, relación entre palabras, localización, descubrimiento de lugares y personas, entre otros; así entonces es posible crear objetos de análisis, **gracias a la intervención de las ciencias de la computación y la tecnología, que de otra manera sería muy difícil obtener**, para dar un paso más allá analizando con más precisión y exactitud, logrando entonces tener un conocimiento profundo sobre los textos en tópicos no evidentes, preservar información valiosa y dentro de lo más destacable, se logran disminuir considerablemente los tiempos empleados para los análisis y lectura de los documentos.

Por medio de la minería de texto sobre las Actas Capitulares de la Catedral de Plasencia, no solo es posible lo anterior, sino que es mucho más sencillo encontrar factores ocultos sobre el ambiente social, político, económico y religioso que se vivía en aquel momento en Plasencia, España, y sus alrededores y que llevaron al curioso desenvolvimiento del lugar y sus pobladores, a pesar de las inusuales condiciones para la época.

Al permitirse un diálogo "dinámico" entre la investigación histórica y las tecnologías emergentes se fomenta la experimentación y la colaboración pues dichas disciplinas juntas, están generando un impacto significativo en el contexto académico desde hace más de una década (ver capítulo 5) .

2.2. Aporte desde la Ingeniería

Debido a que el proyecto es multidisciplinario y se destaca la constante comunicación y colaboración entre las Humanidades y las Ciencias de la Computación y la informática, también es importante recalcar el aporte que se hace desde la ingeniería, que por su puesto tiene un tinte más técnico. El aplicar minería de texto, implica todo un proceso que involucra entendimiento de tecnologías, algoritmos que en su mayoría estadísticos, herramientas, metodologías y otros, que permiten llevar a cabo el proceso de trabajo de la minería (ver capítulo 7) .

A pesar de la suma importancia que tiene la historia que se pueda descubrir, es de interés también la metodología que se llevará a cabo para encontrar las relaciones interpersonales no evidentes, a partir del entorno del lugar, descrito en los documentos base que se están utilizando. La metodología que se aplicará para la gestión de la información con la que se cuenta (en formato no estructurado) nace a

partir de ideas ya existentes que han surgido en otras investigaciones, lo principal es que junto con la colaboración al proyecto ***"Revealing Cooperation and Conflict Project (RCCP)"*** se podrá proponer una metodología apoyada en la tecnología y en La Minería de Texto Histórica, que funcione bien para investigaciones de carácter histórico o afines y que permita analizar colecciones de texto con el fin de obtener un cuerpo de datos con el que se pueda capturar temas claves junto con conceptos para descubrir relaciones ocultas entre su contenido, patrones o tendencias de los datos y la información, sin conocer términos o conceptos fijos que haya usado el autor para expresarse.

Cabe resaltar que debido a la naturaleza del proyecto, aunque esté ligado a la investigación y disciplina histórica, también está ligado a investigación de carácter tecnológico y de cierta manera científico debido a que para poder generar conocimiento a partir del gran volumen de datos e información con el que se trabaja, es necesario ahondar en metodologías, técnicas y herramientas que permitan el análisis, modelamiento, tratamiento, limpieza, calidad y otros aspectos fundamentales de los datos datos y su preparación.

Se pretende hacer un aporte significativo no solo a la comunidad que le interesa la historia y el pasado sino también a la comunidad orientada a las ciencias o ramas de la matemática, estadística, ingeniería y afines, que les pueda interesar éste desarrollo. La metodología que se llegue a proponer junto con una serie de diccionarios para la investigación, que contienen palabras clave de la época (ver capítulo 3) , estarán disponibles para quien requiera utilizarlos en otras investigaciones o incluso para resolver algunos otros interrogantes de ésta misma.

Dentro de las especificaciones de innovación que se plasman en la presentación del proyecto ***"The revealing Cooperation and Conflict Project"*** se indica que las bases de datos del mismo son enteramente de código abierto junto con las codificaciones de las API. Dentro de dichas bases de datos se encontrarán los datos que sean descubiertos por la comunidad a partir de la transcripción e indexación de las Actas Capitulares de la Catedral de Plasencia. Dado que el núcleo del proyecto dirigido por el Dr. Martínez-Dávila está basado en contar con un innovador Sistema de información geográfico (Geographic Information System) (GIS) que involucre datos sociales, económicos y demográficos será entonces vital para el éxito del mismo contar con conocimiento que difícilmente puede ser generado sin hacer uso de técnicas o herramientas para el manejo y entendimiento de grandes volúmenes de

datos.

Capítulo 3

Objetivos

3.1. Objetivo general

Presentar al proyecto **Revealing Cooperation and Conflict Project (RCCP)** una metodología con base tecnológica para optimizar los tiempos y mejorar calidad de análisis de las Actas Capitulares de la Catedral de Plasencia, esto mediante técnicas de **minería de texto**.

3.2. Objetivos específicos

- Conocer, entender y profundizar respecto a los nodos y diferentes funcionalidades con las que la herramienta de software KNIME cuenta para realizar minería de texto.
- Utilizar la herramienta KNIME para realizar la Minería de Texto sobre los documentos denominados Actas Capitulares de la Catedral de Plasencia, lo que se refiere específicamente a realizar la construcción algorítmica que permita gestionar las Actas Capitulares de la Catedral de Plasencia en base a las diferentes funcionalidades de la plataforma.
- Indagar e investigar sobre otros proyectos afines que hayan surgido anteriormente o que surjan paralelamente a éste para de ésta manera contar con visión sobre las diferentes técnicas y metodologías que pueden aplicarse para hacer Minería de Texto y encontrar asociaciones complejas.
- Colaborar en el desarrollo de una nueva línea en la minería de texto: **la minería de texto histórica**.

- Elaborar una metodología para la gestión y análisis de documentos de carácter histórico, teniendo en cuenta las diferencias y similitudes entre el español antiguo y el español moderno.
- Generar tres diccionarios para hacer análisis de documentos de la época (siglo XIV a XVI) de Plasencia, España. Dichos diccionarios quedan para el libre uso de la comunidad y otras investigaciones de la misma índole o que aplique.

Diccionario de nombres	Diccionario de roles	Diccionario de lugares

- Generar un diccionario de *stop-words* (ver capítulo 4.2.2) para el análisis de textos comprendidos en los siglos XIV a XVI de Plasecia, España.

Capítulo 4

Minería de texto

4.1. ¿Qué es?

La minería de texto también es conocida como minería de datos de texto [10], esta consiste en la extracción de información y patrones no triviales contenidos en los diferentes documentos de texto usados para el análisis, con esta se busca identificar, deducir y ampliar el conocimiento sobre los documentos de texto tratados [10].

Tanto la minería de texto como el análisis de texto son una *sombrilla* muy amplia que involucra diferentes términos y técnicas especializadas para el procesamiento de los diferentes datos estructurados y los datos no estructurados [11], si bien estos algoritmos pueden llegar a ser completamente diferentes, en el fondo tienen el mismo objetivo, "realizar una transformación de texto a números" [11], esto con la finalidad de poderlos aplicar completamente a todo el documento.

Es complejo dar una única definición de minería de texto puesto que esta involucra siete áreas de conocimiento de siete campos diferentes, aunque todos se encuentran relacionados por el campo de la estadística tal y como se aprecia en el diagrama de Venn en la figura 4.1 [11].

Las áreas que complementan la minería de texto son:

1. Search and Information Retrieval
2. Document clustering
3. Document classification
4. Web mining
5. Information Extraction
6. Natural Language Processing
7. Concept extraction

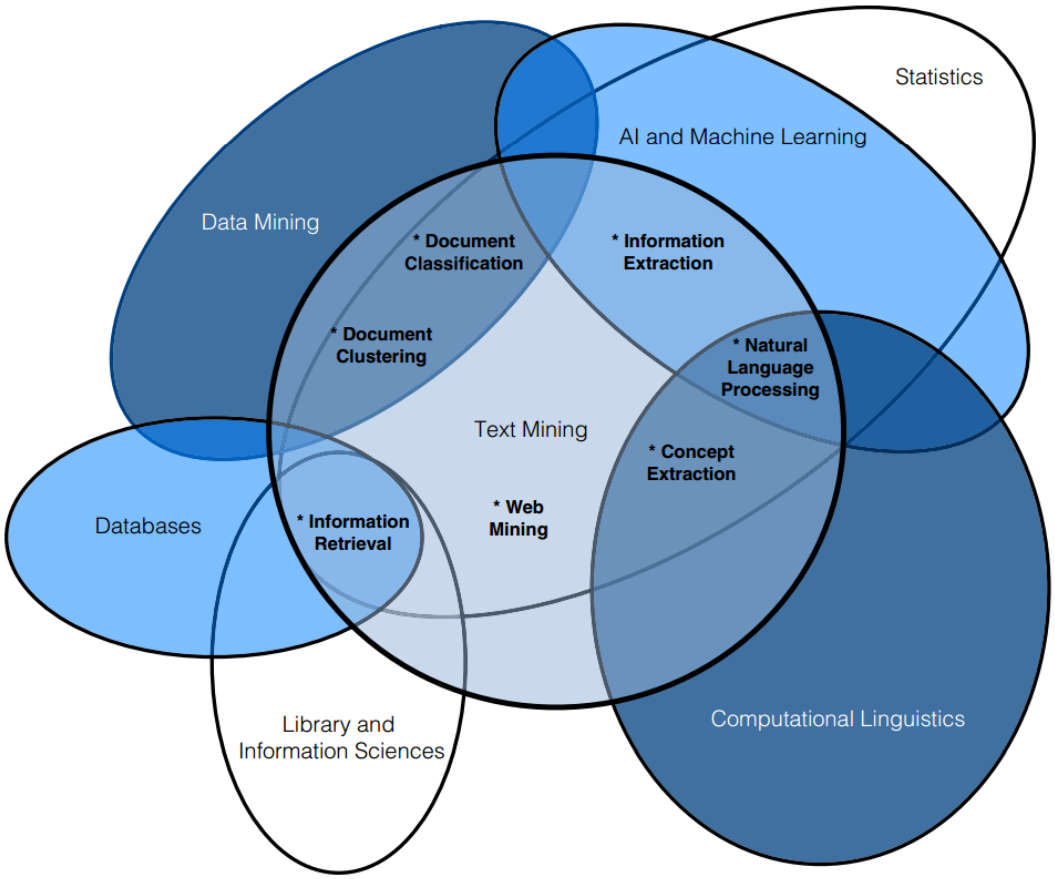


Figura 4.1: Las siete áreas practicas del análisis de textos

Estas siete áreas que existen debido a la comunicación con los campos de estadística, Inteligencia artificial (Artificial Intelligence) (IA) y aprendizaje de máquina, lingüística computacional, minería de datos, Bases de datos (Databases) (DB) y bibliotecas y ciencias de la información, esto es porque gracias a la combinación de las técnicas propias de cada campo es posible generar mejores resultados, en la tabla 4.1 presentada en el libro *Practical Text Mining* [11] se evidencian solo algunos de los temas más comunes en el campo de la minería de texto y el área mayormente usada para 'atacar' dicho tema.

4.2. Ciclo de trabajo

Al igual que la minería de datos, la minería de texto sigue un marco de trabajo, o proceso, muy similar, el cual va desde la definición del problema hasta la implementación de los modelos matemáticos planteados buscando dar solución o respuesta a el problema establecido al comienzo del proceso; la diferencia radica en que sigue los siguientes cuatro pasos dentro del proceso metodológico

Tema	**Área**
Búsqueda de palabras clave	Search and Information Retrieval
Resumen de documentos	Search and Information Retrieval
Agrupación de documentos	Document clustering
Semejanza de documentos	Document clustering
Selección de documentos por características	Document classification
Análisis de sentimientos	Document classification, Web mining
Clasificación de documentos	Document classification
Descubrimiento electrónico	Web mining
Rastreo web	Web mining
Extracción de entidades	Information Extraction
Etiquetado gramatical / PoS	Natural Language Processing
Tokenización / Lematización	Natural Language Processing
Solución de interrogantes	Natural Language Processing, Search and Information Retrieval
Identificación de sinónimos	Concept extraction
Modelamiento de temas	Concept extraction

Cuadro 4.1: Relación entre los temas y las diferentes áreas de la minería de texto

4.2.1. Corpus

También conocido como **text corpus**, en ingles, es el conjunto de documentos que van a ser usados para el trabajo de investigación, generalmente son documentos digitales, aunque pueden ser físicos.

Es importante tener en cuenta que entre más grande sea el **corpus**, será posible que la recuperación de la información se haga de manera más completa.

Para el caso particular de este trabajo el **corpus** estará compuesto por las Actas Capitulares de la Catedral de Plasencia, más específicamente aquellas que ya se

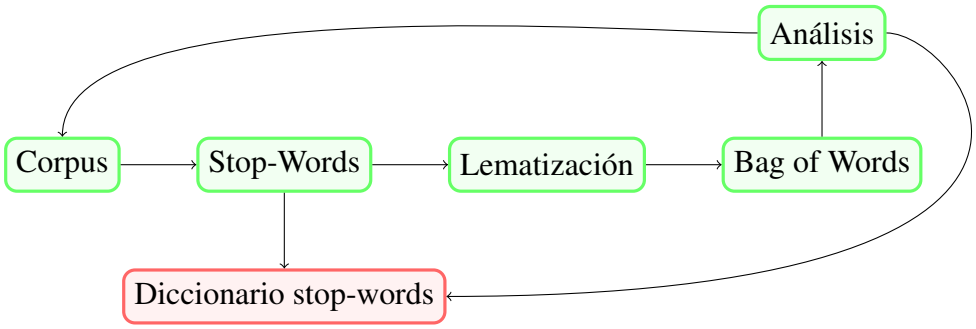


Figura 4.2: Diagrama del ciclo de trabajo

encuentran digitalizadas.

4.2.2. Stop - Words

Este paso consiste en la eliminación de pronombres, artículos y proposiciones, esto debido a que éstos no hacen un mayor aporte al trabajo de análisis, es decir, es indiferente si se encuentran o no en el texto, a esto también se le conocen como palabras vacías. Esto tiene como finalidad ayudar a realizar un análisis con mayor velocidad, en los siguientes pasos.

Ejemplo 1

Antes del proceso (Corpus)	He adoptado una perra con sus cachorros.
Eliminación de los **stop-words**	He adoptado una perra con sus cachorros.
Después del proceso	adoptado perra cachorros

Cuadro 4.2: Ejemplo stop-words

En el ejemplo del cuadro 4.2 se puede observar después de la eliminación de los **stop-words** la finalidad de la oración no se ve afectada, es más rápida de leer y analizar.

Éste es un proceso muy usual, debido a que los motores de búsqueda de Internet lo aplican siempre, un claro ejemplo es Google [12], pues para realizar una búsqueda más rápida y precisa es necesario este paso, puesto que de no realizarse esto incluiría en la búsqueda páginas que hagan uso de estos artículos, pronombres y/o proposiciones, lo cual resulta bastante ineficiente.

Ejemplo 2

Palabras	Lema
Perro, Perros, Perritos, Perra, Perras, Cachorro	Perro

Cuadro 4.3: Ejemplo 1 lematización

Es importante tener en cuenta que casos no se debe de aplicar éste proceso, puesto que se pueden eliminar partes de un nombre, por ejemplo, si tenemos una banda de música llamada *Los de adentro* y le aplicamos este procedimiento, el nombre de la banda se vería reducido a *adentro*, en este caso si cambia completamente el sentido, pues los **stop-words** identificados son una parten fundamental del nombre de la banda [12]; para evitar esto es necesario tener un filtro en el cuál se le indique al proceso cuales no son **stop-words** como en este caso.

Para la correcta identificación de los **stop-words** es necesario contar con un diccionario de los mismos. Para éste se puede hacer uso de uno ya *existente*, crear uno nuevo, esto se puede deber a que *no existe* uno sobre el idioma tratado o porque el existente no se aplica en su totalidad para el caso de estudio.

4.2.3. Lematización

En la lingüística computacional la lematización es un proceso importante y necesario al momento de realizar un procesamiento automáticamente del **corpus**. Es la representación de una normalización de los datos de texto en donde cada palabra se analiza para posteriormente reducirla a un lema [13, 14], la normalización del texto es crucial en los idiomas ricos en palabras, como lo es el caso del español, pues, es posible expresar una misma idea con una gran cantidad de palabras diferentes [14].

Para la realización correcta de éste paso es necesario hacerse con un **diccionario morfológico** lo más completo posible [14], pues de este se depende para lograr un buen resultado para la disminución de errores durante todo el proceso siguiente de la minería de texto.

La lematización consiste en la eliminación de plurales, pasados, futuros, entre otras cosas, esto con la finalidad de dejar solamente la raíz de la palabra, ésta será el lema de la palabra[13, 14].

Ejemplo 3

Después de **stop-words**	adoptado perra cachorros
Proceso de **Lematización**	adoptado perra cachorros
Después del proceso	adoptar perro perro

Cuadro 4.4: Ejemplo 2 lematización

En el ejemplo del cuadro (4.3) el resultado de la lematización es **perro**, puesto que todas las palabras usadas hacen referencia al mismo animal, que es *perro*, es decir, todos estos son sinónimos, diferenciación entre masculino y femenino y plurales.

En el ejemplo del cuadro (4.4) se puede ver claramente el proceso de **lematización**, en el cual se analiza cada palabra y se cambia por el lema correspondiente, tal como se mostro en el ejemplo 4.3. Ya en éste paso es donde, al leer nosotros mismos la frase, esta ya pierde un poco su significado original, puesto que ahora da a entender que se va a adoptar uno o dos perros. Es en este paso donde se suele pensar que al cambiar las palabras el resultado del análisis no va a ser correcto, este pensamiento ocurre debido a que se piensa en un análisis no matemático, puesto que matemáticamente hablando se trata de una simple simplificación de las variables, por lo tanto el resultado del análisis no se ve afectado [13, 14].

Capítulo 5

Antecedentes de minería de texto histórica

5.1. Expressing complex associations in medieval historical documents: the Henry III Fine Rolls Project

Se da inicio a este proyecto en El King's College London (Universidad pública de investigación) desde el Centro de Informática para las Humanidades, el cual específicamente, hace parte del Departamento de Humanidades Digitales de la Universidad. Los responsables de la investigación y divulgación de los resultados de la misma son Arianna Ciula, Paul Spence y José Miguel Vieira. La investigación se centra en el hecho de que, el uso de tecnologías asociadas con la gestión del conocimiento y manejo de información, puede ayudar a expresar y mostrar gráficamente asociaciones complejas entre "entidades" en textos históricos. Para esta investigación el texto ha sido clasificado o tratado en formato XML. En particular, se pretende facilitar la interpretación de información oculta dentro del texto como lo son asociaciones implícitas en el texto.

Para llevar a cabo la investigación se hizo uso de algunas tecnologías y/o metodologías que están directamente ligadas con los objetivos de la misma, un ejemplo de ello es RDF o Resource Description Framework que se entiende como ciertas especificaciones que establece la Word Wide Web Consortium para el modelado de información que se implementa en recursos web y se centra en poner a disposición un modelo para el intercambio de datos en la web. La anterior y otras herramientas se utilizaron con el fin de obtener información explícita sobre personas, lugares y

entidades dentro del texto. Todo lo anterior es demostrado bajo el desarrollo del proyecto denominado *"Henry III Fine Rolls project"* que además, logra exponer una de las problemáticas más importantes que se presentan en los proyectos e investigaciones de carácter humanístico y es la brecha que existe entre la información primaria o las fuentes de información primaria con las que cuenta un historiador y el contexto de interpretación que existe a partir del análisis de los mismos; por medio de la realización de la investigación y del proyecto *Henry III Fine Rolls project*, los autores logran proponer un modelo que representa la complejidad entre las dos partes que componen la brecha y que además facilita el proceso de interpretación de las fuentes de información.

Entrando más en detalle acerca del proyecto *Henry III Fine Rolls project*, es importante destacar el corpus principal sobre el que se trabajó, según el mismo artículo del proyecto, se sabe que existe *"un total de sesenta y cuatro rollos que contienen alrededor de 800 membranas de pergamino, uno para casi todos los cincuenta y seis años del reinado de Enrique III de 1216-72, sobreviven en Los Archivos Nacionales. Cada rollo se compiló en latín por un puñado de escribas. Tomados como un cuerpo de evidencia documental, 'los rollos son de Importancia en el estudio de la historia política, social y económica y del gobierno y la administración a nivel local y nacional' "* [15]. Debido a que los textos estaban en latín, se hace el esfuerzo por publicar algunos fragmentos, sin definición de tiempo, en ingles para que pudieran ser consultados por medio de un sitio web publicado hacia el año 2007, allí también se hicieron publicaciones de índices o diccionarios de lugares, personas y temáticas; traducciones del texto, imágenes del texto, representaciones gráficas de las asociaciones encontradas, entre otros aspectos relevantes.

La fuente de información o los rollos mencionados anteriormente, se manipularon de tal manera que se cumpliera con los estándares para la investigación y además se permitiera una preparación y manipulación de los mismos más sencilla, haciendo uso del marco de trabajo denominado Electronic Text Encoding and Interchange fue posible representar la estructura de los documentos y así mismo generar un recurso que pudiese ser utilizado como guía no solo para extraer información sino para entender el contexto histórico de la época profundizando en aspectos que el investigador/historiador pudiera tener interés como por ejemplo política, economía, cultura. **De los resultados más importantes de la investigación se destaca el hecho de que, *"para el periodo de 1216 a 1224 (un cuarto del periodo cubierto por la primera fase del proyecto) los investigadores ya habían sido capaces de identifi-***

car 3.436 hombres, 499 mujeres, alrededor de 4.070 lugares y 1.159 sujetos." [15].

Los resultados más inmediatos que se obtuvieron al capturar relaciones implícitas en el texto, siguiendo la estructura y resultados obtenidos a partir del marco de trabajo se reflejaron en la obtención de conocimiento a un mayor nivel de profundidad acerca de las personas, los lugares y entidades; la cantidad de información obtenida acerca de ello, se basa en que fue posible que a cada persona se le asociaran inmediatamente componentes referentes a sus relaciones con otras personas, a su ocupación, lugares en común, etc. Después de obtener dicho resultado era necesario continuar con el proceso y entonces obtener las relaciones complejas y no tan evidentes contenidas en el texto. Los datos relevantes se extraen y se hace uso de tecnologías como lo fueron RDF y OWL6 (con RDF se definió la estructura de la información y con OWL se define por medio de programación, la estructura semántica de las relaciones de tal manera que la semántica compleja pueda reconstruirse desde conceptos más simples); en palabras de los autores: *"En términos más prácticos, nuestra elección de RDF / OWL fue impulsada por una serie de factores, entre ellos: nuestro interés en facilitar las interconexiones con otros proyectos de Humanidades Digitales, especialmente aquellos que abarcan períodos y áreas temáticas similares, utilizando estándares internacionales en el corazón de la Semántica Web; La madurez relativa de las herramientas de soporte; Y el hecho de que RDF / OWL puede expresarse en formato XML, lo que permite la fácil re-definición de los datos para la entrega de la web."* [15]

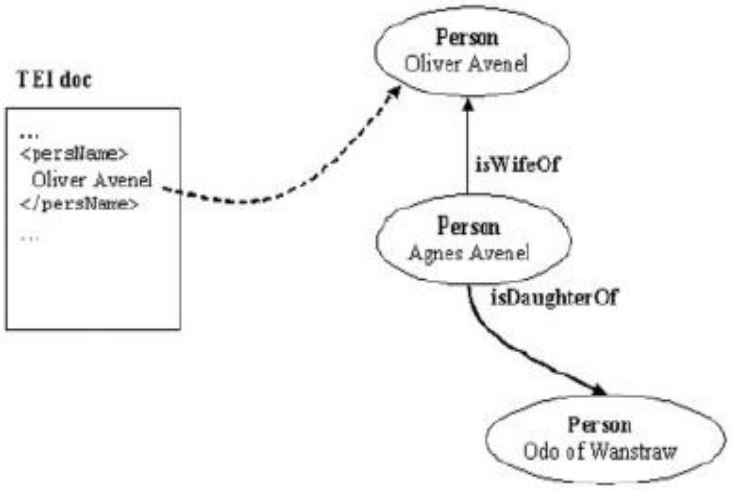


Figura 5.1: Ejemplo de asociaciones a partir de documento estructurado

El siguiente paso fue iniciar con la creación de la ontología para lo cual fue nece-

sario hacer un análisis con el debido detalle requerido para poder representar dicho conocimiento correctamente, los documentos TEI fueron la clave para recrear la ontología y así mismo incluir la información para aquel modelado. Por la naturaleza de la creación de la ontología, los personajes, lugares o entidades no existirían en la misma hasta que no fueran mencionados en el corpus a partir del cual se modela; del XML resultante, se extraía información como por ejemplo relaciones familiares y se traspasaban a la ontología para ir formando el contexto poco a poco, los parentescos o relaciones muy específicas debieron ser incluidas "manualmente" dentro de la ontología.

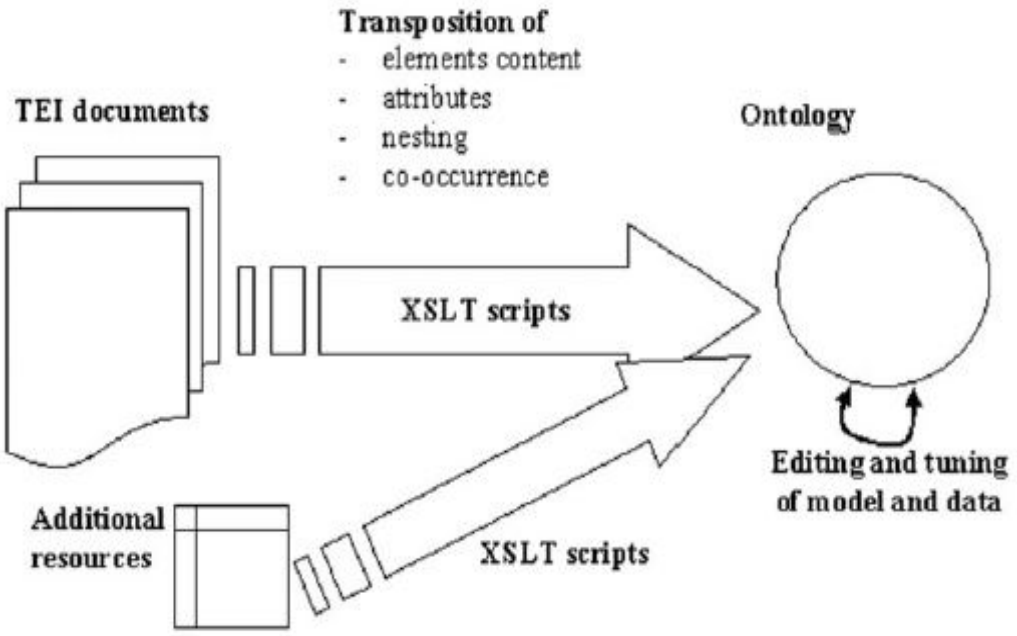


Figura 5.2: Proceso para construir la Ontología

Para nutrir aun más la ontología fue necesario encontrar patrones de comportamiento de los datos y de la información de tal manera que el proceso para encontrar elementos anidados fuera de forma automatizada y se redujeran los tiempos considerablemente, junto con lo anterior también fue posible descubrir diferentes categorías de sucesos al que podía pertenecer una asociación; por ejemplo el que hubiese un anidamiento de un nombre de un lugar dentro del nombre de una persona podía implicar que dicho lugar se refiriese al apellido del personaje (*Nombre_Personaje* de *Nombre_Lugar*), también podía implicar la relación estrecha entre dicho lugar y la persona. El rpoceso anterior muestra una pequeña parte del nivel de automatización al que se pretendía llegar y además lo útil que es al momento de encontrar elementos anidados directamente relacionados con co-ocurrencias y sus relaciones semánticas.

Tener la información en formato XML fue clave porque de cierta manera ratificaban con seguridad el contexto de la información y dan seguridad a la reconstrucción histórica que se lleva a cabo.

Además de todos los beneficios que la construcción de la ontología trajo al proyecto *"Henry III Fine Rolls project"*, también se destaca el hecho de que este modelo podría servir a otras investigaciones de la misma índole y que traten también un tiempo determinado a partir de fuentes históricas. Para expresarlo más fácilmente, la ontología se forma a partir las siguientes especificaciones:

- Los documentos que permiten contextualizar la realidad descrita en los rollos finos utilizados como fuente principal. Se usaron también documentos externos que dieran información afín adicional.
- Se realizó con el fin de describir anidaciones entre personas, lugares y entidades pertenecientes al texto, dar contexto a la posible realidad modelándola y además asegurando de cierta manera el contexto resultante.
- Fue clave para encontrar categorías de roles dentro del texto, finalmente a ello se llegaría. Por ejemplo una categoría válida es el rol que desempeñó una persona que apareció en los textos, por ejemplo Mayordomo, arzobispo, etc; otra categoría válida es un rol que especifica una relación directa con otro rol principal, como por ejemplo Sastre del Rey.
- Para las clasificaciones existen "clases" por medio de las cuales se especifica un rol y posibles subroles. Igual ocurre con las entidades y lugares.

En la visualización 5.3 es posible ver un ejemplo de como resultan las conexiones entre la fuente de datos, la ontología (con "clases incorporadas").

Después de obtener los resultados aplicando el método y todo el proceso propuesto se llega al punto de publicación. Por el hecho de que se trabajó con tecnologías que facilitan el intercambio de datos web, la publicación de los dichos resultados en la página web de la institución fue un poco más sencilla. Los desarrolladores del proyecto hicieron uso de xMod, un marco de publicación diseñado por el mismo Centro de Informática para las Humanidades. El uso de OWL facilitó la capacidad de procesar datos lógicamente para seguir apoyando a la ontología y así encontrar información no evidente en las fuentes (una persona pudo ser "rastreada" dentro de los resultados por su nombre o sus diferentes maneras de aparición en las anidaciones XML, sin la ontología no sería posible. También fue posible clasificar las relaciones

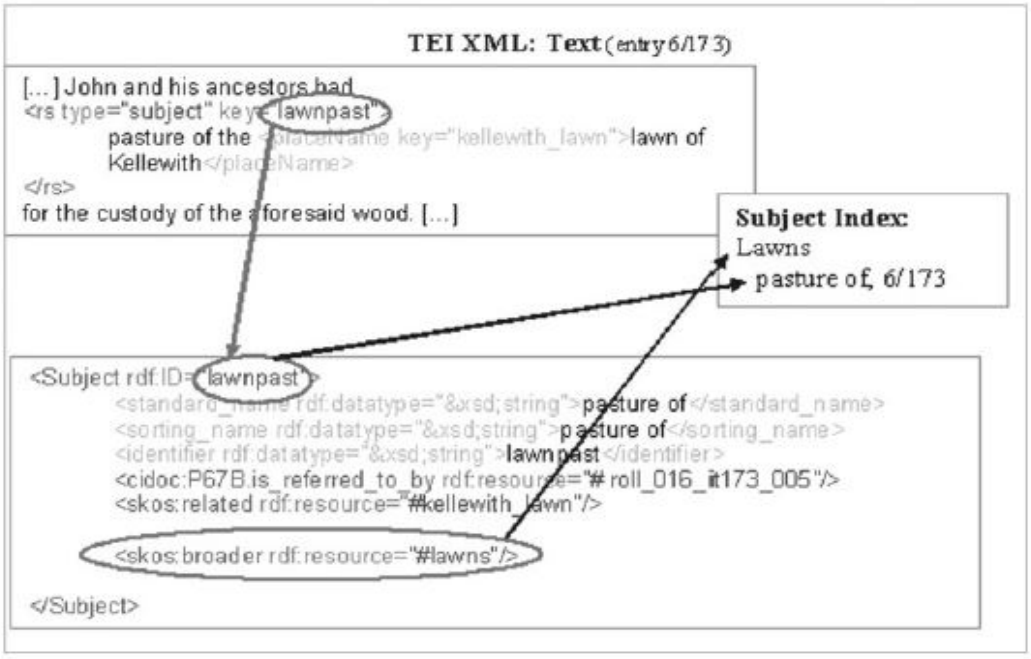


Figura 5.3: Ejemplo de anidaciones entre conceptos, TEI XML

y asociaciones).

En cuanto a los resultados y conclusiones del proyecto, los autores lograron generar un enfoque de transferible a otros proyectos históricos destacando el hecho de que la facilidad que brinda las tecnologías utilizadas hace que la publicación y personalización de publicación de los resultados sea más sencilla y reduzca tiempos. En cuanto a lo que se descubrió desde la fuente de datos, puede ser muy útil para otras investigaciones y otros investigadores ya que puede que las personas, lugares y entidades que aparecen en este caso en los rollos finos, también aparezcan en otros documentos de la misma época y del mismo carácter, lo que es un gran avance en cuestión de iniciar otra investigación. Aunque la ontología no demuestra un contexto totalmente certero, pues la historia depende de la percepción, si tiene un nivel de adaptabilidad bastante destacable, si se diera el hecho de que en otra investigación se descubriera información nueva o se modificara de una u otra manera el contexto dado a la misma, sería evidentemente sencillo permitir que los cambios se hagan efectivos basados en certeza de interpretación.

En cuanto al proceso de análisis de las anidaciones y la relación con la ontología, aun quedaban algunos aspectos por retocar en cuanto a los niveles de anidación, clases y subclases pues los niveles de los mismos en cierto momento dejaban de

tener la lógica y el contexto propuesto. La ontología a parte de esto sí permitió que los investigadores refinaran las asociaciones encontradas apegándose o no al texto, es por ello que da cierta libertad al investigador para que juzgue el contexto e incluso identifique aspectos que no parezcan ser como se indica. Como conclusión valiosa de los autores se destaca *"consideramos que el uso de un enfoque ontológico, junto con el marcaje XML de TEI XML para la codificación de las fuentes, representa un poderoso marco técnico para proyectos históricos, que equilibra el respeto al texto como testigo único con la necesidad de representar su Contexto histórico e interpretación."* [15]

5.2. Visualization of relationships among historical persons from Japanese historical documents

Este proyecto que fue realizado por Fuminori Kimura, Takahiko Osaki, Taro Tezuka, Akira Maeda hacia el año 2013 y surge debido a que los documentos históricos en Japón ya hace un tiempo están siendo digitalizados y por tanto el auge por analizarlos y generar conocimiento en áreas humanísticas va en crecimiento. Surge un percance parecido al problema actual que se tiene con éste proyecto pues los textos claramente no se encuentran en Japonés moderno por lo que la principal justificación para llevar a cabo todo éste proceso es el hecho de presentar una metodología que permita encontrar y visualizar asociaciones poco evidentes entre personajes históricos japoneses a partir de textos históricos, teniendo en cuenta datos sobre lugares y nombres relevantes. Además de lo anterior los autores expresan que es la posibilidad para apalancar el surgimiento de herramientas basadas en componentes tecnológicos que permiten su aprovechamiento al máximo en campos humanísticos en los que sea necesario gestionar textos con japonés pre-moderno ya que hacerlo en las herramientas actuales es absolutamente complicado.

Para realizar la metodología de la que se habla, se hace uso de la **minería de texto histórica** debido a que es posible obtener información que no es muy evidente al analizar manualmente o simplemente al leer el documento, esto con el fin de encontrar relaciones entre las personas según sus actividades y locación geográfica y sacar conclusiones a partir de ello. Se hizo uso de un documento histórico denominado *Hyohanki* del cuál se sabe que es un **diario** escrito por **Noburoni Taira**, correspondiente desde el año **1112 hasta 1187** [16]. Con respecto a éste documento

se resalta el hecho de que es fácil obtener el orden del documento porque las fechas y la escritura diaria definen el mismo.

En este caso, también fue necesario realizar un diccionario que contiene los nombres propios que se mencionan o aparecen en el diario, esto es importante porque de ésta manera se puede dar contexto a dichos nombres para encontrar las asociaciones complejas que se requieren; debido a que los diccionarios que usan las herramientas con las que se aplican las técnicas de minería de texto claramente no contienen "soporte" lingüístico para el japonés pre-moderno es un paso muy importante y, al igual que acá, en éste proyecto que se quiere llevar a cabo es sumamente importante tener los diccionarios ya mencionados en los objetivos específicos debido a que las herramientas actuales (específicamente KNIME) al igual que ocurre con el japonés pre-moderno, tampoco cuentan con "soporte" para el procesamiento del lenguaje natural correspondiente al español antiguo.

Entrando un poco más en detalle con respecto a lo anterior, las técnicas parten de extraer todos los nombres de personas y de lugares existentes en el diario, también se incluyeron los sobrenombres. Para conocer qué tan fuertes eran las relaciones se usó la modalidad de co-ocurrencia entre el nombre de una persona, o su sobrenombre, con el nombre de un lugar, o el sobrenombre de este; la co-ocurrencia ocurre si alguno de dichos nombres propios de personas y de lugares aparecían en el mismo párrafo o en un mismo bloque específico de texto, entonces se considera que existe co-ocurrencia. Se hace mención al hecho de que puede ocurrir que aunque halla co-ocurrencia entre una persona y un lugar, eso no implica que aquella persona hubiera estado o no ahí, aunque el la mayoría de casos ocurre que la persona si estuvo allí en algún momento.

Para realizar la visualización de las relaciones interpersonales usando la información de las co-ocurrencias entre información sobre personas y l se genesneraban vectocacioores de características, en dicho vector cada locación se considera como una dimensión del vector en el espacio, seguidamente se hace uso de el algoritmo de **K-MEANS** formando un total de tres *clusters*. Los clusters en este caso representan que personas de las mencionadas eran seguidoras de un emprerador o de otro según los indicados en el diario, uno correspondiente a los seguidores del emperador Sutoku, otro correspondiente a los seguidores del emperador Goshirakawa y el último mostrando una participación neutral entre ambos emperadores. Los vectores son utilizados para calcular la similitud entre ellos y así poder deducir las relaciones

interpersonales entre personas sin la necesidad de confiar en o recurrir únicamente a los resultados de las co-ocurrencias directas.

El algoritmo correspondiente para aplicar la técnica o metodología propuesta en este caso es el siguiente:

1. Obtener nombres de personas y lugares a partir del texto.
2. Contar la frecuencia de las co-ocurrencias entre cada nombre de una persona y el nombre de un lugar.
3. **Por cada persona:** se construye el vector de características compuesto por el número de co-ocurrencias con el nombre de un lugar que sea componente del vector.
4. Calcular las similitudes entre las personas por medio de la comparación de los vectores ya construidos.
5. Se realizan los clusters usando el método K-means.
6. Realizar las visualizaciones de las relaciones usando las similitudes entre vectores y los resultados de los clusters.

Para la visualización del resultado de los clusters se hace uso de una biblioteca de **Java** la cual es de código abierto y permite dibujar la estructura por medio de grafos. El nombre de aquella biblioteca es **- Java Universal Network/Graph Framework**. *"Es una biblioteca de software que proporciona un lenguaje común y extensible para el modelado, análisis y visualización de datos que se puede representar en forma de gráfico o de la red"* [17]. La arquitectura JUNG soporta una gran variedad de representaciones como grafos dirigidos, no dirigidos, multimodales, bordes paralelos, entre otros. Además de lo anterior facilita la creación de herramientas para analizar datos complejos y las relaciones entre entidades. Se apoya en distintos algoritmos de la teoría de grafos, **minería de datos** y análisis de redes sociales para poder agrupar, descomponer, optimizar, hacer análisis estadístico, etc.

A continuación, en la visualizacion 5.4 se muestra el resultado visual de graficar el resultado de los clusters según la persona y los factores históricos.

En la visualización 5.4 se diferencian las personas según lo siguiente: quienes son seguidores del emperador Sutoku están representados por el nodo cuadrado y quienes siguen al emperador Goshirakawa están representados por el nodo circular.

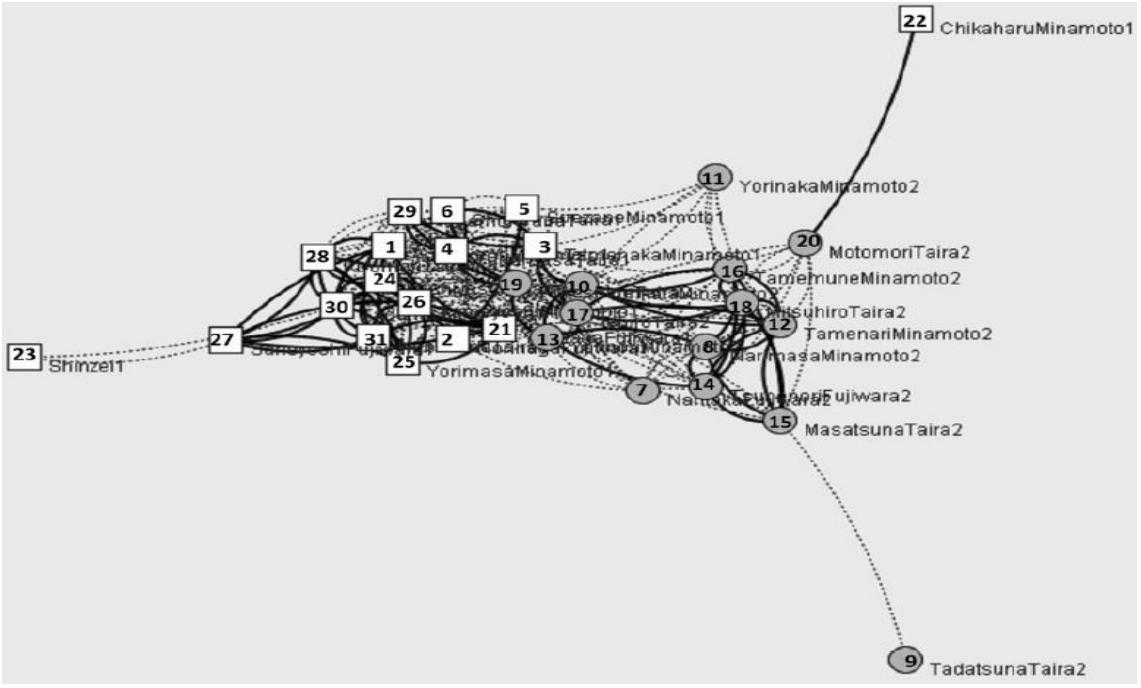


Figura 5.4: Gráfica del relaciones entre personas y factores históricos

El número que está dentro de cada nodo corresponde al mismo número que tiene asociado cada persona en la tabla 5.5. Las líneas se dibujan si la similitud entre los vectores es superior a 0.4; si la línea es punteada quiere decir que la similitud entre dichos dos vectores es entre 0.4 y 0.7 y si la línea es continua quiere decir que la similitud entre los vectores es mayor a 0.7.

En la tabla 5.5 se muestran los resultados de los clusters que resultaron de utilizar la metodología planteada para el desarrollo de la investigación teniendo en cuenta las relaciones interpersonales y los nombres de los lugares, como se mencionó anteriormente son 3 agrupaciones que según las similitudes entre los vectores, también se clasifican según el emperador del que son seguidores o si son neutrales entre ambos:

En la visualización 5.6 ya es posible ver gráficamente las relaciones, ésta vez usando información por locación de la persona y los resultados de la tabla 5.5. Esta vez se evidencian los clusters de la siguiente manera: El cluster 1, quienes son los personajes neutros están representados por el nodo triangular; el cluster 2 corresponde a aquellos que son seguidores de Sutoku representados por el nodo circular; finalmente el cluster 3 corresponde a los seguidores de Goshirakawa representados por el nodo cuadrado.

	Cluster 1	Cluster 2	Cluster 3
Faction of former Emperor Sutoku	1. Nagamori Taira 2. Norinaga Fujiwara 3. Tamenaka Minamoto 4. Tadamasa Taira	7. Naritaka Fujiwara 8. Narimasa Minamoto 9. Tadatsuna Taira 10. Yorikata Minamoto 11. Yorinaka Minamoto 12. Tamenari Minamoto 13. Yorinori Minamoto 14. Tsunenori Fujiwara 15. Tadatsuna Taira 16. Tamemune Minamoto 17. Iehiro Taira 18. Mitsuhiro Taira	21. Yorinaga Fujiwara 22. Chikaharu Minamoto
Faction of Emperor Goshirakawa	5. Suezane Minamoto 6. Shigemori Taira	19. Koreshige Taira 20. Motomori Taira	23. Shinzei 24. Yoshitomo Minamoto 25. Yorimasa Minamoto 26. Tameyoshi Minamoto 27. Saneyoshi Fujiwara 28. Kiyomori Taira 29. Yoshiyasu Minamoto 30. Tadamichi Fujiwara 31. Nobukane Taira

Figura 5.5: Tabla del relaciones entre personas y factores históricos

Al final de la investigación, después de tener los resultados de la misma y los objetivos del proyecto cumplidos, se consulta con conocedores del tema, involucrados en temas de índole humanístico si los resultados son prometedores y expresan conocimiento útil o simplemente la metodología propuesta no es aplicable. Los resultados se evidencian en la tabla 5.7.

Para cada criterio de evaluación presentado, quienes respondieron la *"encuesta"* tienen la posibilidad de responder usando puntajes de 1 a 7, siendo 1 el caso en que estuvieran totalmente en desacuerdo, 4 ni de acuerdo ni en desacuerdo y 7 totalmente de acuerdo. Se concluye entonces de la metodología propuesta lo siguiente:

- Existe credibilidad histórica: La puntuación del método para éste punto supera las expectativas con un 4.50 por lo que se deduce que el método es capaz de mostrar relaciones complejas no evidentes entre personas mencionadas dentro del texto.
- Aprueba ser novedoso ante métodos tradicionales humanísticos: El método es puntuado con 5.25 en cuanto a la novedad de la investigación y el descubrimiento histórico.
- Es posible usarlo en otras investigaciones humanísticas o históricas: La metodología propuesta tiene potencial para poder hacer descubrimientos históricos en investigaciones afines y es puntuado con 5.00

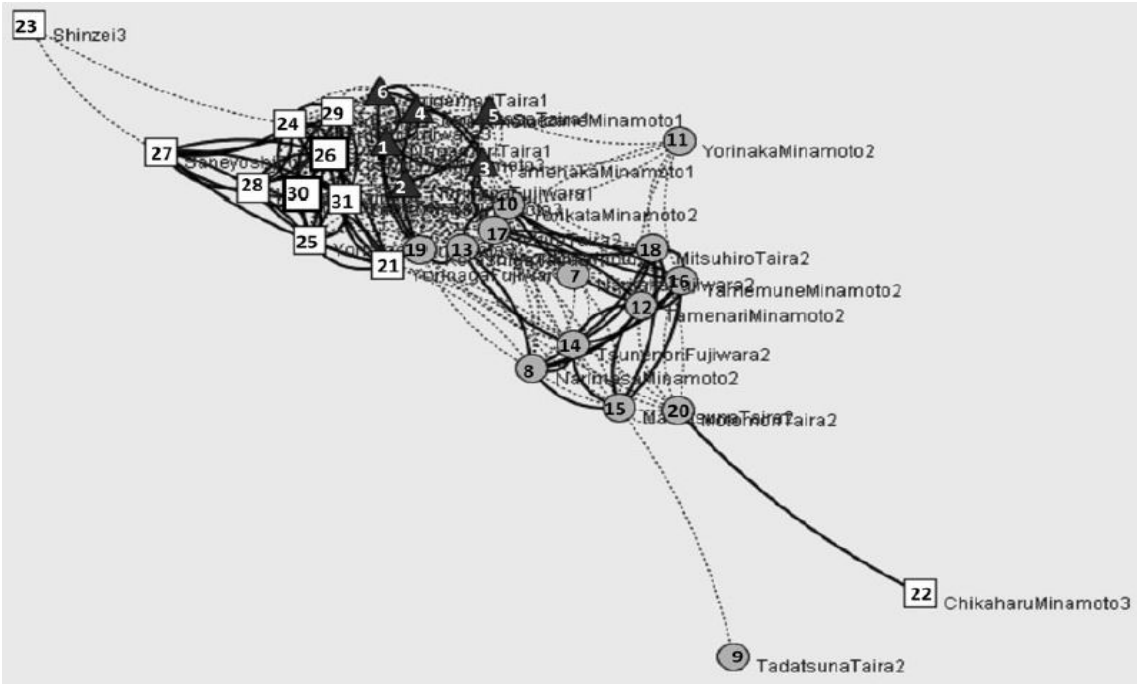


Figura 5.6: Gráfica del resultado del agrupamiento utilizando información por locación

Evaluation criteria	User				Average
	A	B	C	D	
Historical credibility	5	5	4	4	4.50
Novelty for historical research	5	5	5	6	5.25
Availability for historical research	5	5	5	5	5.00
Average	5.00	5.00	4.67	5.00	4.92

Figura 5.7: Tabla del resultado del agrupamiento utilizando información por locación

- Se comprueba la efectividad de usar la locación de personas para hacer agrupamientos, el resultado es mucho más preciso y confiable.

5.3. Information Access to Historical Documents from the Early New High German Period

Actualmente, refiriéndose a literatura y textos que circulan por la web, una gran cantidad pertenece a la categoría de libros y documentos históricos pero aun así, todavía hay una gran cantidad de ellos incluyendo también repositorios de texto, con-

siderados incluso patrimonio cultural mundial, que siguen sin presentarse en formato digital; es por esto que ha surgido la necesidad de realizar esfuerzos solucionar dicho problema y hacer que dicha información sea accesible y pública, no sólo desde las humanidades sino desde otras disciplinas como la informática.

La digitalización de dichos textos históricos es la oportunidad perfecta para poner en práctica técnicas de acceso y manejo de grandes volúmenes de información, como por ejemplo Search and Information Retrieval o Minería de Texto. Hoy en día la colaboración entre las humanidades y la ciencia ha crecido más que nunca y el trabajo se ha visto simplificado gracias a los avances que han surgido para poder presentar y compartir información (tanto homogénea como heterogénea) de investigación y educativa considerada valiosa.

Andreas Hauser, Markus Heller, Elisabeth Leiss, Klaus U. Schulz y Christiane Wanzeck son quienes llevaron a cabo el informe inspirados en el interés de permitir acceso electrónico a miles de documentos y literatura lo que no solo aliviana el trabajo de historiadores, filósofos, lingüistas o Paleólogos sino que también permite compartir conocimiento de forma masiva con todos aquellos interesados y que no necesariamente son expertos. Para la presentación de sus resultados enfocaron sus esfuerzos primeramente, en mostrar las diferentes observaciones que hicieron con respecto al hecho de que muchos de estos textos no están en lenguas modernas (difieren en muchos aspectos del lenguaje moderno) y como eso influye a la hora de realizar técnicas de Search and Information Retrieval o minería de datos y obtener conclusiones. En segundo lugar se enfocaron en demostrar el desarrollo de su trabajo, focalizados en demostrar estrategias que ayudan a relacionar las palabras y el comportamiento del lenguaje moderno con las variaciones correspondientes de las lenguas antiguas. La fuente principal de información sobre la que trabajaron fueron una colección de documentos perteneciente al periodo del Nuevo alto alemán temprano.

El problema principal que surge al querer indexar textos de forma digital es la cantidad de formas o variantes que pudo haber tenido una palabra o incluso un grafema para escribirse según el periodo pues no necesariamente se hizo de la misma manera durante toda la historia. Las técnicas modernas de Search and Information Retrieval y minería de datos hasta ahora están evolucionando para contemplar dichos obstáculos y los trabajos recientes son con los que se ha dado inicio a la resolución de este tipo de inconvenientes. Las siguientes preguntas son claves para empezar a

tratar el problema de manera certera y profunda: [18]

1. *¿Qué tipo de alteraciones o variantes pueden ser identificadas en la ortografía de los distintos lenguajes? ¿Cómo se podrían describir dichas variaciones formalmente?*

2. *¿Cuáles son las consecuencias directas en los distintos campos como lo son Search and Information Retrieval y la minería de texto?*

3. *¿Cómo pueden ser adaptadas las técnicas existentes para hacer frente al problema haciendo uso de textos históricos?*

Los textos históricos generalmente se tratan de documentos que contienen información sobre administración, por ejemplo diarios de gestión judicial o administración de bienes por parte de la iglesia, los mismos se han escrito generalmente con el fin de informar a posibles lectores futuros, dichos textos pueden encontrarse en diferentes formatos, cada uno con sus ventajas y desventajas. Una opción es encontrar los documentos digitalizadas pero por medio de imágenes de los libros originales lo que, aunque si da acceso a los mismos, limita en todo sentido acceder a la información que allí esta contenida. Existe también el caso en el que el texto histórico se encuentra transcrito (generalmente usando una representación de Unicode como UTF-8). La tercera opción es que se encuentre disponible por medio de su estructura y anidaciones, generalmente en formato Extensible Markup Language (XML), Electronic Text Encoding and Interchange (TEI).

Para cumplir con los objetivos y lograr generalizar los beneficios fue necesario para los autores identificar cuatro puntos importantes a tener en cuenta que serían claves para suplir las necesidades de las diferentes categorías de posibles interesados y beneficiados con el avance del proyecto: [18]

- *Los lingüistas estarán interesados en el lenguaje del texto y desearán generar análisis profundos sobre el corpus lingüístico como por ejemplo respecto a las concordancias, esquemas estadísticos de distribución en el texto, etc.*

- *Los Paleólogos estarán interesados en la información sobre las propiedades externas y no textuales de la fuente histórica.*

- *Los historiadores estarán interesados en trabajar directamente sobre el contenido de la fuente, tanto en la fuente original como en la fuente digitalizada, también sobre la información del contexto histórico.*

- *El publico en general puede caer en la misma categoría que el historiador a pesar de que sus consultas no pueden basarse en un enfoque metódico.*

Los requerimientos de cada categoría debían ser atendidas con cuidado, las consultas que se quisieran hacer desde el texto original seguramente no tendrían mayor problema para ser contestadas pero, aquellas consultas que implicaran el contexto a la información o información paleográfica fueron consideradas de un nivel más alto de complejidad que por el momento requerían aún gran cantidad de trabajo manual. El acceso directo a los textos y el cambio de lenguaje implican entonces que las técnicas utilizadas pudieran no dar los mejores resultados ya que se basa en identidad del término de búsqueda o en las co-ocurrencias; como ya se indicó anteriormente las variables para un solo grafema son muchos y eso dificulta obtener resultados con real certeza. No solo para este proyecto sino en los similares, se deben tener en cuenta aspectos como la variación fonética, semántica, entre otros.

Entrando en detalle específicamente acerca del cambio del lenguaje histórico que los autores identificaron sobre los documentos fuente sobre los que se trabajó, destacan el hecho de que el alemán pasó por varias etapas que cronológicamente hicieron que fuera cambiando hasta llegar al punto de que por ejemplo, cada grafema tenga una serie de variantes.

Las siguientes son las etapas cronológicas de desarrollo y transformación por las que ha pasado el alemán:

- Antiguo alto alemán, del ingles Old High German (OHG): aproximadamente hasta c. 1100. Está presente en las rúnicas, glosas y textos religiosos.
- Medio alto alemán, del ingles Middle High German (MHG): aproximadamente hasta 1350 o 1400. Tiempos de poesía cortesana y épica.
- Nuevo alto alemán, temprano del ingles Early New High German (ENHG): aproximadamente hasta c. 1650. Se amplían los horizontes de textos que lo incluían y se evidencian grandes variaciones dialécticas.
- Nuevo alto alemán, del ingles New High German (NHG): aproximadamente desde c. 1650, encuentra sus fundamentos en el ENHG.

A continuación se presentan las variaciones lingüísticas, explicadas por niveles.

1. **Fonológico/gráfico:** Los cambios gráficos y las diferentes variables existentes para los grafemas se incrementa cuando de textos históricos se trata. Las variantes generalmente se tratan de elementos estilísticos.

Grapheme	Variants
<a>	< á, å, ah, aa, ai, ae, å >
<e>	< eh, ee, ei, ey, å, e̊, ä >
<i>	< j, y, ÿ, ie, iee, i͛, ij, ye, ih, jh, ieh, yh >
<o>	< oh, ó, oe, oi, oy, oo >
<u>	< ú, ů, ů, v, w, uh, wh, ůh, uy >
<ä>	< å, e, a, æ, ae, äh >
<ü>	< ů, u, ů, v, ű, ÿ, y, w, ue, üe, üh, uy>
<ö>	< o̊, ó, o, öh, oe, öe, o̊e, œ >

Figura 5.8: Ejemplo Variantes en grafemas vocálicos

2. **Morfológico:**El grado de variabilidad para este nivel es alto, entre más antigua sea la referencia más complejidad se representa y a medida que se desarrolla el lenguaje también se desarrolla la formación del plural.

3. **Léxico:** Los cambios que ha este nivel se refieren son de mucha importancia para el entendimiento del propio documento, los significados de una palabra al día de hoy pueden no ser los mismos significados del pasado, más de un significado para una palabra puede aparecer en el texto. Para el alemán los significados de las palabras cambiaron según la etapa cronológica, por ejemplo, *urlaub* que en inglés significa “vacaciones”, significaba “permiso” en la etapa de OHG y MHD; vuelve a cambiar su significado a “despedida” para la etapa ENHG.

4. **Sintáctica:** Las frases en los textos históricos se comienzan a ver de apoco con cierta extensión considerable debido a que surgen nuevas posibilidades para hacerlo. En cuanto a la sintaxis los cambios no son en tal escala como las extensiones, antes de la época NHG no había un manejo marcado de signos de puntuación.

Se destaca el hecho de que para que apliquen las diferentes variaciones hay que tener en cuenta el tamaño y las condiciones del texto pues puede que maneje periodos diferentes en su contenido o no lo haga o lo haga de manera particular. La ventaja de tener los textos en formato digital según los autores: *“Los corpus de texto digital nos permiten evaluar los procesos de cambio de lenguaje sobre el pilar de una base de datos completa y así recoger todas las variaciones lingüísticas relevantes”* [18].

En el desarrollo del informe y del proyecto, se ratifica de nuevo, el mayor problema a tratar era la variabilidad del lenguaje, al aplicar los conceptos de anidación y los métodos de minería de textos basados en frecuencia se corría con el riesgo de obtener resultados alejados de los que realmente deberían de ser; así mismo al aplicar técnicas de anotación o agrupación. Una posible solución contemplada era el *hacer uso de Recuperación de información (Information Retrieval) (IR) específicamente orientado a los textos históricos*, para poder llevar a cabo las diferentes computaciones y algoritmia que permitiría trabajar el lenguaje antiguo los investigadores contemplaron posibles soluciones, la primera posibilidad que surgió era la creación de diccionarios especializados que incluyeran cada palabra moderna con sus respectivas variantes encontradas (incluyendo información valiosa como tiempo o lugar), la principal desventaja de esta posibilidad era la gran cantidad de tiempo que esto podría consumir junto con el hecho de que se obtendría un diccionario estático no totalmente compatible con textos diferentes a la fuente definida para este caso.

Como segunda solución se contemplaba el *generar relaciones basadas en reglas*. Las diferencias entre el lenguaje moderno y antiguo se le presentarían a la máquina por medio de un conjunto de reglas. La posibilidad online estimaba el proceso de incluir una palabra para así generar posibles variantes de búsqueda; con la posibilidad offline se intentaría normalizar las variantes históricas e indexarlas aplicando las reglas de manera inversa.

Como tercera solución los autores proponen el *encontrar relaciones basados en similitud de palabras*, se especifica un formato con las similitudes entre las palabras. Se presenta una palabra y como salida se mostraría un conjunto de resultados que cumplen similitud suficiente con dicha palabra consultada.

Entrando en detalle sobre el desarrollo del informe, los autores especifican que su foco de interés era la adquisición de impresiones del siglo XIV - XVII ya que son bastante numerosos y se podría obtener una base de datos robusta. Si se conseguían textos de este tipo, se accedería a crónicas, sermones o documentos legales con posibilidad de digitalización automática. El proceso a seguir para llevar a cabo la investigación fue: [18]

1. *Crear manualmente un pequeño corpus*
2. *Manejar la ortografía y las variaciones compuestas*
3. *Crear un diccionario electrónico utilizable*

4. *Incorporar morfología y sintaxis (del lenguaje)*
5. *incorporar la estructura del documento y metadatos*
6. *Utilizar todo lo anterior para mejorar el Optical Character Recognition (OCR) y digitalizar más textos*

Resulta entonces como fuente principal de la investigación el siguiente corpus: *"De una selección de 23 textos de la época del Nuevo Alto Alemán, once han sido digitalizados. Cuatro de éstos han sido etiquetados para incluir información sobre la categoría, la nueva traducción del Alto Alemán. Los 11 textos representan un total de unas 18.000 líneas y 130.000 palabras (tokens)"* [18]. Se crea un diccionario en donde se recolectaron las correspondencias entre las antiguas y nuevas variantes en el texto *Dyll Vlnspiegel* y como entregable destacable los investigadores han recopilado una lista de sucesos que explican las correspondencias entre las palabras del alemán moderno y el antiguo: [18]

- *Formación de una nueva palabra*
- *Palabras en latín*
- *Variaciones en la separación por sílabas*
- *Formación parcial de una nueva palabra*
- *Variación en prefijos y sufijos*
- *Variaciones en la tipografía*
- *Variaciones fonéticas*
- *Nuevos caracteres*

Con el fin de obtener mejores resultados en la investigación se empezó a diseñar una estrategia para optimizar la precisión al obtener coincidencias pretendiendo incluir formaciones de nuevas palabras, palabras en latín y variaciones en la separación por sílabas de las palabras. Lo anterior llevaría a una variación en lo ya trabajado llevando a que algunos resultados sean ilegibles por el sistema desarrollado por lo que se pasaría a manejar un conjunto de reglas de adaptación para tratar específicamente las siguientes variaciones: Formación parcial de una nueva palabra, variaciones en la separación por sílabas, nuevos caracteres, variaciones en prefijos y sufijos y variaciones en la tipografía.

Para la construcción del diccionario mencionado anteriormente se ha realizado una base de datos que apoyada en Structured Query Language (SQL) se centra en proporcionar estadísticas de co-ocurrencias de las palabras en el texto, presentar variables ortográficas y coincidencias sobre documentos con estadísticas también.

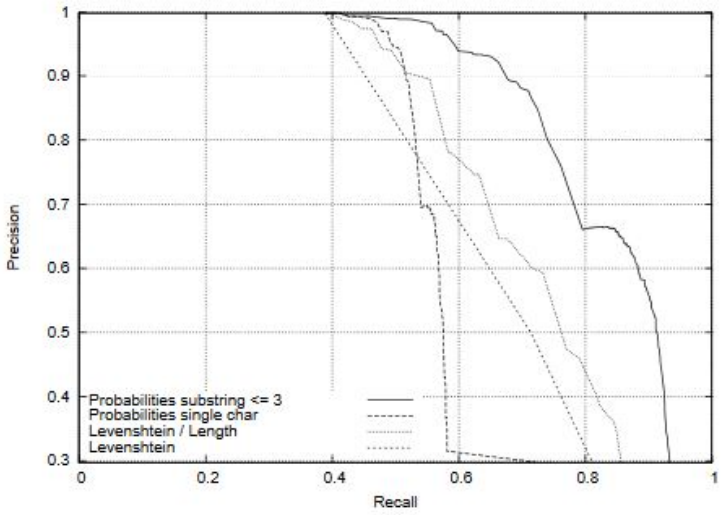


Figura 5.9: Resultados del sistema al encontrar relaciones

Como principales conclusiones de la investigación se encuentra el hecho de que efectivamente las variaciones en los lenguajes modernos y antiguos pueden influir negativamente en los resultados de los sistemas construidos a partir de Recuperación de información (Information Retrieval) (IR) o Minería de texto, modificando los resultados a resultados inciertos. En cuanto a la digitalización de textos históricos, también se presentan ciertos inconvenientes, por ejemplo, haciendo uso de Optical Character Recognition (OCR) se obtienen resultados muy deficientes debido a que la caligrafía de lo textos históricos no es la misma que se ve actualmente, es mucho más compleja y con sin fin de variantes visuales.

En cuanto al uso de Optical Character Recognition (OCR) se destaca como principal inconveniente el que los software identifican mal las letras lo que tergiversa los resultados, los investigadores en ese entonces trataban de diseñar una medida de confianza para definir si el caracter que reconoció el software es el correcto o no.

Otra conclusión importante fue que los documentos históricos electrónicos estaban generalmente basados en formato Extensible Markup Language y era un poco complicado hacer búsquedas “inteligentes” a partir de las entradas o palabras al

sistema, para lo anterior los investigadores pretendían desarrollar un marco de trabajo multiplataforma que lograra soportar las consultas Extensible Markup Language y lograr resultados desde la perspectiva de la indexación y construcción de relaciones directas.

5.4. Information Retrieval from Historical Corpora

Actualmente el numero de documentos antiguos digitalizados están aumentando de una manera significativa, estos documentos difieren a los actuales de tres maneras, el vocabulario y la ortografía han cambiado y en algunos casos era inconsistente, los objetivos de esta investigación fueron identificar los cuellos de botella de la recuperación de la información de los corpus históricos y encontrar soluciones [20].

En este trabajo se realizo un experimento para encontrar las diferencias entre textos antiguos y modernos, el corpus utilizado fue una colección de textos legales holandeses y belgas conocidos como Antwerpse Compilatae y los Gelders Land en Stadsrecht, que datan de los siglos XVI y XVII. Los textos se dividieron en 393 documentos, con un total de 371862 palabras. Se pusieron a disposición en forma electrónica mediante reconocimiento óptico de caracteres (OCR) y en parte mediante introducción manual de datos. [20]. Luego se le pidió a 6 expertos en el campo que formularan cada uno 5 necesidades de información, éstas necesidades se convertirían en consultas y se consultarían en un sistema de IR(Interpretación simultanea) de modelo vectorial básico, también se les pidió a cada uno que señalaran que documentación podría ser útil para cada una de las consultas, dentro de la inspección de los documentos que no fueron recuperados por el sistema, mostraron dos características, la primera los documentos contienen uno o más de los términos de la consulta pero con ortografía diferente y la segunda los documentos no contienen los términos de la consulta específicos, sino similares, dentro de la primera característica sucede porque el modelo vectorial solo determina relevante si dentro del documento se encuentran los términos exactos de la consulta, entonces se entra a evaluar cada termino teniendo en cuenta que aunque tenga ortografía difiera, tengan una relación permitida, es aquí donde se encuentra un cuello de botella y unos de los problemas mas complejos en la digitalización de documentos antiguos, al ser la ortografía inconsistente y al aplicar técnicas de OCR la magnitud de los errores va a ser mayor que con los documentos modernos. Como se puede apreciar en la siguiente imagen las variaciones de las palabras y los sinónimos es mas alta en documentos antiguos que en documentos

modernos.

	Historical texts		Modern texts	
Word	var	syn	var	syn
advocaat	4	9	0	4
arbitrages	2	1	0	0
borgtocht	9	0	0	4
dadingen	4	0	0	3
gehuwd	3	2	0	1
notaris	2	4	0	0
panding	1	0	0	1
persoon	0	2	0	0
procederen	2	1	0	1
procureur	1	1	0	0
rechter	5	0	0	1
rente	4	5	0	1
secretaris	2	5	0	0
termijn	3	0	0	0
verbintenissen	6	0	0	0
vonnis	7	2	0	2
vrouw	1	2	0	10

Figura 5.10: Variantes y sinónimos encontrados en documentos históricos y en documentos modernos

La segunda característica se da por la incapacidad del modelo de reconocer los sinónimos de una consulta, al no cumplir las condiciones del sistema, los documentos que contienen uno o más sinónimos de un término de consulta, pero no el término de consulta exacta, se consideran incorrectos, en la recuperación de los documentos modernos también se da el mismo problema aunque en ésta muestra se da la particularidad que los términos de los documentos históricos tiene mas sinónimos que los modernos. A partir de los resultados de la inspección de los documentos, concluimos que hay dos principales cuellos de botella en la recuperación de información con textos históricos [20]:

- *Ortografía variable e incoherente: el cuello de botella de ortografía*
- *Uso de un vocabulario diferente: el cuello de botella del vocabulario*

Para solucionar el problema de botella del vocabulario se hace uso de un tesauro, pero esta solución no se implemento porque no existía un tesauro adecuado para el corpus de texto utilizado en esta investigación y también fue difícil su construcción por el limitado tiempo de los expertos, entonces se decidió concentrarse en la ortografía cambiada e incoherente. Otra posible solución era hacer uso de heurística de transformación para transformar las formas de las palabras viejas en modernas, las heurísticas se basaron fuertemente en las reglas aplicadas por el algoritmo holandés

stemming. Este algoritmo transforma las palabras en sus raíces basándose en sus sufijos, prefijos e infijo. La antigua forma de la palabra «uuytricht-inghe» tiene, por ejemplo, el sufijo «inghe», el prefijo «uuyt» y el infijo «richt». Sin embargo, los sufijos, prefijos e infijos de formas de palabras antiguas podrían no coincidir exactamente con los requeridos por el algoritmo de derivación. Por lo tanto, el algoritmo no funcionará en formas de palabras antiguas. Sin embargo, utilizando heurísticas, las formas de palabras antiguas pueden ser transformadas de tal manera que sus sufijos, prefijos e infijos puedan ser combinados y transformados por el algoritmo de derivación. La antigua forma de la palabra «uuytrichtinghe» puede, por ejemplo, transformarse en «uitrichtingen» [20].

Se desarrollaron tres tipos de heurística, heurística para la transformación del sufijo, heurística para la transformación del prefijo, y heurística para la transformación de infijo. La razón para el desarrollo de tres conjuntos específicos de heurísticas es que la transformación de una combinación específica de caracteres depende de si esta combinación de caracteres es un sufijo, un prefijo o un infijo. Un ejemplo es la heurística para la transformación del sufijo $ff \rightarrow f$, lo que indica que los caracteres 'ff' tienen que ser transformados en 'f' cuando forman el sufijo de una palabra. Sin embargo, si forman el prefijo de un infijo de una palabra, no siempre se pueden transformar [20].

Para poder elegir el mejor de los diferentes procedimientos de combinación (n-gram matching, stemming, programación dinámica y heurística de transformación) se llevó a cabo un experimento. Basándonos en los resultados de Robertson y Willett elegimos n = 3 para la adaptación de n-gramos y como enfoque de programación dinámica particular el algoritmo de Wagner-Fischer. La configuración de este experimento fue la siguiente. Todos los términos de consulta se extrajeron de las consultas de cada experto (ver sección 1). Para cada uno de estos términos de consulta, las variantes del término fueron recuperadas de la lista completa de términos de documento extraídos de los documentos de la colección bajo consideración. La recuperación de las variantes de cada término de la consulta se hizo una vez por cada método a ser probado. Para cada método, esto resultó en una lista de términos de documento para cada término de consulta. Si un término de documento era en realidad una variante del término de la consulta, se calificó como pertinente. Si no lo era, se denotaba como no pertinente. El anterior procedimiento se realizo en varias combinaciones de las cuatro posibles, dando como resultado 8 tipos de combinaciones posibles donde para cada uno se hizo un recuadro comparativo de desempeño.

Retrieval method	Comp. recall	Precision
Tri	70.4	57.9
Tri – stem	74.0	62.5
Tri – prepro	74.8	53.7
Tri – stem – prepro	82.1	57.8
Wag	67.2	12.1
Wag – stem	70.6	13.2
Wag – prepro	73.8	11.9
Wag – stem – prepro	77.4	12.1

Legend: Tri: Trigram matching, Wag: Wagner-Fischer, Stem: stemming, Prepro: preprocessing

Figura 5.11: Resultados de la evaluación de desempeño

Como conclusiones finales, se demostró que existen dos problemas de cuello de botella de ortografía y de vocabulario para docuentos que datan de los siglos XVI y XVII, para el cuello de botella de ortografía se uso procedimientos de combinación de trigramas, derivación, pre-procesamiento y programación dinámica, el pre-procesamiento consistió en aplicar heurísticas para mapear las viejas formas de las palabras a las modernas, para hacer que las palabras fueran adecuadas para derivar un stemmer moderno. Posteriormente, este nuevo sistema se comparó con un sistema basado en vectores estándar en una tarea de recuperación de información, el rendimiento del nuevo sistema fue significativamente mejor. Vale la pena resaltar que este sistema solo aplica para textos holandeses, lo que hace que no funcione para otros idiomas, aunque se puede investigar en qué medida otros idiomas pueden beneficiarse de este enfoque.

5.5. Topic Modeling on Historical Newspapers

Dentro del campo de minería de texto histórica, por lo general la fuente de información proviene de texto antiguos que pertenecían a la iglesia o que eran de difícil acceso, dentro de este proyecto en su primera fase, se dio la particularidad de usar periódicos antiguos con el fin de ayudar a la investigación histórica. Para los investigadores históricos una de las fuentes de información mas interesantes son los periódicos, para este proyecto se rescataron periódicos que constan del siglo XIX al siglo XX en Estados Unidos, cabe resaltar la gran importancia de estos, ya que dentro del contenido de estos los norteamericanos debatían toda clase de temas que hacían parte de la sociedad y la vida cotidiana, un aspecto de gran importancia de los periódicos es que documentan todas las experiencias humanas que ninguna otra

fuente de información puede tener. A pesar de su gran valor, se puede considerar que los periódicos antiguos son una mina de oro de información que no ha sido explotada y que hasta ahora esta siendo tenida en cuenta en trabajos de esta índole, pero no ha sido lo suficientemente explotada por la sencilla razón que son volúmenes de información inimaginables y que a la final le va a dificultar mucho al investigador recorrer semejante volúmenes de información.

A pesar de la gran cantidad de información que no se esta utilizando, iniciativas de The National Endowment for the Humanities (NEH) y The Library of Congress (LOC) están patrocinando proyectos para digitalizar todos los periódicos sobrevivientes de los Estados Unidos desde 1836 hasta el presente, la finalidad de este proyecto es poder ofrecer una gran cantidad de periódicos históricos en formato digital. Actualmente las herramientas de búsqueda se vuelven mas eficientes para los investigadores históricos, sin embargo el problema permanece hay, la gran cantidad de información sigue siendo enorme y esta en constante crecimiento, entonces este proyecto como objetivo principal busca identificar los temas mas importantes durante un periodo de tiempo determinado, la idea es agrupar varios temas en conjuntos con patrones inusuales, lo cual puede llevar a investigaciones nunca antes hechas y a revelar información nunca antes vista. El modelado de temas se puede aplicar para un año en particular o para un periodo de años. [21]

El modelo se basa en dos modelos, el primero un modelo de análisis semántico latente probabilístico (pLSA) y el segundo un modelo latente de asignación de Dirichlet (LDA), ambos modelos probabilísticos consideran cada documento como una mezcla de temas. Los modelos descomponen la colección de documentos en grupos de palabras que representan los temas principales. También se uso la herramienta MALLET que es un paquete basado en Java para el procesamiento de lenguaje natural estadístico, clasificación de documentos, agrupación, modelado de temas, extracción de información y otras aplicaciones de aprendizaje de máquinas a texto. [22]

Para aplicar el modelo de temas se escogió una colección de periódicos históricos publicados en Texas desde 1829 hasta el 2008, dentro de la colección habían 232.567 de paginas que pertenecían a 180 años. Para dividir los años se separaron en periodos o en años en específicos por características propias, el primer conjunto de datos fue periódicos de 1865 a 1901 porque en esta época los texanos buscaron reconstruir su economía de posguerra invirtiendo fuertemente en la producción de algodón en

todo el estado. El algodón se consideró una inversión segura, por lo que los texanos produjeron lo suficiente durante este período para convertir a Texas en el mayor productor de algodón de los Estados Unidos en 1901. Sin embargo, la sobreproducción durante ese mismo período empobreció a los agricultores de Texas al bajar el precio del algodón y por lo tanto un gran porcentaje se declaró en bancarrota y perdió sus tierras. Como resultado, los algodoneros enojados en Texas durante la década de 1890 se unieron a un nuevo partido político, los Populistas, cuyo objetivo era utilizar al gobierno nacional para mejorar las condiciones económicas de los agricultores. Este esfuerzo fracasó en 1896, aunque representó una de las mayores revueltas políticas de terceros en la historia de los Estados Unidos. Entonces todos estos temas se esperan que sean mencionados en un conjunto de 52.555 páginas en más de 5.902 ejemplares. [21]

El segundo conjunto fueron todos los periódicos de 1892 este fue el año de la formación del Partido Populista, que una gran parte de los campesinos de Texas se unió para las elecciones presidenciales de 1892. Los Populistas buscaron que el gobierno federal estadounidense se involucrara activamente en la regulación de la economía para evitar que los productores de algodón se endeuden más. Este conjunto de datos consta de 1.303 páginas sobre 223 ejemplares.

El tercer conjunto fueron los todos los periódicos de 1893 ya que una depresión económica importante golpeó los Estados Unidos en este año, devastando la economía en cada estado, incluyendo Texas. Esto perjudico el problema del algodón dentro de la economía de los estados y aumentó los esfuerzos de los populistas dentro de Texas para impulsar reformas políticas importantes para abordar estos problemas. Lo que vemos en 1893, entonces, es un gran estrés que debería incrementar las tendencias dentro de la sociedad texana de ese año. Este conjunto de datos consta de 3.490 páginas en más de 494 ejemplares.

El cuarto conjunto de datos fue periódicos de 1929 a 1930, estos años representaron el inicio y el comienzo de la Gran Depresión. La economía de los Estados Unidos comenzó a colapsar en Octubre de 1929, cuando el mercado bursátil se estrelló y comenzó una serie de fracasos económicos que pronto derrumbaron casi toda la economía de los Estados Unidos. Texas, con su ya débil dependencia económica del algodón, estaba tan devastada como cualquier otro estado. Como tal, este período estuvo marcado por discusiones acerca de cómo salvar tanto la economía del algodón de Texas como la posible intervención del gobierno en la economía para prevenir

la catástrofe. Este conjunto de datos consta de 6.590 páginas sobre 973 ejemplares. [21]

Cabe resaltar que para muchos investigadores y expertos en la historia de los Estados Unidos, la economía y el algodón eran temas de gran importancia, además con eso fue el ascenso y la caída del Partido Populista durante la década de 1890, cuando los agricultores trataron de usar el sistema político como un medio para enfrentar sus problemas económicos. Entonces se espera encontrar todos estos temas como dominantes en los periódicos de ese periodo de tiempo.

Antes de aplicar el modelo de temas sobre la información de los periodos de los periódicos, se hizo un proceso de pre-procesamiento de la información la cual tenia algunos inconvenientes, uno de ellos fue la falta de puntuación y separación de los artículos de una misma página, para empezar se hizo una corrección ortográfica usando un diccionario de Aspell y de-hyphenates para identificar las palabras separadas por guiones, luego de tener la ortografía corregida se procede a utilizar el Stanford Named Entity Recognizer el cual clasifica las palabras en cuatro grupos, person(Persona), organization(Organización), location(Lugar) y Miscellaneous(Varios), el software usado para el desarrollo del proyecto proporciona esta funcionalidad, luego las palabras que no entraban en estos grupos, se pasaron por un stemmer en ingles el cual toma la raíz de la palabra. Omitir alguno de los pasos anteriores tenia un gran impacto en el resultado por lo cual se llego a la conclusión de pasar la información varias veces por el modelo por cada una de las fases para obtener mejores resultados. [21]

Los resultados arrojados por el modelo fueron validados por un historiador experto en la historia de Texas y el resultado fue muy satisfactorio dentro de los temas, el lenguaje orientado al mercado es el tema dominante, lo que es exactamente lo que el experto esperaba como historiador de esta región. Se puede ver, por ejemplo, que gran parte de la economía del algodón estaba orientada a abastecer a los molinos industriales de Inglaterra. La palabra "Liverpool" que es el nombre del puerto inglés a donde fue enviado el algodón de Texas, aparece bastante a través de las muestras. Estos resultados sugieren un alto grado de precisión en la identificación de temas dominantes e importantes en el corpus. Como conclusiones finales se puede usar el modelo de temas como una herramienta muy poderosa para obtener los temas mas importantes dentro de un volumen de datos muy grande, pero con el inconveniente de no saber cuales de estos temas son importantes para el investigador en este

caso todos los referentes a Texas durante los periodos establecidos, es decir solo el profesional puede saber cual o cuales son los temas relevantes y cuales no, reduciendo la cantidad de información a procesar y haciendo que no se requieran técnicas de pre-procesamiento de la información tan robustos.

Capítulo 6

Herramientas para minería de texto

6.1. KNIME Analytics Platform

Konstanz Information Miner (KNIME) es una plataforma de análisis, generación de informes e integración de datos de código abierto, integra diversos componentes para Machine Learning y minería de datos a través de su concepto de pipelining modular de datos [6.1]. Una interfaz gráfica de usuario permite el montaje de nodos para el procesamiento de datos (ETL: Extracción, Transformación y Carga), para el modelado, análisis y visualización de datos. Con más de 1000 módulos, cientos de ejemplos listos para ejecutar, una amplia gama de herramientas integradas y la más amplia selección de algoritmos avanzados disponibles.[23]

Figura 6.1: Pipelining modular de datos

Una de las ventajas que posee KNIME es la gran cantidad de nodos nativos y las contribuciones de una enorme comunidad mundial de científicos de datos, también permite trabajar con varios tipos de datos como archivos de texto simples, bases de datos, documentos, imágenes, redes e incluso datos basados en Hadoop pueden combinarse dentro del mismo flujo, su amigable interfaz gráfica fácil de aprender significa que la codificación es opcional. KNIME está escrito en Java y está basado en Eclipse y hace uso de sus métodos de extensión para soportar plugins

proporcionando así una funcionalidad adicional.

KNIME cuenta con soporte para una gran cantidad de tipos de datos como Extensible Markup Language (XML), JSON e imágenes, posee funciones matemáticas y estadísticas, algoritmos avanzados de aprendizaje predictivo y de máquina, control de flujo de trabajo, combinación de herramientas para Python, R, SQL, Java, Weka, realizar vistas e informes interactivos de datos, los flujos de trabajo KNIME se pueden utilizar como conjuntos de datos para crear plantillas de informes que se pueden exportar a formatos de documento como doc, ppt, xls, pdf y otros.

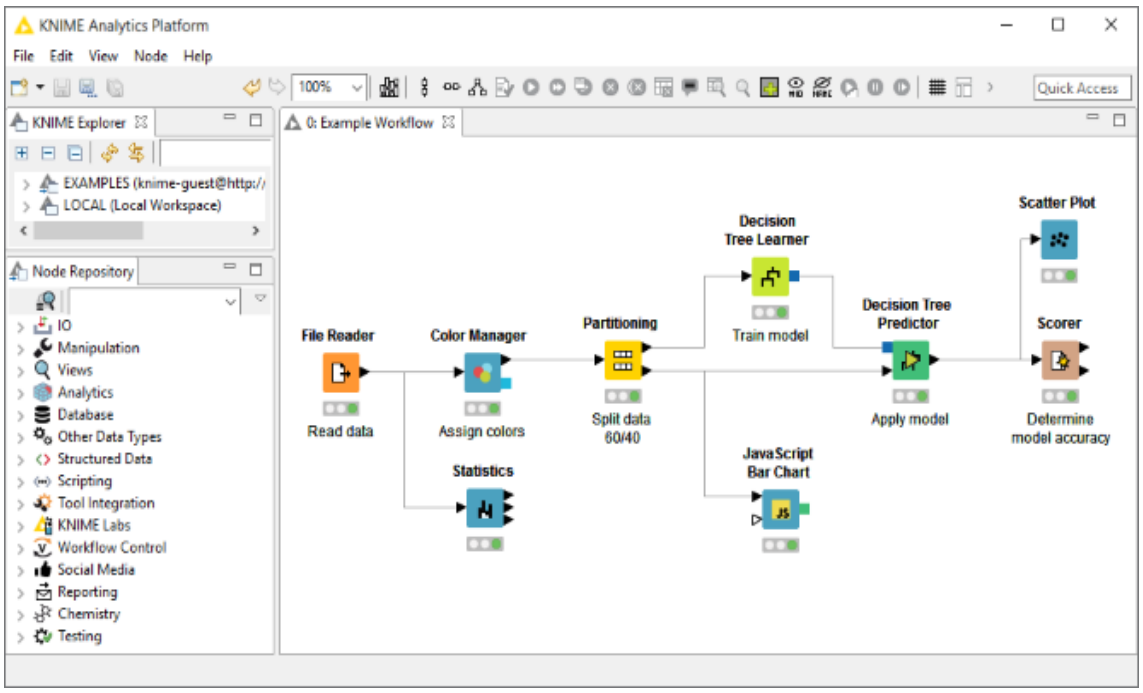


Figura 6.2: KNIME Analytics Platform

Por ser una plataforma intuitiva de código abierto, fácil de usar, tener la capacidad de crear de forma visual flujos o tuberías de datos, ejecutar selectivamente algunos o todos los pasos de análisis, modelar y generar vistas interactivas fue escogida para la realización de este proyecto, ya que posee las herramientas y complementos necesarios para la finalización del mismo.

6.2. RapidMiner

RapidMiner es una plataforma de software que proporciona un entorno integrado para el aprendizaje automático, minería de datos, minería de texto, análisis negocios

y análisis de predictivo. Se utiliza para aplicaciones empresariales y comerciales, así como para la investigación, la educación, el entrenamiento, el prototipado rápido y el desarrollo de aplicaciones y soporta todas las etapas del proceso de minería de datos, incluyendo la preparación de datos, la visualización de resultados, la validación y la optimización. RapidMiner se desarrolla en un modelo de núcleo abierto.[24]

RapidMiner utiliza un modelo cliente/servidor con el servidor ofrecido como (SaaS)Software as a Service o en infraestructuras en la nube. RapidMiner proporciona el 99 % de una solución analítica avanzada a través de marcos basados en plantillas que aceleran la entrega y reducen los errores eliminando casi la necesidad de escribir código. RapidMiner ofrece procedimientos de minería de datos y aprendizaje de máquinas, incluyendo: carga y transformación de datos (extracción, transformación, carga (ETL)), procesamiento y visualización de datos, análisis predictivo y modelado estadístico, evaluación e implementación. RapidMiner está escrito en el lenguaje de programación Java. RapidMiner proporciona una GUI para diseñar y ejecutar flujos de trabajo analíticos. Estos flujos de trabajo se llaman procesos en RapidMiner y consisten en múltiples operadores. Cada operador realiza una sola tarea dentro del proceso, y la salida de cada operador forma la entrada de la siguiente. Alternativamente, el motor puede ser llamado desde otros programas o utilizado como API. Las funciones individuales se pueden llamar desde la línea de comandos. RapidMiner proporciona esquemas de aprendizaje, modelos y algoritmos y puede extenderse utilizando scripts R y Python. [25]

6.3. R-Programming y R-Studio

R es un lenguaje y un entorno para la informática estadística y los gráficos. R proporciona una amplia variedad de modelos estadísticos como modelos lineales y no lineales, pruebas estadísticas clásicas, análisis de series de tiempo, clasificación, agrupación y técnicas gráficas. R es una implementación de software libre del lenguaje S pero con soporte de alcance estático. Se trata de uno de los lenguajes más utilizados en investigación por la comunidad estadística, siendo además muy popular en el campo de la minería de datos, la investigación biomédica, la bioinformática y las matemáticas financieras. Una de las fortalezas de R es la facilidad con la que se pueden producir parcelas de calidad de publicación bien diseñadas, incluyendo símbolos matemáticos y fórmulas donde sea necesario. Se ha prestado gran atención a los valores predeterminados de las opciones de diseño menores en gráficos, pero el usuario conserva el control total. R está disponible como Software Libre bajo

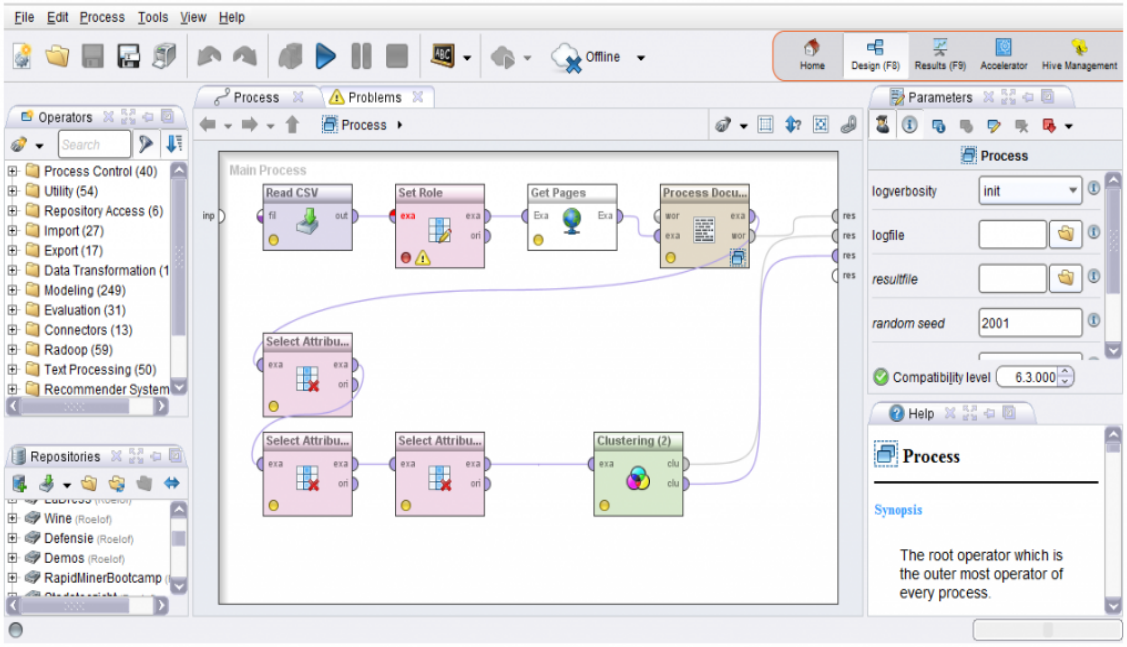


Figura 6.3: RapidMiner Studio

los términos de la GNU General Public License de la Free Software Foundation en forma de código fuente. Compila y ejecuta en una amplia variedad de plataformas UNIX y sistemas similares (incluyendo FreeBSD y Linux), Windows y MacOS. [26]

RStudio es un entorno de desarrollo integrado (IDE) libre y de código abierto para R. Uno de los aspectos más poderosos del uso de R es que se puede descargar paquetes gratuitos para varios tipos de análisis. El análisis de texto todavía se encuentra en etapa de desarrollo, pero es muy prometedor.

6.4. Weka

Weka es un software de código abierto emitido bajo la GNU General Public License. Contiene una colección de algoritmos de aprendizaje automático para tareas de minería de datos y interfaces gráficas de usuario para facilitar el acceso a estas funciones.

Los algoritmos pueden ser aplicados directamente a un conjunto de datos o llamados desde su propio código Java. Weka contiene herramientas para pre-procesamiento de datos, clasificación, regresión, clustering, reglas de asociación y visualización. También es adecuado para desarrollar nuevos esquemas de aprendizaje de máquinas.[27]

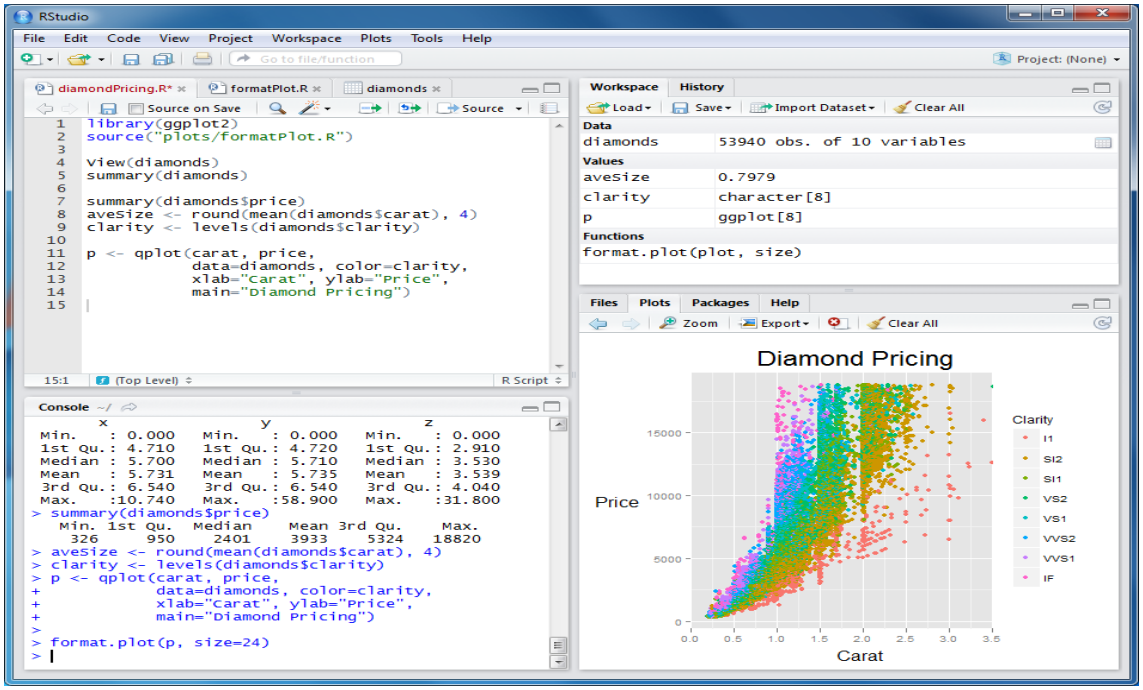


Figura 6.4: RStudio

Algunas de las ventajas de usar Weka es la portabilidad, ya que se implementa completamente en el lenguaje de programación Java y por lo tanto se ejecuta en casi cualquier plataforma de computación moderna. Posee una colección completa de preprocesamiento de datos y técnicas de modelado. La facilidad de uso gracias a sus interfaces gráficas de usuario. Weka soporta varias tareas estándar de minería de datos, más específicamente, pre-procesamiento de datos, agrupación, clasificación, regresión, visualización y selección de características. Todas las técnicas de Weka están basadas en la suposición de que los datos están disponibles como un archivo plano o relación, donde cada punto de datos se describe por un número fijo de atributos (normalmente, atributos numéricos o nominales, pero también se admiten otros tipos de atributos) . Weka proporciona acceso a las bases de datos SQL utilizando Java Database Connectivity y puede procesar el resultado devuelto por una consulta de base de datos. No es capaz de hacer minería de datos multi-relacional, pero hay software separado para convertir una colección de tablas de base de datos enlazadas en una sola tabla que es adecuada para el procesamiento utilizando Weka.

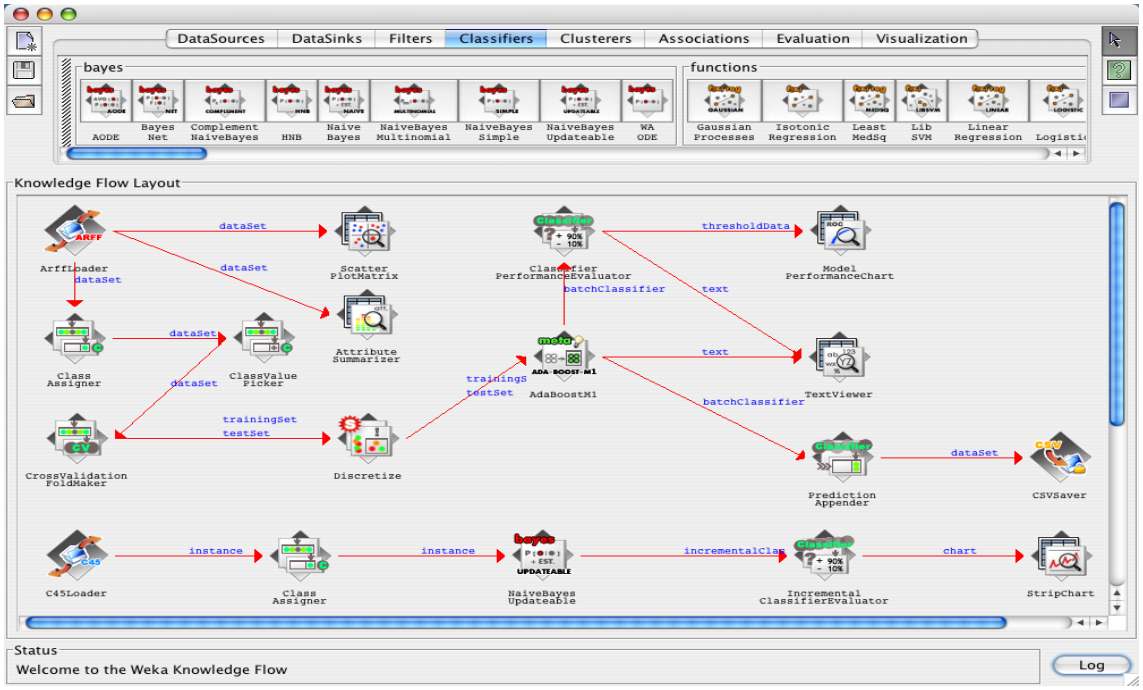


Figura 6.5: Weka Knowledge Flow

Capítulo 7

Proceso del trabajo

7.1. Metodología de trabajo

Existen muchos modelos para el desarrollo de proyectos de Minería de datos tales como SEMMA (Sample, Explore, Modify, Model, Assess), DMAMC (Definir, Medir, Analizar, Mejorar, Controlar) o CRISP-DM (Cross Industry Standard Process for Data Mining), pero el mas utilizado en ambientes académicos o industriales es CRISP-DM, para este proyecto se utilizo varias fases de ésta metodología que ayudaron en la realización del mismo, excluyendo el caso de negocio, ya que se encuentra en un contexto académico.

La metodología CRISP-DM es un enfoque multinacional basado en estándares para describir, documentar y mejorar continuamente los procesos de minería de datos y los procesos asociados al almacenamiento de datos e inteligencia empresarial. El marco de trabajo identifica seis pasos en el proceso de minería de datos, como se muestra en la siguiente imagen. Se añadió un séptimo paso de rendimiento para la mejora continua de procesos a través de un plan sucesivo que se compone en analizar, implementar y realizar iteraciones a los proyectos de minería de datos.[11]

1. **Fase 1 : Determinar el propósito del estudio**
   Como en cualquier proyecto, se define el problema a resolver, lo cual requiere entender el caso de negocio, identificar la oportunidad, interactuar con los implicados para entender a fondo la situación, establecer los objetivos y metas de la minería de texto y generar el plan del proyecto.

2. **Fase 2 : Compresión de la naturaleza de los datos**

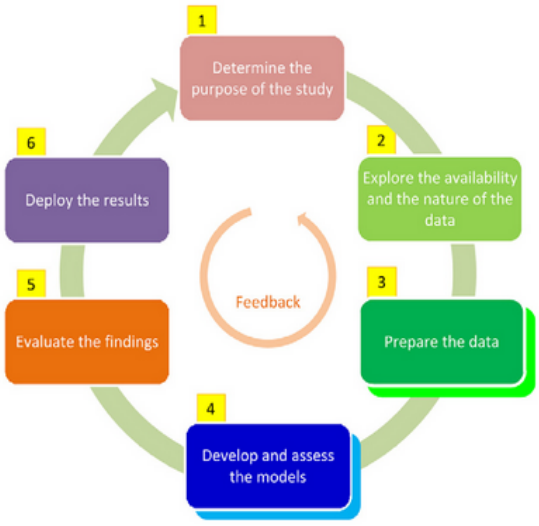


Figura 7.1: Diagrama CRIPS-DM

Una vez el propósito del estudio esta claro se debe evaluar la disponibilidad, accesibilidad y aplicabilidad de los datos necesarios para la realización del estudio, para esto se debe tener en cuenta de donde provienen los datos, si son internos o externos de la empresa y el tipo de datos por ejemplo si son digitales o físicos.

3. **Fase 3 : Preparación de los datos**
   La tercera fase se divide en cuatro fases:

   - Establecer el Corpus: El propósito de esta actividad es recolectar todos los documentos que son relevantes para el contexto del problema, dos atributos a tener en cuenta de los datos son la calidad y la cantidad. Se pueden utilizar técnicas manuales o automatizadas para la recolección de la información y ésta información pueden ser archivos como PDF, tablas de Excel, documentos de Word, presentaciones de PowerPoint, emails, web posts, archivos HTML o notas. Una vez recopilada toda la información, se deben organizar y transformar si es el caso para ser procesados por computador, la organización de los datos puede ser tan simple como guardarlos en una carpeta.
   - Limpieza de los datos: La probabilidad de que los datos del corpus estén sin errores es baja, por eso se debe realizar un trabajo de limpieza el cual puede incluir reemplazar valores, realizar correcciones ortográficas, reemplazar campos vacíos por nulos, inclusive se pueden incluir técnicas de reemplazo sofisticadas. Al finalizar el trabajo de limpieza debería existir un documento

donde se detalla todos los cambios realizados y los problemas de calidad que se identificaron en los datos, esté documento sirve para medir el impacto en los resultados de los cambios que se realizaron durante la limpieza de datos.

- Construcción de datos: En esta actividad se debe documentar si se realizan modificaciones a la información para llevar acabo cálculos, por ejemplo si se agregaron columnas o nuevos registros.
- Integración de los datos: Por lo general en este punto la información proviene de varias fuentes, la idea de esta actividad es hacer la integración de la información para tener una sola fuente.

4. **Fase 4 : Desarrollar y evaluar los modelos**
   Una vez se tiene el corpus listo ya se pueden identificar patrones en la información, para esto se debe seleccionar la técnica de modelado a utilizar, diseñar las pruebas, construir el modelo y por ultimo evaluar el modelo utilizado. Existen muchas técnicas de modelado pero no todos se adaptan a las necesidades del problema, se debe escoger en función de los tipos de variables involucradas. Al final se debe tener especificado el modelo que se usara.

5. **Fase 5 : Evaluación de los datos obtenidos**
   Después de explorar los datos y encontrar patrones en la información, se debe evaluar la calidad de los resultados, no solo se evalúan los modelos aplicados sino también los procesos que se usaron para crearlos, como tarea de esta fase se debe evaluar el valor de los modelos usados para resolver el problema planteado para el caso de uso. Luego revisar a profundidad el modelo utilizado, para detectar posibles problemas antes de salir a producción y también para revisar como mejorar el proceso para posibles proyectos en el futuro. Esta fase concluye con la revisión completa del modelo listo para desplegarlo.

6. **Fase 6 : Despliegue**
   Cuando el modelo esta listo para ser usado, se debe especificar una estrategia para desplegarlo en función del negocio, dento de la estrategia se debe definir el resumen de los detalles del proyecto en general, los reportes obtenidos y los resultados, también se debería incluir todos los problemas asociados durante todas las fases, como problemas con los datos, malas experiencias e incluir recomendaciones para proyectos similares que se puedan desarrollar en el futuro.

7. **Fase 7 : Mejora continua**
   El proceso de minería de datos es cíclico, es decir que se debe desarrollar un plan de control y mantenimiento que asegura la mejora continua del plan en cada iteración, en donde se haga uso apropiado de las estrategias definidas en los pasos anteriores y detecte problemas en el desempeño del modelo.

7.2. El contexto de los datos

La base de datos contenía todos los datos correspondientes al MOOC, ya antes mencionado (ver capítulo 1.3) , del año 2014. Los datos contenidos en dicho son las transcripciones de las Actas Capitulares de la Catedral de Plasencia digitalizadas por parte de los estudiantes del mismo, adicionalmente, como se trataba de un curso, cada transcripción realizada por parte de los estudiantes presenta una calificación, la calificación de dichas transcripciones se encuentra dada por dos (2) estudiantes quienes previamente realizaron la misma actividad, cada uno daba una nota y esta era promediada, en caso de haber unas notas bastante diferentes entraba un tercer estudiante a calificar, esto con la finalidad de dar la nota más adecuada posible.

7.2.1. Estructura de los datos

La base de datos es un conjunto de archivos en formato HTML debidamente ordenado de acuerdo a como se dicto el curso, es decir, en el caso de las transcripciones, estas se encuentran en una carpeta la cual contiene a su vez las transcripciones por cada alumno inscrito en el MOOC, tal como se ve en la imagen 7.2, en su respectiva carpeta con la identificación del mismo dentro del curso, como se ve en la imagen 7.3, dentro de esta se encuentran las calificaciones dadas por los otros dos o tres estudiantes del MOOC, tal como se aprecia en la imagen 7.4, y finalmente, dentro de la carpeta de cada calificador se encuentra un archivo HTML que contiene el mismo archivo del estudiante, con la única diferencia es que este trae la calificación y menciona quien lo califico y a quien califico.

En el MOOC participaron más de 10.000 personas, por lo cual la base de datos es bastante grande, teniendo en cuenta que la calidad de los mismos no es la mejor, puesto que la información almacenada esta enfocada en la visualización desde un navegador web, debido a esto entro los archivos se encuentran lo que se le conoce como *"archivos basura"* puesto que no aportan nada de información necesario, sino que por el contrario inflan el peso de la base de datos, entre estos archivos se

Figura 7.2: Carpetas con todas las actividades del curso

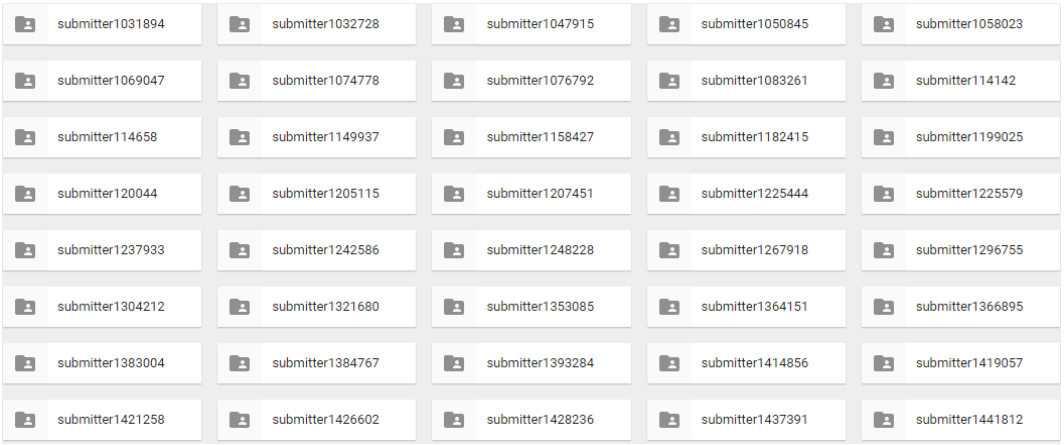


Figura 7.3: Carpetas de los estudiantes de cada actividad

encuentran archivos CSS.

Además de contar con *"archivos basura"* las transcripciones se encuentran llenas de etiquetas HTML además de contar con caracteres **unicode** cuando dicho carácter no se encuentra dentro de la codificación ASCII estándar, esto tal y como se puede apreciar en la imagen 7.5, de tal manera que para poder realizar el trabajo de minería es necesario realizar primero una limpieza de los datos y dejarlos de manera ordenada, de tal forma que sea posible realizar la construcción del documento.

Para poder realizar una limpieza del documento y filtrar cada transcripción es necesario comprender primero la estructura de las transcripciones, como ya se ha

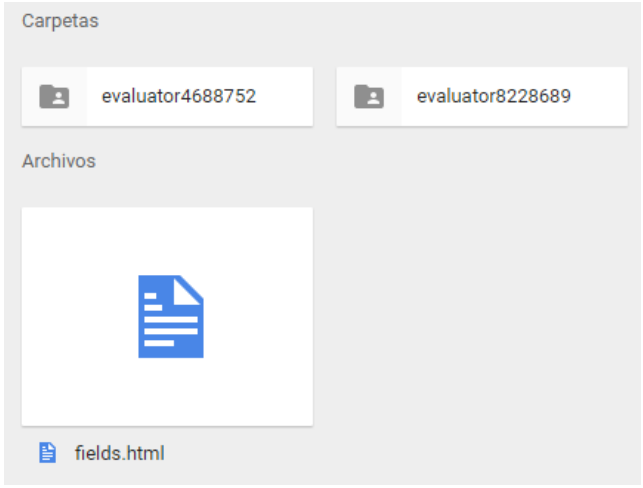


Figura 7.4: La actividad del estudiante con los calificadores de la misma

```
<!DOCTYPE html>
<html lang="en">
  <head>
    <meta charset="utf-8">
    <title>Submission 1782 from Anon (coursera_user_id: 1076792, session_user_id: 955c526538ef899518e8409cb5b1008ae2d93944)</title>
    <link href="../../../export.css" rel="stylesheet">
  </head>
  <body>
    <h1>Submission 1782 from Anon (coursera_user_id: 1076792, session_user_id: 955c526538ef899518e8409cb5b1008ae2d93944)</h1>
    <div class="field-name" id="1c88c8cdf8a113fd"><b>Assignment Statement:</b><span style="font-size: 14.4444446563721px;"> By submitting this assignment,
I affirm that I will abide by the Revealing Cooperation and Conflict Project's agreement to not reproduce or distribute any images of manuscripts. </span>
<br><br><b><i>The student will type the following statement in the answer box below:</i></b><br><b><i>"I will not reproduce or distribute any images of
manuscripts."</i></b><br><br><span style="font-size: 14.4444446563721px;">Your response will not impact your ability to earn sufficient credit in the course to
earn a Statement of Accomplishment. However, if you type in the statement you will earn additional points.</span><br></div>
    <div class="field-value">"I will not reproduce or distribute any images of manuscripts"<br /><br />Group 6<br />Manuscript Image B5H20<br />Diego Alonso y
Gonzalo Fernandez<br />vecino del dicho lugar con una<br />cuba de madera buena de tener<br />vino con una azada e un aza-<br />don e la qual dicha viña e
cuba<br />e azada e azadon dejo al dicho<br />cabillo Don Diego Blazquez Dean<br />la qual dicha viña e cuba e aza-<br />da e azadon arrendo desde hoy<br />dia
del dicho cabillo falta en to-<br />dos los años e dias de su vida del<br />dicho cabillo Alonso Ximenez<br /><br />Group 6<br />Manuscript Image B5H21<br
/>recrescieren; sobre lo qual amoy<br />las partes otorgaron dos contratos<br />ambos de una forma ante Pedro<br />Gonzalez Racionero Notario pu-<br />blico
Apostolico testigos que fueron<br />presentes Gonzalo Martinez Per-<br />tiguero e Juan Fernandez cleri-<br />go dela Magdalena y Aparicio<br />Fernandez
Clarigo de San Pedro<br />que estaban presentes capellanes<br />dela dicha Yglesia = Petrus Gonza-<br />lez, Porcionarins : Notario Appco [Apostolico]<br /><br
/>Group 6<br />Manuscript Image B5H22<br />raices e muebles que los dichos<br />señores cabillo e Benefiiados han<br />en la Villa de Troxillo e en Xara-<br
/>iz Aldea e termino dela dicha Vi-<br />lla; los quales dichos bienes tiene<br />arrendados del dicho cabildo Mar-<br />tin Alonso, clarigo de la Yglesia<br
/>de Santiago, dela dicha Villa, e de<br />Logro san delindados so ciertos linde-<br />ros que son estos que aqui dira en<br />la dicho Villa de Troxillo unas
ca-<br />sas en par dela venta dela dicha</div>

    <div class="field-name" id="d8792a8f26e92e"><p><u><span style="font-size: 11pt;"><b>Transcribe Your Week 12 Bundle and Submit it Here - Due September 5,
2014</b></span></u></p><p>Please type your transcription and submit it using the assignment link (hyperlink). <b>Remember - For full credit for the week's
assignment, you must transcribe at least 3 selections (images). <a href="https://class.coursera.org/medievalspain-
001/wiki/Week_11%3A_Transcription_Project_Page" target="_blank" title="Link: https://class.coursera.org/medievalspain-
001/wiki/Week_11%3A_Transcription_Project_Page">Week 12 Manuscripts are located here.</a></b>You may transcribe additional selections (up to 17 more images) for
extra credit and special recognition in the Revealing Cooperation and Conflict Project. <b>You can choose any three selections to transcribe. Choose those
selections that are most appropriate for your skill level.</b></p><p><br></p><p><b>Required Transcriptions:</b></p><p>Please type in your transcription of the
selected manuscript images. For each page you should complete the follow parts:<br></p><p></p><ol><li><b><i>Required Part One: On the first line, type in your
Group Number. </i></b>(Your peer reviewer will need this to locate the image you transcribed.) For example, if your last name is "Garcia" , then you would
type in "Group 8".<br><br></li><li><b><i>Required Part Two: On the second line, type in your image number.</i></b> For example, "Manuscript Image B1D23"
(Your peer reviewer will need this to locate the image you transcribed.)<br><br></li><li><i><b>Required Part Three: Transcribe at least Three (3) Selections
(Images). Choose Any Three (3) Images.</b></i></li></ol><ul><li>Do your best! We are not expecting perfection. We are hoping that you will be able to transcribe
at least 50 percent of what you see on the page. </li><li>Start your transcription with the first full line of readable text. End your transcription with
the last full line of readable text. (In some cases, your image will have pieces of sentences.)</li><li>Type in the exact letters you see. Please do not spell
out abbreviations.</li><li>Type in lines exactly as you see them.</li><li>Please do not add any additional punctuation. Do type in any punctuation that is on
the page.</li><li>If you cannot read an individual letter or number, then you should type a single period symbol " . " to indicate you are not certain what
letter/number is reported.</li><li>If you cannot read multiple letters or numbers, please use a period symbol " ." for each letter you cannot read. If you
cannot read an entire word, please type in 5 period symbols " …. "</li></ul><p></p><div style="font-size: 14.4444446563721px;"><br></div><p><span style="font-
size: 11pt;"><b>Extra Credit Transcriptions:</b> </span><span style="font-size: 15px;">If you wish to earn extra credit on this assignment, transcribe as
many of the remaining manuscript images in your assigned bundle. You should type in your extra transcriptions below your required transcriptions.</span></p>
</div>
    <div class="field-value">Group 6<br />Manuscrip Image B5H23<br />to por parte del cabildo dela dicha<br />Yglesia le fuera dicho que el dicho<br />Cabildo
tenia ciertas casas arren-<br />dadas en la Villa de troxillo e en<br />sus Aldeas e terminos que perteneci-<br />an ala Su Mesa Capitular del otro<br
/>cabillo; las quales decian que queri-<br />an dar a cemo perpetuamente po-<br />derlo precio en cada un año por qu-<br />anto entendian que era necesario de
```

Figura 7.5: Transcripciones digitalizadas "sucias"

mencionado en capítulos pasados, cada transcripción esta dividida en párrafos y cada párrafo cuenta con un código que lo identifica dentro del documento. El código está compuesto entre 4 a 6 caracteres donde el primer carácter es **'B'**, seguido de un carácter numérico, el tercer carácter pasa a ser alfabético mayúsculo, el cuarto carácter es de tipo numérico, el quinto y sexto carácter son opcionales y son también de tipo numérico, es decir, el código sigue el siguiente patrón: $B\#[A-Z]\#\#\#$

Todas las actividades del curso que se encuentran relacionadas se pasan a una base de datos en hojas de cálculo; en esta hoja de cálculo se encuentran los códigos de los estudiantes, las transcripciones realizadas y otros datos como el compromiso que adquirieron los estudiantes al momento des realizar la transcripción. Dichas transcripciones cuentan con el mismo problema antes mostrado en la imagen 7.5, se encuentran llenos de etiquetas HTML, caracteres unicode, caracteres de escape para las tabulaciones, espacios, etc. una gran cantidad de espacios entre palabras, códigos HTML para la representación de un espacio sencillo, entra otras cosas.

7.2.2. Limpieza de datos

Para la limpieza de las transcripciones y poder generar el documento para su posterior análisis se uso **python 3** y se tomo la base de datos en la hoja de cálculo.

Sabiendo que en el idioma no se encuentran los caracteres **'<'** ni **'>'** y además estos forman parte de las etiquetas de código HTML se procede con su eliminación de todos los textos de la siguiente forma:

```
1  def deleteTag(s):
2      startsTag = s.find('<'); endsTag = s.find('>')
3      if startsTag == -1 or endsTag == -1: return None
4      elif len(s) == endsTag: return s[:startsTag]
5      else: return s[:startsTag] + ' ' + s[endsTag+1:]
6
7  def deleteHTMLTags(s):
8      ok = True
9      while ok:
10         aux = deleteTag(s)
11         if aux is None: ok = False
12         else: s = aux
13     return s.strip()
```

De forma similar se realiza la eliminación de los demás elementos que no pertenecen a las transcripciones: Seguidamente a la limpieza de las transcripciones es necesario identificar cada una, para ello es necesario primero identificar cual es el código de las transcripciones, para así saber cual es la transcripción:

```
def findCode(s):
    startCode = 'B'
    code = ''
    sIndexCode = s.find(startCode)

     # If 'B' not exist
    if sIndexCode == -1: return code
    if  s[sIndexCode+1].isdigit(): # Verify B#
        # Verify B#[A-Z]#
        if s[sIndexCode+2].isalpha() \
        and s[sIndexCode+3].isdigit():
            code = s[sIndexCode:sIndexCode+4]
            # Verify B#[A-Z]##
            if len(s) > sIndexCode+4 \
            and s[sIndexCode+4].isdigit():
                code = s[sIndexCode:sIndexCode+5]
                # Verify B#[A-Z]###
                if len(s) > sIndexCode+5 \
                and s[sIndexCode+5].isdigit():
                    code = s[sIndexCode:sIndexCode+6]
    else: code = findCode(s[sIndexCode + 1:])
    return code.strip()
```

Junto con el código anterior se van generando las transcripciones; estas se almacenan en tuplas, tal que el primer elemento de la tupla sea el código de la transcripción y el segundo elemento sea la transcripción que tiene dicho código, este se hace de manera recursiva con la finalidad de tener todas las transcripciones existentes en una cadena:

```
def extractTranscriptions(s, transcriptions=[]):
    code = findCode(s)

    if code != '':
        indexCode = s.find(code)
        # new s without code
        s = s[indexCode+len(code):].strip()
        code2 = findCode(s)

```

```
10          if code2 != '':
11              indexCode2 = s.find(code2)
12              aux = s[:indexCode2]
13              transcriptions.append((code, aux.strip()))
14              # new s without previous msnuscript
15              s = s[indexCode2:]
16          else:
17              transcriptions.append((code, s.strip()))
18      else:
19          return transcriptions
20
21      return extractTranscriptions(s, transcriptions)
```

Finalmente y para evitar duplicados en las transcripciones, los duplicados se deben a que varios estudiantes presentaron las mismas transcripciones, estas se almacenan en una lista para posteriormente crear un archivo en base a esta de la siguiente forma:

```
1 t = []
2 ...
3 aux = extractTranscriptions(dataClean(s, elements2del))
4 t += [e for e in aux if e not in t]
5           ...
```

Adicionalmente como existen varios elementos que se deben de eliminar de las transcripciones para limpiarlas completamente, y estos elementos no fueron contemplados directamente en el código, es necesario la utilización de un archivo dedicado a esto, donde se listan todos los elementos a eliminar de las transcripciones, este, en primera instancia contiene:

- I will not reproduce or distribute any images of manuscripts.
- \.
- \"
- \,

En el repositorio en GitHub se puede tener acceso al código completo usado para esta limpieza, además del archivo adicional necesario para la eliminación de elementos *"basura"*, el código se encuentra completamente documentado y con ejemplos,

la dirección URL del repositorio es `https://github.com/escuela-ing/RCCP`

Finalmente, el documento obtenido, después de haber realizado el proceso de la limpieza se presenta tal y como en la imagen 7.6, aunque aún presenta pequeños fragmentos de "suciedad", esta claro que ya es completamente legible y es posible comenzar a trabajar con el mismo.

	Codes	Transcriptions
0	B2T7	ren al dicho Tesorero, si el dicho Cabildoy Beneficiados lo contrario feciesen, é loque dicho es, non compliesen, é ambaslas dichas partes otorgar
1	B2T8	moneda que ágora corre, ó dela moneda que corriere al tiempo delas paga aqui en la dicha ciudad en pazé en salbo a su costa por di de SantaMa
2	B2T9	el suyo quales yo el dicho Notario lesnotare á vista de Letrados. Testigos PedroFernandez é Rui Gonzalez Racioneros dela dicha Yglesia, é Alons
3	B2T10	ni por menos, ni por el tanto. É ambaslas dichas partes otorgaron dos contratos en forma. Testigos Pedro Fernandezé Rui Gonzalez, é Monso Go
4	B2T11	Yglesia = Petrus Gonzalez PorcionariusNotario Apostolico En la Ciudad de Plasencia viernes diez y to ocho dias de Noviembre año del nasArren
5	B2T12	(Blank page)Group 10 (M. Zidovec)Manuscript Image
6	B2T13	La autoridad ordinaria arrendo del dho.Cabildo el Corral y meson que está todo caido que dicen el Meson de ayuso en el arrabal desta dicha Ciud
7	B2T14	C i S Fasta aqui havia enido é tenia en rentaSancho Ortiz de A…ñiga, Canonigo deladicha Yglesia; los cuales bienes en Molino moliente y corrien
8	B2T15	Obligó a si y a sus bienes y a todas Sentencias de Santa Yglesia é fizo juramento de decir y aclarar todos losdichos vienes é de non encobrir del
9	B1D66	Algunos Beneficiados dela dicha Iglesia prerendian haber los tales oficiosdisien do que les pertenescen e otrosBeneficia dos tienen la grande opir
10	B1D71	la dicha Iglesia sobre sus Dignidadex personargos canongias y Trebendas, y Raciones y Beneficios Eccos. Que losdhos Beneficiados sienen y p
11	B1D86	ses e temporales presentes e fusunos 85 e otorganon dos contraros firmes por mi el dicho Gutierre Gonz~No.° para cadaimo delas dichas partes
12	B3P59	[First line not visible] ra de juicio y por ese mesmo fecho in curran en Sentencia de excommunion de la qual no puedan haver absolucion sin prim
13	B3P60	dellos que lo non guardasen é por su sen tencia definitiba rescivirla por si mes mo en escritos, lo pronunció y mandó declanó y confirmó diciendo
14	B3P51	[first line missing] Dean y Cabillo; para lo qual todo y cada cosa dello obligaron á ello y á todos sus vienes havidos y por ha ver muebles y raices
15	B3N3	en una linde traviesa, é en un mojon q,e está cerca della, é cerca de unas Encimas é fasta aquí es lindero la dicha tierra del dicho Gonzalo Berdu
16	B3N7	. Gonzalez; é dela otra parte casas de fijos de MartÍn Sanchez, é dela otra parte la dicha Calle, é el traslado de este testamento, tienelo Gil Martí
17	B3N5	la figuera heredad de fijo de Esteban Sa chez, é por parte de ayuso de dicho camino, é dela otra parte contra Albalat here dad de Gonzalo Berduq

Figura 7.6: Documento obtenido del proceso de la limpieza

7.2.3. Flujos de trabajo

EL primer paso a realizar ya una vez se cuenta con el documento limpio, es realizar un flujo en KNIME con la finalidad de terminar el filtro de palabras y caracteres *"basura"*, para esto, el flujo realizado es un **Tag Cloud**, tal como se puede observar en la imagen 7.7 y en la imagen 7.8, el proceso realizado posterior a la construcción del flujo, consiste en realizar una observación al resultado del mismo e identificar las palabras que "no sirven", junto con los nombres de las diferentes personas que se puedan haber encontrado, y cada vez que se encontraban, el flujo se volvía a hacer, pero esta vez se le añadían los filtros ya antes mencionados, esto con la finalidad de obtener la mayor cantidad de nombres posible.

7.3. Correlaciones

Una vez se cuenta con los diccionarios es posible realizar una búsqueda de las correlaciones existentes entre A y B, siendo A y B diccionarios.

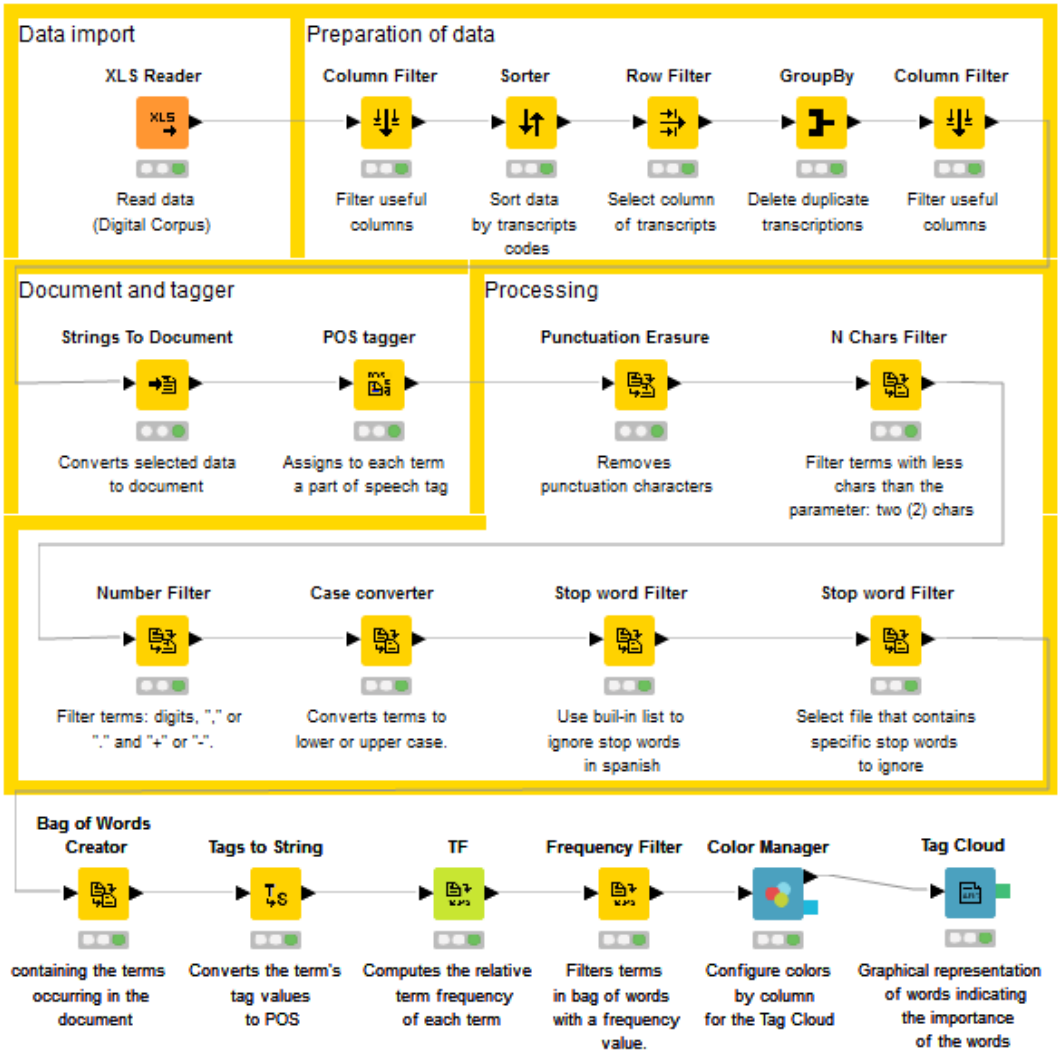


Figura 7.7: Flujo en KNIME de Tag Cloud para limpieza y extracción

Las correlaciones dadas en este punto son dependientes del orden de entrada de los diccionarios, es decir, la correlación planteada se encuentra dada por la siguiente formula matemática:

$$\text{Si } A \neq B \text{ entonces}$$
$$A \to B \neq B \to A$$
$$\text{Si } A = B \text{ entonces}$$
$$A \to B = B \to A$$

Esto teniendo en cuenta que tanto A como B representan el conjunto de elementos que componen cada diccionario, en la practica para este tipo de relaciones, en la cual se cuentan la cantidad de los elementos de B que se encuentran en la misma transcripción que al menos un elemento de A es mediante el siguiente código.

Figura 7.8: Tag Cloud para limpieza y extracción

```
for e in transVal.iteritems():
    s = str(e[1])
    for ed1 in ld1:
        if s.find(ed1) != -1:
            for ed2 in ld2:
                try: correlations[ed1][ed2] += s.count(ed2)
                except:
                    try:correlations[ed1][ed2] = 0
                    except:correlations[ed1] = {ed2 : 0}
```

En el código anterior, e representa una transcripción, $ed1$ representa un elemento del primer diccionario ingresado y $ed2$ representa un elemento del segundo diccionario.

Capítulo 8

Conclusiones

Al realizar el proyecto Minería de texto histórica - colaboración al proyecto 'Revealing Cooperation and Conflict Project' fue posible resaltar el hecho de que a pesar que existen muchos ámbitos de estudio y de investigación, todos ellos no tienen porqué desarrollarse por separado. Es este el caso en el que desde la informática fue posible colaborar en el desarrollo de un proyecto humanístico de tal manera que se fomento participación y como ya se mencionó, un dialogo fluido entre ambas disciplinas. Hay dos cosas importantes a rescatar, en primer lugar la profundización que se ha hecho para la rama de la minería de texto histórica y más específicamente con la producción de los diccionarios que apoyan directamente la investigación de Dr. Martínez-Dávila pues dicha lista de palabras claves y descubrimiento de correlaciones entre las mismas permite, en la medida y forma de su uso, entre otras cosas, verificar o tener más certeza sobre las propuestas y descubrimientos que se han hecho.

En segundo lugar es posible afirmar que teniendo los textos históricos digitalizados listos para su análisis, por medio de una metodología que permita conocer el corpus, adaptarlo según la información necesaria, manipularlo de tal manera que solo la información relevante quede disponible para el investigador, haciendo tratamientos basados en el lenguaje y en palabras clave, utilizando técnicas estadísticas y matemáticas, para finalmente obtener resultados gráficos, si agiliza el proceso que el experto debe llevar a cabo para descubrir información. Lo anterior se ve apoyado en el hecho de que con hacer uso de la metodología con el corpus de Actas Capitulares de la Catedral de Plasencia es posible identificar lugares, nombres, apellidos, palabras de la época, abreviaciones de términos, entre otros, que permitirán ir forjando una idea del texto que se está utilizando sin la necesidad de descubrirlos por medio de la lectura detenida.

Después de realizar todo el análisis usando la metodología queda en manos del profesional interesado, ya sea historiador, paleógrafo o geógrafo, utilizar los resultados obtenidos de manera que tenga el poder de saber que información va a tener en cuenta y cual información no es útil para su trabajo.

También se concluye que efectivamente el uso de técnicas de minería de texto sobre documentos históricos se ve afectado por el lenguaje antiguo en que están escritos los mismos, las técnicas se basan en el uso de lenguajes modernos, en este caso, difiriendo bastante de los de la fuente. Para hacer minería de texto histórica hay que tener muy presente que será necesario conocer aquel lenguaje antiguo del que se esté dependiendo. Para el uso de la metodología realizada fue necesaria la creación manual de listas de palabras a ser ignoradas al momento de analizar el corpus pues no hacían parte del lenguaje y desviaban los resultados hacía incongruencias e incluso alteraciones de la información.

Teniendo en cuenta la naturaleza del corpus, el proceso de preparación y limpieza de los datos fue complejo debido al formato en el que fueron entregados los mismos. No solo influyó el hecho de que se tenía una gran cantidad de datos desordenados pues el libro de las Actas Capitulares de la Catedral de Plasencia no estaba digitalizado como tal sino se tenían fragmentos digitalizados provenientes de los resultados del MOOC que realizó Dr. Roger Louis Martínez-Dávila , también influyó el hecho de que las transcripciones no venían totalmente limpias y listas para su uso, incluían gran cantidad de código HTML filtrado entre las mismas lo cual fue un gran problema a la hora de evidenciar la real información ya que si se incluía el texto de esta manera sobre la metodología realizada en la herramienta KNIME, sería imposible obtener resultados correctos pues la máquina entendería que aquellos códigos también hacen parte de la terminología del texto y lo incluiría en el análisis como cualquier otra cadena de texto.

Aparte de lo anterior, otro problema con respecto a la información suministrada fue que, no incluía un patrón claro que permitiera desarrollar fácilmente un algoritmo para limpiar rápidamente el texto sino que fue necesario descubrir la mayor cantidad de patrones que seguía la serie de transcripciones para poder lograr limpiar y organizar las mismas, el desarrollo del algoritmo tampoco fue sencillo y requirió incluso transformación de caracteres y texto a los símbolos del lenguaje español.

Apéndice A

Diccionarios

Se presentan los siguientes diccionarios de palabras clave, descubiertas a partir de Actas Capitulares de la Catedral de Plasencia acorde con los objetivos específicos propuestos para el proyecto:

- Diccionario de nombres
- Diccionario de roles
- Diccionario de lugares

Con respecto al diccionario de palabras claves que incluye términos relacionados con los roles de la época, se presenta también un diccionario de roles en aspectos generales, que no incluye especificaciones demasiado profundas encontradas dentro de la información.

A.1. Diccionario de Nombres

A continuación se muestra el Diccionario de nombres que surge del análisis de las Actas Capitulares de la Catedral de Plasencia, es posible encontrar los nombres de los personajes con las ocurrencias frecuentes (en sus diferentes variaciones).

1. Abaen Zapatero
2. Abraen
3. Abraen Zapatero
4. Adorraesmen
5. Alban Sanchez de Foledo
6. Albaro de San Pedro
7. Albar Sanchez
8. Albar Sanchez de Toledo
   Eino del Rey
9. Alcalde

10. alíed Arradamen
11. Alfgonso Sanchez
12. Alfonnes Canonicus
13. Alfonsies Ferrandez
14. Alfonso de Garganta la Olla
15. Alfonso de Organ
16. Alfonso Emo
17. Alfonso Feltez
18. Alfonso Fernandez
19. Alfonso Ferrandez
20. Alfonso Ferrandez de Logroño
21. Alfonso Ferrandez de Sogrono
22. Alfonso Garcia
23. Alfonso Garcia Fesoreno
24. Alfonso Gomez
25. Alfonso Gomez Cabrero
26. Alfonso Gonzalez
27. Alfonso Gonzalez Criado
28. Alfonso Gonzalez de Amasco
29. Alfonso Gonzalez de Amusco
30. Alfonso Gonzalez de Medina
31. Alfonso Jimenez
32. Alfonso Martin
33. Alfonso Martinez
34. Alfonso Martinez dela Corraliza
35. Alfonso Maxnin
36. Alfonso Sanches de Anuncio
37. Alfonso Sanchez
38. Alfonso Sanchez de Amusco
39. Alfonso Sanchez Molinero
40. Alfonso Sanchez Montero
41. Alfonso Til
42. Alfonsus
43. Alfonsus (Alfonso)
44. Alfonzo Ferrandez
45. Alfonzo Gomez
46. Ali Arradamen
47. Ali Moro
48. Alonso Fellez
49. Alonso Ferrandez Panaigua
50. Alonso Ferrandez Paniagua Caballero
51. Alonso Gonzalez
52. Alonso Gonzalez de Amusco
53. Alonso Gonzalez de Arranco
54. Alonso Gonzalos
55. Alonso Gutierrez
56. Alonso Martinez
57. Alonso Sanchez
58. Alonso Sanchez del Montalban
59. Alvar Blazquez

60. Alvar Garcia de Grimaldo
61. Alvaro de Salazar
62. Alvaro de San Pedro
63. Alvarus Decanus Placentinus Alfonsus
64. Amat Moro Bexarano
65. Amat Moro Teredor
66. Andreas Perez
67. Andres Domingez
68. Andres Dominguez
69. Andres Domintuez Bachiller
70. Andres Gonzalez
71. Andres Martin
72. Andres Perez
73. Andrez Gonzalez
74. Andrés Pérez
75. Antonio Grae
76. Antonio Martinez
77. AntonioMartinez
78. Antonio Martinez Capellan
79. Antonius Martinus
80. Anton Perez
81. Aparicio Martin
82. Arcediano de Plasencia
83. Arcediano de Trasillo
84. Arediano de traxillo
85. Bachillen
86. Bachiller Rui Garcia Canonigos
87. Bano gomez
88. Bano Gomez
89. Bano Gomez Racioneros
90. Barco Ferrandez
91. Barco Gomez de Saabedar
92. Barrolome San Chez
93. Barrtolome Sanchez
94. Bartolomea
95. Bartolome Sanchez
96. Bartolomé Sánchez
97. Basco Gomez
98. Basco Gomez de Almaraz
99. Basco Gomez Racionero
100. Basco Gonzalez
101. Bascon Gomez
102. Basio Gómez
103. Beatriz
104. Beatriz Solen
105. Beatriz Soler
106. Belasan
107. Belasan Ortiz Porcionarius
108. Belascus
109. Belasio Ferrandez de Madrigal

110. Benita Benito Perez
111. Benito Martin Perez
112. Benito Perez
113. Benito Sanchez
114. Benito Sanchez Osorio
115. Blanco Perez
116. Blasco Ferrandez
117. Blasco Gomez
118. Blasco Gomez de Alma
119. Blasco Gomez de Almaraz
120. Blasco Perez
121. Capellan
122. CAROLUS IV
123. Christobal Sanchez
124. Christobal Sanchez Canonigo
125. Christobal Sanchez Conomigo
126. Clemeinte
127. Cofradia de Santi-Espiritus
128. condera
129. Cristóbal Sánchez
130. Dean Cabillo
131. Dean Don Diego Blazquez
132. Dean Don Juan Sanchez y Sacristanes
133. Dean y Cabillo
134. Demetia
135. Deneria
136. Didacus Martíednez
137. Diego Alfonso
138. Diego Alfonso Molinero
139. Diego Bermejo
140. Diego Blasco
141. Diego Blascp
142. Diego de Matos
143. Diego de Riorsonsillo
144. Diego Esteban
145. Diego Esteban de Xaariz
146. Diego Ferrandez
147. Diego Ferrandez dela Coralla
148. Diego Ferrandez Subchantre
149. Diego Ferrandez Troton
150. Diego Garcia
151. Diego Garcia Bexenano
152. Diego Garcia Bexerano
153. Diego Gómez
154. Diego Gomez
155. Diego Gomez de Almanaz
156. Diego Gonzalez Compañenos
157. Diego Gonzalez de Carvajal
158. Diego Marcos
159. Diego Marrin Sabrador

160. Diego Martíednez
161. Diego Martin
162. Diego Martinez
163. Diego Martinez del Barco
164. Diego Martinez Racionero
165. Diego Muñoz
166. Diego Nuñez,
167. Diego Nuñez
168. Diego Obispo de Plasencia
169. Diego Perez
170. Diego Perez de Granada
171. Diego Perez Torreno
172. Diego Perez Torrero
173. Diego Pimentel Machacon de Deleitosa
174. diego Rodriguez de Carvajal
175. Diego Rois
176. Diego Sanchez
177. Dn Albaro de Morray
178. Dn. Alfonso Garcia
179. Dn. Alfonso Garcia Arcediano de Troxillo
180. Dn. Allaro
181. Dn Alvaro
182. Dn Alvaro de Morroy
183. Dn. Alvaro de Salazar
184. Dn Gil Martinez Arcediano de Traxillo
185. Dn Gonzalo Garcia de Carvajal
186. Dn. Gonzalo Gutierrez
187. Dn. Martin Fernandez Arcediano
188. Dn Martin Ferrandez Bachiner
189. Dn. Miguel Sanchez
190. Dn Miguel Sanchez Arcediano
191. Dn.Til Martinez
192. Doña Benita
193. Doña Gracia
194. Doña Leonor
195. Doña Mari-Perez
196. Doña Nuña
197. Doña Pasquala
198. Doña Plata
199. Doña Sol
200. Doña Sol muger de Pasaron
201. Doña Ximena
202. Doctor Don
203. Doctor Don Jil Martinez de Soria
204. Doctor Garcia Lopez
205. Doctor Juan Fernandez

206. Domingo Martin
207. Dona Bartolomea
208. Dona Beatriz
209. Dona Benita
210. Don Abraen
211. Dona Gracia
212. Dona Leonor
213. Don Alfonso Garcia
214. Don Alfonso Garcia Arcediano de Fraxillo
215. Don Alfonso Garcia Fesoxero
216. Don Aloaro de Salazar
217. DonAlonso Rodriguez de Alabienda
218. Don Alvaro de Salazan
219. Don Alvaro de Salazar Dean
220. Dona Pasquala
221. Don Arradamen
222. Don Arradamen Moro
223. Dona Teresa
224. Doñn Bartolomea
225. don Clemeinte
226. Don Clemeinte
227. Don Clemente dela Figuera
228. Don Diego Blazqz
229. (Don) Garcia Lopez deCarvajal
230. Don Gonzalo de Astuniga
231. Don Gonzalo Gutierrez dela Calleja Chansre
232. Don Gonzalo Obispo
233. Don Juan Inez Chantre
234. Don Juan Sanchez Chantre
235. Don Marcos
236. Don Martin Fernandez de Sonia
237. Don Martin Ferran
238. Don Martin Ferrandez
239. Don Martino Ferrandez
240. Don Migan
241. Don Miguel Inez
242. Don Miguel Sanchez
243. Don Millan
244. Don Pedro Ferrandez
245. Don Pedro Obispo
246. Don Rodrigo de Carvafal
247. Don Rodrigo de Carvajal
248. Don Ruarcos
249. Don Rui Garcia de Salamanca
250. Don Yaco Albeha
251. Don Yusaf
252. Duran González
253. el Arcediano Don Timon

254. el conde
255. Eluego
256. Elvira
257. Ervira
258. Estaban Sanchez
259. Esteban Fernandez
260. Esteban Hernandez el Bote
261. Esteban Hernandez Jil Martinez
262. Esteban Perez
263. Esteban Sanchez
264. Estevan Sanchez
265. Eyo Alfonso Ferrandez
266. Facen Herrador
267. Falabera Berracol
268. Fancino Hil el Viglo
269. Fancisco Fernandez de Foritzo
270. Fatigos Arias Salado
271. Fernandas Garcia
272. Fernandez Lassre
273. Fernando Garcia
274. Fernán Martínez
275. Ferran
276. Ferran Albarez
277. Ferran Albarez de Montalban
278. Ferran Alfonso
279. Ferran Alfonso Bachiller
280. Ferran Alfonso Clerigo de San Esteban
281. Ferran Alfonso de Garganta la Olla
282. Ferran Alvarez
283. Ferran deLogroño,
284. Ferrandez Bravo
285. Ferrandez Cabillo E Aldonso
286. Ferrandez Canonigo
287. Ferrandez del Barco
288. Ferrandez Mojon Viejo
289. Ferrando Gomez
290. Ferran Dominguez
291. Ferran Dominguez de Salamanca
292. Ferran Garcia
293. Ferran Garcia del Parco de Castano
294. Ferran Garcia del Parco del Castaño
295. Ferran Gonzalez
296. Ferran Gonzalez de Villanueba
297. Ferran Gonzalez Racionero
298. Ferran Martinez
299. Ferran Martinez Bachiller
300. Ferran Martinez Racionero

301. Ferran Martinez Sobchantre Gonzalo de Salamanca
302. Ferran Perez
303. Ferran Pérez
304. Ferran Rodrigues
305. Ferran Rodriguez
306. Ferran Sanchez
307. Ferran Sanchez de Tarandilla
308. Ferran Ximenez el mozo
309. Ferran Ximenez el Mozo
310. Ferran Yniguez
311. Ferrigo Martin Fernandez de Logroño
312. Ferrigos
313. Ferrigos Martin
314. Ferron Martinez Lohantre
315. Festigos Juan Ferrandez dela Magalena
316. Festigos Sancho Ortiz de Astañiga
317. Foribio Ferrandez
318. Foribio Sanchez
319. Foribio Sanchez del Barco Racio
320. Foribio Sanchez y Garcia Lopez
321. Foribus
322. Foribus Martinez de Villarba
323. Foribus (Martinzede Villarba Canonigos)
324. Forivio Ferrandez
325. Forivio Ferrandez Esno
326. Francisco Ferrandez
327. Fran Martinez Canonigor
328. Fray
329. Gabriel Sanchez
330. García Alvarez de Toledo
331. García Gonzalez
332. Garcia Alfonso
333. Garcia de Castrobende
334. Garcia de Castroberde
335. Garcia de Plasencia
336. Garcia Feranrarius Placentinus
337. Garcia Fernandez
338. Garcia Fernandz de miranda
339. Garcia Ferrandez
340. Garcia Gonzalez
341. Garcia Gonzalez de Herrea
342. Garcia Lopez
343. Garcia Lopez de Carvafal
344. Garcia Lopez de Carvajal
345. Garcia Lopez Diego Martinez
346. Garcia Lopez Jastre
347. Garcia Lopez Jaure
348. Garcia Lovo

349. Garcia Marquez
350. Garcias Ferrandez
351. Garcpia Fernandez
352. Gil Esteban
353. Gil Fernandez
354. Gil Martinez
355. Gil Martinez de Soria
356. Gil Martínez
357. Gómez Fernando
358. Gomez Fernandez
359. Gomez Fernandez de Albalad
360. Gomez Fernandez de Albalade
361. Gomez Ferrandez
362. Gomez Ferrandez canonigo
363. Gomez Ferrandez de Albalád
364. Gomez Garcia
365. Gomez Gerrandez
366. Gomez Gonzalez
367. Gonazlo Bendigo
368. Gonzalbus (Gonzales) Gutierrez
369. Gonzalbus Gutierrez
370. Gonzalez
371. Gonzalez Porcionarius
372. Gonzalo Alfonso
373. Gonzalo Alfonso Penniguero
374. Gonzalo Bendugo
375. Gonzalo Berduga
376. Gonzalo Berdugo
377. Gonzalo Blazquez
378. Gonzalo de Astonga
379. Gonzalo de Ballejo
380. Gonzalo de Salamanca
381. Gonzalo Ferrandez
382. Gonzalo Ferrandez de Sevilla
383. Gonzalo fijo de Blasco Sanchez
384. Gonzalo García de Carvajal
385. Gonzalo Garcia
386. Gonzalo Gutierre
387. Gonzalo Gutierrez
388. Gonzalo Gutierrez dela Callefa Chasmtrem
389. Gonzalo Gutierrez de la Callejas Tesorero
390. Gonzalo Gutierrez della Calleja
391. Gonzalo Martin
392. Gonzalo Martin de Albalad
393. Gonzalo Martinez
394. Gonzalo MartinezPerriguero
395. Gonzalo Martinez Persiguero y Aparicio
396. Gonzalo Martinez Pertiguero

397. Gonzalo Martinez Poriguero y Albano
398. Gonzalo Martinez Portiguero
399. Gonzalo Matas
400. Gonzalo Monso
401. Gonzalo Monto Portigueno
402. Gonzalo Perez dela Camana
403. Gonzalo Perez dela Camara
404. Gonzalo Ruiz
405. Gonzalo Ruiz Capellanes
406. Gonzalo Sanchez
407. Gonzalo Xil Ferrus
408. Gonzalo Yanez
409. Gregorio Martinez
410. Gunierre Gonzales
411. Guriox de Gonzalez
412. Gutierre Gonzalez
413. Gutierre Gonzalez del Barco
414. Gutierrez Ferrandez
415. Gutierrez Gonzalez
416. Gutierroz Gonzalez
417. Gutierr.z Gonzalez
418. Hamados Alfonn
419. Iague Sancho de Albalad
420. Jean Criado
421. Jil Esteban
422. Jil Fernandez
423. Jil Ferrandez
424. Jil Gutierrez
425. Jil Gutierrez Nacionero
426. Jil Gutiertez
427. Jil Martinez
428. Joanes
429. Joanes Ferrandez
430. Joanes (Juan)
431. Juan
432. Juana Gonzalez
433. Juan Alfonso
434. Juan Alfonso Zapatero
435. Juan Alonso
436. Juan Alonso de Fiasillo
437. Juan Clemente Racionero
438. Juan de Alva
439. Juan de Carvajal
440. Juan de Guadalfajara
441. Juan del Barco
442. Juan de Logroño
443. Juan de Obiedo
444. Juan de Perea
445. Juan de San Juan
446. Juan Duran

447. Juanes Canonicus
448. Juan Feliz Cierrigo de San Nicolas
449. Juan Fernandez
450. Juan Fernandez de la Corchuela
451. Juan Ferrandez
452. Juan Ferrandez
453. JuanFerrandez
454. Juan Ferrandez Capellanes
455. Juan Ferrandez de Betamos
456. Juan Ferrandez de Betanzos
457. Juan Ferrandez de Cabrenos
458. Juan Ferrandez de Cabreros
459. Juan Ferrandez de Velvis Clerigo
460. Juan Firrandez
461. Juan Garcíeda
462. Juan Garcia
463. Juan Gomez
464. Juan Gomez Bofon
465. Juan Gonzalez Capellan
466. Juan Gutierrez
467. Juan Julian
468. Juan Martin
469. Juan Martinez
470. Juan Martín Merino
471. Juan Mateos
472. Juan Mateos Capellan
473. Juan Mateos Clerigo de Santiago
474. Juan Nisñez de Villalobos
475. Juan Nuñez
476. Juan Nunez
477. Juan Paneagua
478. Juan Paneagua
479. Juan Perez
480. Juan Perez de Buenabentura
481. Juan Rodríguez de Fuentevero
482. Juan Rodriguez
483. Juan Rodriguez de Sevilla
484. Juan Rodriguez de Sevilla
485. Juan Rodriguezez
486. Juan Ruiz
487. Juan Ruzz deCamargo
488. Juan Sacristan
489. Juan Sanchez
490. Juan Sanchez de Avila
491. Juan Sanchez de Segovia
492. Juan Sanchez Jil
493. Juan Sanchez Notario
494. Juan Sanchez Racioneros

495. Juan Sanchez Vicario
496. Juan San Pedro Gonzalez Racionero
497. Juan Simon
498. Juan Sobrino de Diego Martinez del Barco
499. Juan Timon
500. Juez
501. Leyer Arcediano
502. Llorencio Perez
503. Llorencio Perez
504. Lope de Ruvian
505. Lope Ortiz de Montoya
506. Ludovicus Porcionarius
507. Lugan
508. Luis Alfonso
509. Luis Alfonso de Sevilla
510. Luis Alonso de Sebilla
511. Luis Ferrandez
512. Luis Ferrandez Bachillen
513. Luis Gonzalez
514. Luis Gonzalez de Turones
515. Luis Sanchez Vicario
516. Mansion Perez
517. Maras Sanchez Ballon
518. Marcos Sanchez (Ballon)
519. Maria Alfonso
520. Maria Basio
521. Maria Blanco
522. Maria Blasio la Rebellada
523. Maria Blasu la Habellada
524. Maria de Candelaria
525. Maria de Candeleta
526. Maria Ferrandez
527. Maria Gonzalez
528. Maria Maria Martin
529. Maria Martin
530. Maria Sanchez
531. Maria Sanchez su muger
532. Mari-Diego (La Bóona)
533. Mari-Perez
534. Mari Sanchez
535. Marta Ferrandez
536. Martin Blasco
537. Martin de San Esteban
538. Martin Diego
539. Martin Diego Corcho
540. Martin Diego Corcho
541. Martin Dominguez
542. Martinez del Barco

543. Martinez de Villarba Canonigos
544. Martin Fernandez
545. Martin Ferrandez
546. Martin Francisco Carbonero
547. Martin Gonzalez de Albalad
548. Martin Perez
549. Martin Sanchez
550. Martin Yañez
551. Martin Yanez
552. Martín Blasco
553. Martín Sanchez
554. Mateo Sanchez Ballestero
555. Mateo Sanchez (Ballon)
556. Mateo Sanchez Ballon
557. Mayordomo del Obispo
558. Mayordomos
559. Mencia Sanchez
560. Menga Munos
561. Menga Muñoz
562. Mera Capiralar
563. Meton Perribado
564. Miguel de Froxillo
565. Miguel Lopez
566. Miguel Lopez de Buendia
567. Miguel Martinez
568. Miguel Sanchez
569. Miguel Sanchez de Asetuna
570. Miguel Sero
571. Miguel Soté
572. Miguel Yague
573. Monar Juan Garcia
574. Monso Gonzalez
575. Nicolas de Castro
576. Nicolas Fernandez
577. Nicolas Ferrandez
578. Nicolas Ferrandez de Sandobal
579. Nicolaus Fernandez
580. Nicolaus (Nicolas) Fernandez
581. Nicolavo
582. Notario publico
583. Nuño Alfonso
584. Obispo de Plasencia
585. oficios decinales
586. Parecs Gonzles
587. Pedo Gonzalez
588. Pedro Cabildo
589. Pedro Chamizo
590. Pedro de Alfaro
591. Pedro de Trejo
592. Pedro de Valladolid

593. Pedro de Valladolid de avia fajo
594. Pedro Feirandez
595. Pedro Fernandez
596. Pedro Ferradez
597. Pedro Ferrandez
598. Pedro Ferrandez del Hoyo
599. Pedro Ferrandez de Logroño
600. Pedro Ferrandez de Soria Racionero de la dha
601. Pedro Ganzalez Rocionero
602. Pedro Garcia Sobrado
603. Pedro Gonsalez Racionero
604. Pedro Gonzales
605. Pedro Gonzales Racionero
606. Pedro Gonzalez
607. Pedro Gonzalezo
608. Pedro Gonzalez Racionero
609. Pedro Gutierrez
610. Pedro Jil fijo de Alfonso Jil de Granada
611. Pedro Juan el Ballon
612. Pedro Juan Ferran
613. Pedro Lopez
614. Pedro Martinez
615. Pedro Sanchez
616. Persigueso
617. Pesrus Ferrandez
618. Pesrus (Pedro) Ferrandez
619. Pestus Gonzalez
620. Petrus Fernandez
621. Petrus Ferrandez
622. Petrus Ferrandez Portionarius
623. Petrus Gonzales
624. Petrus Gonzales Poraisanius
625. Petrus Gonzalez
626. Petrus Gonzalez
627. Petrus Gonzalez Poriconarius
628. Petrus Gonzalo
629. Petsus Forrandez
630. Plata
631. Porcionarius
632. Ramin Nuñez
633. Ramir Nu
634. Ramir Nuñez
635. residentes en la dicha Yglesia
636. Rey e Reyna
637. Roder
638. Rodericcus
639. Rodericus
640. Rodericus Archidia
641. Rodericus Breala Canonicus

642. Rodericus (Rodrigo)
643. Rodericus(Rodrigo)
644. Rodrigo Alfonso
645. Rodrigo Alfonzo
646. Roman Gondillo
647. Rui Garcia
648. Rui Garcia de Salamanca
649. Rui Gomez
650. Rui Gonzalez
651. Ruiz de Camargo
652. Ruy Gonzalez Racionero
653. Sadovicces
654. Sadovicces (Sandobal)
655. Salbador Jesuchristo
656. Sanchez Ballon
657. Sancho
658. Sancho de Albalad
659. Sancho Ertiz
660. Sancho Ertiz de Zuñiga
661. Sancho Gomez
662. Sancho Martinez
663. Sancho Ortiz
664. Sancho Ortiz Caronigos
665. Sancho Pasqual
666. Sancho Sanchez
667. Sanelu Gomez
668. San Esteban Petrus Gonzalez
669. San Juan
670. San Juan de Turio
671. Santa Maria
672. Santo Perez
673. Senor Cabildo
674. Senor el Rey
675. Señor de Oropesa
676. Señor el Conde de Astuñiga
677. Sol
678. Sope Diaz
679. Sra Pata
680. Su Señoría Eoblogo
681. Teresa Fernandez
682. Teresa Lope
683. Teresa Lopez Martinez
684. Terran Marrniz
685. Til Fernandez
686. Till Garcia Juana Martinez
687. Tome Gil de Juacos
688. Tome Gil de Quacos
689. Toribio Fernandez
690. Toribio Ferrandez
691. Toribio Sanchez
692. Torivio Ferrandez

693. Torivio Sanchez
694. Trian de Carvasol
695. ullí Moro
696. Vuclaus
697. Ximenez de Troxillo
698. Yague Sanchez
699. Yague Sancho
700. Ybañez Sancho
701. Ynes Gonzalez
702. Ysabel Alonso
703. Yusaf
704. Yusefe Arañon
705. Yusefe Champus
706. Zebrian Pimentel

A.2. Diccionario de Lugares

A continuación se muestra el Diccionario de Lugares que surge del análisis de las Actas Capitulares de la Catedral de Plasencia, es posible encontrar las denominaciones de los lugares con las ocurrencias frecuentes (en sus diferentes variaciones).

1. Albal
2. Albalad
3. Albaladejo
4. Albalat
5. Aldeanueba
6. Aldeanueva
7. Alion de la Erguijuela de Arrañuelo
8. Alion dela Moeda del Yuglar
9. Alion de Saucedilla
10. Almaraz
11. Almonte
12. altozano de los frailes
13. Alvalar
14. Alveda del Yoglar
15. Amasco
16. Amat
17. Amat Moro
18. Amusco
19. Aragon
20. Arcadia
21. Arranuelo
22. arroyo Barbadon
23. arroyo de Barbadon
24. Arroyo de Barbadon
25. arroyo de la Cañada que dicen
26. arroyo de la Habaza
27. arroyo de la Pardala

28. Arroyo de la Pardala
29. arroyo de lavin
30. arroyodel fojarastan
31. arroyo del Tamujan
32. arroyo del Tamujar
33. arroyo de Pajarejos
34. arroyo de Valde-Miguel
35. arroyo de Valde-Miguel Diego
36. arroyo que viene ala fuente
37. Ayuntamientos
38. Bara
39. Barco
40. Bartolomé Sanchez Serradilla
41. Bejar
42. Benero de Doña Mayor
43. Bentosilla
44. Berrocal
45. Berrocal de Garcia Lopez
46. Bexan
47. Bexar
48. Boadilla
49. Bodega
50. Bueso
51. cañada de los Cebollares
52. cañada de Ventosilla
53. Caballerias
54. Cabreros
55. Calamoco
56. Calle de Carretas
57. calle de Diego Gomez de Almanaz
58. Calle de Dn Marcos
59. Calle de Don Marcos
60. calle de fijos de Diego Gomez
61. calle de Garcia
62. Calle de Gracia
63. calle de la Gracia
64. Calle de la Rua
65. calle de la Zapateria
66. Calle de la Zapateria
67. calle del Concejo
68. calle del Rey
69. Calle del Rey
70. Calle del Sol
71. Calle del Sol cerca de la Plaza
72. calle de Santa María
73. calle de Santa Marída
74. Calle de Santa Maria
75. Calle de Talavera
76. calle deTroxillo
77. Calle de Troxillo
78. calle de Truxillo

79. Calle de zapateria
80. Calle de Zapateria
81. Calle de Zapteria
82. Calleja Fesorero
83. calle major que bá al Portigo
84. calle primera
85. Calle publica de Concejo
86. Calle publica del Rey
87. calle que dicen de Don Marcos
88. calle Santa Maria
89. calles públicas del Rey
90. Calle tras Santa Maria
91. Calzones
92. camino a Mirabel
93. camino de Coria a Plasencia
94. camino de Galisteo
95. Camino de la plaza
96. Camino del Carril
97. camino de Plasencia
98. camino de Plasencia a Coria
99. Camino Judeno
100. camino que ba de Xaraicejo a Plasencia
101. campo del Calamoco
102. Campo de Santa Maria
103. canada delos Ganados
104. capella de Sant Pablo
105. capilla de San Pablo
106. Capilla de San Pablo
107. Capilla de San Pablo
108. Capilla de Santa Catalina
109. Capilla de Sant Paulo
110. Carvajal
111. Casa del Dean
112. Casa de Mari-Domingo
113. casa en el arrabal de la dicha Villa
114. Casar de Elvira
115. casares del Deanazgo
116. casas dela Ensinilla
117. Casas dela Ensinilla
118. casas del Cabillo dela Universidad de-los Clerigos
119. casas de Santa Maria del Campo
120. Cerraldo
121. Cerro del Padron
122. cerro de Valde-Miguel Diego
123. Chintes
124. Chistes
125. Ciudad cerca dela Dejera
126. Ciudad de Truxillo
127. Ciudad Real
128. Cofradia de Santi-Espiritus

129. Collado
130. colmenas
131. Coria
132. Corral de cerrar Ganados
133. corrales
134. corte de Roma
135. Dehesa del Berrocal
136. Dehesa del canchal
137. Dehesa del Canchal
138. Dehesa de los Cebollares
139. Dehesa de Meajadas
140. Dehesa de Mehafadas
141. Dehesa de Santa Maria del Campo
142. Deheza Santa Maria del Campo
143. Delirosa
144. Eglesia de Plasencia
145. Eglesia de San Esteban
146. el arroyo dela fuente delos Vallestero
147. el arroyo de los Maruteros
148. el arroyo de Pajarejos
149. el arroyo Guadaperal
150. el Cadozo de Dona Pasquala
151. el camino que bá de Plasencia al Tanrin
152. el cauce lin
153. el cerro Camino de las Casas
154. el Gamonal
155. el Parral
156. El Pueblo de Xadricejo
157. el Puerto del Caballo
158. el terreno de Doña Mayor
159. Enguijuela de Arrañuelo
160. Ensitrilla
161. Ergüijuela
162. Erguijuela
163. Erquijuela
164. Ferrigos
165. Fesoreria
166. Fesorero
167. Florencia
168. Froxillo
169. Fuente de Ybañe Sancho
170. fuero de Rivera
171. Galisteo
172. Garganta de Degüella Cabras
173. Granada
174. Guadalfajara
175. guijo
176. huerta del obispado

177. huertas
178. huerto de Zebrian Pimentel
179. Iglesia de San Juan del Arrabal
180. Iglesia de San Martin
181. Iglesia de San Martín
182. Iglesia de San Nicolás
183. Iglesia de Santiago
184. Juacos
185. La bodega
186. laCalle de D.n Marcos
187. La calle de Troxillo
188. la calle de troxillo arroyo del tamusjan
189. la Calle Real
190. la Carniceria
191. la Casa dela Pardala
192. La clousstra
193. La Delirosa heredamiento de Basco Gomez de Almaraz
194. la dicha
195. la fuente de Ybañe Sancho
196. lagares
197. La Habaza
198. la Panaderia
199. La Pardala
200. la Perdiguera
201. la Peñuela
202. la puerta del Castillo
203. las casas de Santa Maria del Campo
204. Latacorno
205. la torre de piedra del Concejo
206. la villa de Plasencia
207. la Villa de Plasencia
208. linderosde
209. linderos dela una parte el de Rio de Almonte
210. Logrono
211. Logroño
212. Logrosio
213. los canchos
214. los Sierros
215. Lugarde Garces
216. Maerrnots
217. Magdalena
218. Majuelo
219. Malpartida
220. Maluense
221. Mardigal
222. Medellin
223. Mirabel
224. Moeda del Juglar

225. mojon
226. Molinos del Segura
227. Monasterio de San Yldefonso
228. Monroy
229. Montalban
230. Montanches
231. Monte
232. monte de la Perdiguera
233. Orellana
234. Osada
235. Paente de Nieblas
236. Pajarejos
237. Palacios
238. peñascal gran de de gijos
239. Perdiguera
240. Pesebres
241. Piedra
242. Plascencia
243. Plasencia
244. Plasencia
245. Plaza de Plasencia
246. Plazos
247. Portezuelo
248. Portigo
249. Pozo llamado Ollalla
250. pueblo de Meafadas
251. Pueblo de Meafadas
252. Pueblo de Meajadas
253. Pueblo San Esteban
254. Pueboo deXaraicejo
255. puerta del Castill
256. puerta del Castillo
257. puerta de Talabera
258. puerto de Castaño
259. Puerto del Caballo
260. Puerto del Castano
261. Puerto de Mirabese
262. Puerto de Mirabete
263. Quacos
264. Reportillo
265. Reyno
266. Ribera de Garces
267. Riobermejo
268. Rio Bermejo
269. Riobermejo Aldea
270. rio de Almon
271. rio de Almonte
272. Rio de Fejo
273. rio de Tajo
274. Rio de Tajo
275. Rio de Tejo

276. Rio Tajo
277. Rio Tejo
278. Riotortillo
279. rio Xerete
280. rio Xerete (Jerte)
281. Romango fallo
282. Río zortillo
283. Salamanca
284. San Esteban
285. SanJil
286. San Juan
287. San Juan de Turis
288. San Martin
289. San Pablo
290. San Pedro
291. Santa Iglesia de Roma
292. Santa Maria
293. Santa Maria del Campo
294. Santa Maria de Troxillo
295. Santa Mariala Catedral dela dicha
296. Santa Yglesia de Roma
297. Santiago
298. Santiago del Arrabal
299. Saucedilla
300. Segovia
301. Segura
302. Sevilla
303. Sienna
304. Sierro
305. Soria
306. Tajo
307. Tajo cerca dela Rodeznera
308. Tajo Jarrin
309. Talaban
310. Talabera
311. Tanrin
312. Tarrín
313. teuocillo
314. tierra de herederos del Esteban Sanchez
315. Tierras de San Lazaro
316. Toledo
317. torre de piedra del Concejo
318. Torre de Vigo
319. Tozo
320. Trofillo
321. Troxillo
322. Trujillo
323. Truxillo
324. Vade-Miguel
325. Valdefinete

326. Valdejinete
327. Valde Milana
328. Valdoliba
329. Valle de Doña Sol
330. Valle de Dona Sol
331. Valle de los Regueros
332. Valle de los Requeros
333. Valle de Milana
334. Vallejinete
335. Vallestero
336. Viña de Alfonso Alvarez
337. Viña dela Sierra
338. Viña del Caballo
339. Viña del Hombliguillo
340. Viña del Hospital
341. Viña de los Clerigos de Santa Maria
342. Viña de Rodrigo Diaz
343. Viña la Blanca
344. viña la Salgada
345. viñas
346. Villa de Bexar
347. Villa de Galinco
348. Villa de Galisteo
349. Villa de Galvisco
350. Villa deTroxillo
351. Villanba
352. Villarbos
353. Villarejo
354. Vina del Hospital
355. Xaar
356. Xaariz
357. Xaraiz
358. Xarcaido
359. Xerete (Jerte)
360. Yglesia Cardenal
361. Yglesia Catedral
362. Yglesia Catedral de Santa Maria de Plasencia
363. Yglesia de Plasencia
364. Yglesia de San Esteban
365. Yglesia de San Eueban
366. Yglesia de San Juan del Arrabal
367. Yglesia de San Pablo
368. Yglesia de Santa Maria
369. Yglesia de Santa Maria
370. Yglesia de San Vicente
371. Yusafe

A.3. Diccionario de Roles

A continuación se muestra el Diccionario de Roles que surge del análisis de las Actas Capitulares de la Catedral de Plasencia, es posible encontrar las denominaciones de los roles con las ocurrencias frecuentes (en sus diferentes variaciones).

1. Administracion
2. Albar Sanchez como su Fiador
3. Alcalde
4. Alfonso Arias Alcalde
5. Alfonso Ferrandez de Logro que son de los doce caballeros
6. Alfonso Ferrandez Racionero
7. Alfonso Martinez Criado de Gonzalo Ruiz
8. Alfonsus Ferrandez: Porcionarius
9. Algeiacil
10. Alguacil de qualquier ciudad o villa Eclesiastico
11. Alguacil de qualquier ciudad o villa Seglar
12. Alonso Gonzalez de Amusco Racionero dela dicha Yglesia
13. Andres Perez Tesorero dela dicha Yglesia
14. Arcipreste Racionero
15. Arediano de Plasencia e de Bexar
16. Arrendador
17. Arresidador
18. Bachiller
19. Bachiller en Leyes
20. Bano Gomez Racionero
21. Basco Gomez Racionero
22. Beneficiado dela dicha Yglesia
23. Beneficiados dela dicha Yglesia
24. Bicario de Soria
25. Caballeria
26. Cabildo y Beneficiados
27. Cabillo dela Eglesia Catedral de Plasencia
28. Canongia: Petrus Ferrandez
29. Canonicus = Alfonsus
30. Canonicus = Foribus
31. Canonicus = Joanes
32. Canonigos
33. Canonigos y Racioneros dela dicha Yglesia
34. Capellan Clerigo de Fragon
35. Capellanos
36. Chantre = Alfonsus
37. Christobal Sanchez Canonigo

38. Clerigo de Santiago del Arrabab
39. concejil
40. concejo que llaman Realenco
41. condera
42. Criado
43. Dean
44. debdor
45. Decanus Placentinus = Rodericus Archidia
46. Deudor
47. doctor
48. Doctor en Decreros
49. Doeton = Nicolaus Fernandez
50. Dona Benita e de Benito Perez hermano del dicho Alfonso Martin
51. Don Alfonso Garcia Arediano de traxillo
52. Don Gonzalo Gutierrez dela Calleja Chansre
53. Egidius Archidiaconus deTruxillo
54. el conde
55. el Doctor Garcia Lopez de Carvajal
56. el Mayordomo del Dean y Cabildo
57. El obispo Don Pedro
58. entregador de qualquier ciudad o villa Eclesiastico
59. entregador de qualquier ciudad o villa Seglar
60. Escrivano
61. Escuderos
62. Escuderos
63. Esno publico en Talabera
64. Ferran Alfonso Clerigo de San Esteban
65. Ferran Martinez Racioneros
66. Ferran Rodriguez e Juana Gonzalez su muger
67. Ferran Ximenez el Mozo
68. Fesarerarius = Roder
69. Fiador
70. fijo
71. Foribus Martinez de Villarba Canonogos
72. Garcia Ferrandez como principal deldon y Arrendador
73. Gerran Martinez de Boadilla Nortario publico Apostolico
74. GilMartínez Clerigo de Albaláde
75. Gonzalo Gutierrez dela Calleja Tesorero
76. Gonzalo Ruiz Clerigode Santiago del Arrabal

77. Gutierre Gonzalez del Barco es Notario publico
78. Herederos
79. Hermano
80. Juan Ferrandez de Betamos Doctor en Decreros
81. Juan Mateos Clerigo de Santiago
82. Juez de qualquier ciudad o villa Eclesiastico
83. Juez de qualquier ciudad o villa Seglar
84. Jurado de qualquier ciudad o villa Eclesiastico
85. Jurado de qualquier ciudad o villa Seglar
86. jurisdicion Eclesiastica
87. Jurisdicion Eclesiastica
88. La autoridad ordinaria
89. Labrador dela dicha Viña
90. la corte de Senor el Rey de qualquier Ciudad o Villa
91. Licenciado en Leyes
92. Licenciado en Leyes Arcediano de Plasensia e de Bexar
93. lluenio Senor el Rey
94. los dichos Senores
95. los Pobladores del Pueblo de Meajadas
96. Maestros
97. Mayor-domo
98. Mayordomo dela dicha Yglesia
99. mayordomos en la dicha ciudad
100. Merino de qualquier ciudad o villa Eclesiastico
101. Merino de qualquier ciudad o villa Seglar
102. Mozos del coro dela dicha Yglesia
103. muger
104. Nicolas Ferrandez Clerigo de Quacos
105. Notario
106. Notario de nuestro Señor el Eleito
107. Notario publico
108. Obispado
109. obispado eda dicha cuidad de Plasenciadado
110. Obispo de Plasencia
111. oficial
112. oficios decinales
113. otras personas singulares
114. pagador
115. Pagador
116. Pedro de Alfaro Criado de Andres Perez

117. Pedro Ferrandez Racionero dela dicha Yglesia
118. Pedro Gonzalez Notario
119. Pedro Gonzalez Racionerode la dicha Yglesia Notario Apostolico
120. Pedro Gonzalez Racionero Notario Apostolico de Yermo de Basco
121. Pedro Obispo
122. pobladores del Pueblo de Meafadas
123. Porcionarius Notario Apostolico-En la Ciudad de Plasencia
124. Porcionarius Pedro Ferradez
125. Porcionarun = Pesrus Ferrandez
126. Porcionarun Plasentinus = Sadovicces
127. Porcionus
128. Primo
129. Provisor
130. Provisor
131. publico Apostolico y testigos
132. Racionero deladicha Yglesia = Pedro Martinez
133. Racionero Juan Nuñez
134. RacioneroNotario Apostolico
135. residentes en la dicha Yglesia
136. Rexidores
137. Rui Garcia Canonigo en la dha Yglesia Cardenal
138. Rui Garcia de Salamanca Bachiller en Leyes
139. Señor Provisor
140. Señor Rui Garcia Vicario
141. sobrino
142. Statuto
143. Su Senoria
144. Tendero
145. Tesorero
146. testamentarios
147. Testigos
148. ullí Moro Carpintero
149. Vecinos
150. vecinos de Medellin
151. vecinos de Plasencia
152. vecinos de Talabera
153. Venerables Senores Don Alvaro de Salazan Dean, e Don Rodrigo de Carvajal
154. Vicario
155. Villalovos Canonigo de la dicha Yglesia

A.4. Diccionario de Roles - general de la época

A continuación se muestra el Diccionario de roles general de la época que surge del análisis de las Actas Capitulares de la Catedral de Plasencia, es posible encontrar las denominaciones de los roles con las ocurrencias frecuentes (en sus diferentes variaciones).

1. Administracion
2. Alcalde
3. Algeiacil
4. Alguacil Eclesiastico
5. Alguacil Seglar
6. Arcipreste Racionero
7. Arediano
8. Arrendador
9. Arresidador
10. Bachiller
11. Bachiller en Leyes
12. Beneficiado dela dicha Yglesia
13. Bicario de Soria
14. Caballeria
15. caballeros
16. Cabildo y Beneficiados
17. Cabillo dela Eglesia Catedral de Plasencia
18. Canonicus
19. Canonigo de la dicha Yglesia
20. Canonigos
21. Canonigos y Racioneros dela dicha Yglesia
22. Capellan Clerigo de Fragon
23. Capellanos
24. Carpintero
25. Chantre
26. Clerigo
27. Clerigo de Albalá
28. Clerigo de Quacos
29. Clerigo de San Esteban
30. Clerigo de Santiago
31. concejil
32. concejo que llaman Realenco
33. condesa
34. Criado
35. Dean
36. debdor
37. Decanus Placentinus
38. Deudor
39. doctor
40. Doeton
41. el conde

42. el Mayordomo del Dean y Cabildo
43. entregador Eclesiastico
44. entregador Seglar
45. Escuderos
46. Esno publico en Talabera
47. Fiador
48. fijo
49. Gonzalo Ruiz Clerigode Santiago del Arrabal
50. Herederos
51. Juez Eclesiastico
52. Juez Seglar
53. Jurado Eclesiastico
54. Jurado Seglar
55. Jurisdicion Eclesiastica
56. Labrador dela dicha Viña
57. la corte de Senor el Rey
58. Licenciado en Leyes
59. lluenio Senor el Rey
60. los dichos Senores
61. los Pobladores del Pueblo de Meajadas
62. Maestros
63. Mayor-domo
64. Mayordomos
65. Merino Eclesiastico
66. Merino Seglar
67. Mozos del coro dela dicha Yglesia
68. muger
69. Nortario publico Apostolico
70. Notario
71. Notario Apostolico
72. Notario de nuestro Señor el Eleito
73. Notario publico
74. obispado eda dicha cuidad de Plasenciadado
75. Obispo de Plasencia
76. oficial
77. oficios decinales
78. otras personas singulares
79. pagador
80. Pagador
81. Pobladores
82. pobladores del Pueblo de Meafadas
83. Porcionarius
84. Porcionarius Notario Apostolico-En la Ciudad de Plasencia
85. Porcionarun Plasentinus
86. Porcionus
87. Primo

88. Provisor
89. publico Apostolico y testigos
90. Racionero dela dicha Yglesia
91. Racionero deladicha Yglesia
92. RacioneroNotario Apostolico
93. residentes en la dicha Yglesia
94. Rexidores
95. Señor Provisor
96. Statuto
97. Su Senoria
98. Tendero
99. Tesorero
100. Tesorero dela dicha Yglesia
101. testamentarios
102. Testigos
103. Vecinos
104. vecinos de Medellin
105. vecinos de Plasencia
106. vecinos de Plasencia
107. vecinos de Talabera
108. Venerables Senores
109. Vicario

Apéndice B

Código Python de limpieza de las transcripciones

El siguiente código escrito en **python 3** es el usado para la limpieza y separación de las transcripciones, de tal manera que sea posible realizar todos los trabajos siguientes:

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
Created on Tue Nov 1 09:59:39 2016

@author: Juan Sebastian Martinez Serna,
         Diana Maria del Pilar Socha Diaz
"""

import pandas as pd
import ast

# s = "ferdand\\u00E9"
# prind s.decode('unicode-escape')
# >>> fernandé

def deleteTag(s):
    '''
    Delete the first HTML tag of the string,
    all tags starts with '<' and ends with '>'
```

```
21
22      PARAMETERS
23      ----------
24      s : STRING
25          string to delete the first HTML tag
26
27      RETURNS
28      -------
29      str : string without the first HTML tag
30      None: if 's' no contains a HTML tag
31
32      EXAMPLES
33      --------
34      in["Hello<tag>world"]
35      out["Hello world"]
36
37      in["Hello <tag> again <tag2>"]
38      out["Hello  again <tag2>"]
39
40      in["Hello without HTML tags"]
41      out[None]
42      '''
43      startsTag = s.find('<'); endsTag = s.find('>')
44      if startsTag == -1 or endsTag == -1: return None
45      elif len(s) == endsTag: return s[:startsTag]
46      else: return s[:startsTag] + ' ' + s[endsTag+1:]
47
48  def deleteHTMLTags(s):
49      '''
50      Delete all HTML tags
51
52      PARAMETERS
53      ----------
54      s : STRING
55          string to delete all HTML tags
56
57      RETURNS
```

```
58      -------
59      str : string without HTML tags
60
61      EXAMPLES
62      --------
63      in["Hello<tag>world"]
64      out["Hello world"]
65
66      in["Hello <tag> again <tag2>"]
67      out["Hello   again"]
68
69      in["Hello without HTML tags"]
70      out["Hello without HTML tags"]
71      '''
72      ok = True
73      while ok:
74          aux = deleteTag(s)
75          if aux is None: ok = False
76          else: s = aux
77
78      return s.strip()
79
80
81  def changeCodeToSpace(s):
82      '''
83      Change all HTML codes ' ' for a single
84      space ' ', if the string has no a code, do
85      nothing, i.e. returns the same string
86
87      PARAMETERS
88      ----------
89      s : STRING
90          string to delete all HTML codes ' '
91
92      RETURNS
93      -------
94      str : string without HTML codes ' '
```

```
95
96      EXAMPLES
97      --------
98      in["Hello world"]
99      out["Hello world"]
100
101     in["Hello   again  "]
102     out["Hello   again"]
103
104     in["Hello without HTML codes"]
105     out["Hello without HTML codes"]
106     '''
107     spaceCode = ' '
108     return s.replace(spaceCode, ' ').strip()
109
110
111 def findCode(s):
112     '''
113     Find and return the first code of manuscripts.
114     CODE Struct -> B#[A-Z]###
115
116     PARAMETERS
117     ----------
118     s : STRING
119         string to find code
120
121     RETURNS
122     -------
123     str : string with the first code
124
125     EXAMPLES
126     --------
127     in["Hello B2C59 world"]
128     out["B2C59"]
129
130     in["Hello B5N100 again B8U31"]
131     out["B5N100"]
```

```
132
133     in["Hello without codes"]
134     out[""]
135     '''
136     startCode = 'B'
137     code = ''
138     s = s.find(startCode)
139
140     if s == -1: return code # If 'B' not exist
141     if  s[s+1].isdigit(): # Verify B#
142         # Verify B#[A-Z]#
143         if s[s+2].isalpha() and s[s+3].isdigit():
144             code = s[s:s+4]
145             # Verify B#[A-Z]##
146             if len(s) > s+4 and s[s+4].isdigit():
147                 code = s[s:s+5]
148                 # Verify B#[A-Z]###
149                 if len(s) > s+5 and s[s+5].isdigit():
150                     code = s[s:s+6]
151     else:
152         code = findCode(s[s + 1:])
153
154     return code.strip()
155
156
157 def deleteDoubleSpace(s):
158     '''
159     Delete all double space from the string 's'
160     and put a single space
161
162     PARAMETERS
163     ----------
164     s : STRING
165         string to delete all double space
166
167     RETURNS
168     -------
```

```
169         str : string without any double space
170
171         EXAMPLES
172         --------
173         in["Hello world"]
174         out["Hello world"]
175
176         in["Hello   again "]
177         out["Hello again"]
178
179         in["Hello without double space"]
180         out["Hello without double space"]
181         '''
182         while s.find('  ') != -1: s = s.replace('  ', ' ')
183         return s.strip()
184
185
186 def deleteNegativeSymbol(s):
187         '''
188         Find and remove all negative symbol
189         with or without a space
190
191         PARAMETERS
192         ----------
193         s : STRING
194             string to delete all negative symbol
195             with or without a space
196
197         RETURNS
198         -------
199         str : string without any negative symbol
200
201         EXAMPLES
202         --------
203         in["Hello- world"]
204         out["Helloworld"]
205
```

```
206         in["Hello-again "]
207         out["Helloagain"]
208
209         in["Hello without negative symbol"]
210         out["Hello without negative symbol"]
211         '''
212         s = s.replace('- ', '').replace('-', '')
213         return s.replace('_','').strip()
214
215
216     def extractTranscriptions(s, transcriptions=[]):
217         '''
218         Build a list of tuples with 2 items by touple
219         from string 's', the first item is the code of
220         transcription and the second item is the
221         transcription
222
223         PARAMETERS
224         ----------
225         s : STRING
226             string that contains the transcriptions
227             and its codes
228         transcriptions : LIST of TUPLES
229             have a tuple with two elements, the
230             first is the code, and the second
231             is the transcription
232
233         RETURNS
234         -------
235         l : a list that contains the transcriptions
236             with its codes
237
238         EXAMPLES
239         --------
240         in["asd B3C4 asdoaijfia"]
241         out[[('B3C4', 'asdoaijfia')]]
242
```

```
    in["B3F5 asfsdf B2U3 sadsd"]
    out[[('B3F5','asdsdf'), ('B2U3','sadfd')]]

    in["String without a code"]
    out[[]]
    '''
    code = findCode(s)

    if code != '':
        indexCode = s.find(code)
        # new s without code
        s = s[indexCode+len(code):].strip()
        code2 = findCode(s)

        if code2 != '':
            indexCode2 = s.find(code2)
            aux = s[:indexCode2]
            transcriptions.append((code, aux.strip()))
            # new s without previous msnuscript
            s = s[indexCode2:]
        else:
            transcriptions.append((code, s.strip()))
    else:
        return transcriptions

    return extractTranscriptions(s, transcriptions)


def loadExcel(path, sheetname='Culled Data'):
    '''
    Load a Excel file

    PARAMETERS
    ----------
    path: STRING
        string with the Excel path to load
    sheetname: STRING
```

```
280         the sheetname to load
281
282     RETURNS
283     -------
284     xl : DataFrame
285         A DataFrame that contains all
286         values from excel file
287     '''
288     xl = pd.ExcelFile(path).parse(sheetname=sheetname)
289     return pd.DataFrame(xl)
290
291
292 def loadCSV(path):
293     '''
294     Load a Excel file
295
296     PARAMETERS
297     ----------
298     path: STRING
299         string with the Excel path to load
300     sheetname: STRING
301         the sheetname to load
302
303     RETURNS
304     -------
305     xl : DataFrame
306         A DataFrame that contains all
307         values from excel file
308     '''
309     cvs = pd.read_csv(path, encoding='utf-8')
310     return pd.DataFrame(cvs)
311
312
313 def saveCSV(df, path, name='Transcriptions.csv'):
314     '''
315     Create a CSV file with the transcriptions
316     and its codes
```

```
317
318     PARAMETERS
319     ----------
320     df : DataFrame
321         contains the transcriptions with
322         its codes to save
323     path : STRING
324         the path where to save the file
325     name : STRING
326         the file name
327     '''
328     df.to_csv(path + name)
329
330
331 def createDataFrame(l, colCode='Codes',
332                     colTranscriptions='Transcriptions'):
333     '''
334     Create a DataFrame from a list of tuples
335     that contains the transcription with its
336     respective code
337
338     PARAMETERS
339     ----------
340     l : LIST of TUPLES
341         contains the transcription with its
342         respective code
343     colCode : STRING
344         column name for codes
345     colTranscriptions : STRING
346         column name for transcriptions
347
348     RETURNS
349     -------
350     df : DataFrame
351         The respective DataFrame with
352         names and values by column
353     '''
```

```
354     columnCodes = []
355     columnTrans = []
356 
357     for (cod, trans) in l:
358         columnCodes.append(cod)
359         columnTrans.append(trans)
360 
361     df = pd.DataFrame()
362     df[colCode] = columnCodes
363     df[colTranscriptions] = columnTrans
364 
365     return df
366 
367 
368 def deleteTabs(s):
369     '''
370     Delete all tabs from the string and
371     replaced by a single space
372 
373     PARAMETERS
374     ----------
375     s : STRING
376         the string to remove all tabs
377 
378     RETURNS
379     -------
380     s : STRING
381         the string without any tab
382 
383     EXAMPLES
384     --------
385     in['Hello\t\tWorld\t']
386     out['Hello  World']
387 
388     in['\tHello \t\t\tWorld']
389     out['Hello    world']
390 
```

```
391         in['Hello again']
392         out['Hello again']
393         '''
394         return s.replace('\t', ' ')
395 
396 
397 def dataClean(s, e=[]):
398         '''
399         Clean the string, i.e. Remove HTML codes,
400         double space and negative symbols
401         and delete all elements indicates
402 
403         PARAMETERS
404         ----------
405         s : STRING
406             string without clean
407         e : LIST
408             list of elements to remove of the string
409 
410         RETURNS
411         -------
412         s : STRING
413             a clean string
414 
415         EXAMPLES
416         --------
417         in["Hola <asd> wo- rld"]
418         out["Hola world"]
419 
420         in["Heloo my <ds>  <sfe> world"]
421         out["Heloo my world"]
422 
423         in["Hello again again"]
424         out["Hello again again"]
425         '''
426         for rmv in e: s = s.replace(rmv, ' ')
427 
```

```
428     s = deleteTabs(s)
429     s = changeCodeToSpace(s)
430     s = deleteDoubleSpace(s)
431     s = deleteNegativeSymbol(s)
432     s = deleteHTMLTags(s)
433     if s.find('  ') != -1 : s = deleteDoubleSpace(s)
434 
435     return s
436 
437 #================================================
438 
439 def main():
440 
441     pathLoad  = input('Excel file : ')
442     sheetName = input('Sheet name : ')
443     colName   = input('Column name to work : ')
444     pathSave  = input('Path to save work files : ')
445     path2Del  = input('Path to text file to remove elements: ')
446 
447     if sheetName == '': xls = loadExcel(pathLoad)
448     else: xls = loadExcel(pathLoad, sheetName)
449 
450     li = list(xls[colName])
451     t = []
452 
453     listFailed = []
454 
455     with open(path2Del, 'r') as f:
456 
457         elements2del = f.readlines()
458 
459         for s in li:
460             try:
461                 s = str(s).replace('"', '')
462                 s = ast.literal_eval('"' + s + '"')
463                 aux = extractTranscriptions(
464                         dataClean(s, elements2del))
```

```
465                 t += [e for e in aux if e not in t]
466             except:
467                 listFailed.append(s)
468 
469     df = createDataFrame(t)
470     saveCSV(df, pathSave)
471 
472     if len(listFailed) != 0:
473         saveCSV(pd.DataFrame(listFailed),
474                 pathSave, 'Failed.csv')
475 
476 main()
```

Apéndice C

Código Python para las correlaciones

El siguiente código escrito en **python 3** es el usado para la generación de todas las correlaciones existentes según el orden de entrada de los diccionarios

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
Created on Sun Jan 09 21:02:37 2017

@author: Diana Maria del Pilar Socha Diaz
         Juan Sebastian Martinez Serna
"""
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

def loadExcel(path, sheetname='Transcriptions'):
    '''
    Load a Excel file
    PARAMETERS
    ----------
    path: STRING
        string with the Excel path to load
    sheetname: STRING
        the sheetname to load
```

```
22      RETURNS
23      -------
24      xl : DataFrame
25          A DataFrame that contains all values from excel file
26      '''
27      xl = pd.ExcelFile(path).parse(sheetname=sheetname)
28      return pd.DataFrame(xl)
29
30  def saveCSV(df, path, name='cor.csv'):
31      '''
32      Create a CSV file with the cor
33
34      PARAMETERS
35      ----------
36      df : DataFrame
37          contains the transcriptions with its codes to save
38      path : STRING
39          the path where to save the file
40      name : STRING
41          the file name
42      '''
43      df.to_csv(path + name)
44
45  def main():
46
47      pathLoad  = input('Excel file : ')
48      sheetName = input('Sheet name : ')
49      dict1     = input('Dictionary 1 : ')
50      dict2     = input('Dictionary 2 : ')
51      colName   = input('Column name : ')
52      pathSave  = input('CSV to save : ')
53
54      trans = loadExcel(pathLoad, sheetName)
55
56      with open(dict1, 'r', encoding='latin-1') as d1:
57          with open(dict2, 'r', encoding='latin-1') as d2:
58              ld1 = d1.readlines()
```

```
59              ld2 = d2.readlines()
60
61              cor = {}
62
63              transVal = trans[colName]
64              for e in transVal.iteritems():
65                  s = str(e[1])
66                  for ed1 in ld1:
67                      if s.find(ed1) != -1:
68                          for ed2 in ld2:
69                              try:
70                                  cor[ed1][ed2] += s.count(ed2)
71                              except:
72                                  try:
73                                      cor[ed1][ed2] = 0
74                                  except:
75                                      cor[ed1] = {ed2 : 0}
76
77              valores = []
78
79              for d1 in dict1:
80                  row = []
81                  for d2 in dict2:
82                      try: row.append(cor[d1][d2])
83                      except: True
84                  valores.append(row)
85
86              df = pd.DataFrame(data=valores,
87                                columns=dict1, index=dict2)
88              saveCSV(df, pathSave)
89              print('OK')
90
91  main()
```

Siglas

D

DB Bases de datos (Databases). 25

E

ENHG Early New High German. 44

G

GIS Sistema de información geográfico (Geographic Information System). 20

I

IA Inteligencia artificial (Artificial Intelligence). 25

IE Extracción de información (Information Extraction). *Glosario:* Information Extraction

IE Information Extraction. *Glosario:* Information Extraction

IR Recuperación de información (Information Retrieval). 46, 48, *Glosario:* Search and Information Retrieval

K

KNIME Konstanz Information Miner. 57, *Glosario:* KNIME

M

MHG Middle High German. 44

MOOC Curso masivo abierto y en linea (Masive Open Online Course). 14, 17, 66

N

NHG New High German. 44

NLP Procesamiento de lenguaje natural (Natural Language Processing). *Glosario:* Natural Language Processing

NLP Natural Language Processing. *Glosario:* Natural Language Processing

O

OCR Optical Character Recognition. 47, 48, *Glosario:* OCR

OHG Old High German. 44

P

PoS Discurso de la palabra (Part of Speech). 26

R

RCCP Revealing Cooperation and Conflict Project. 6, 7, 12, 13, 16, 18, 20, 22

S

SQL Structured Query Language. 48

T

TEI Electronic Text Encoding and Interchange. 43, *Glosario:* Electronic Text Encoding and Interchange

U

UTF-8 8-bit Unicode Transformation Format. *Glosario:* UTF-8

X

XML Extensible Markup Language. 43, 58, *Glosario:* Extensible Markup Language

Glosario

A

Actas Capitulares de la Catedral de Plasencia Son los documentos donde eran almacenados todos los datos del día a día sobre las transacciones de los diferentes negocios que se realizaban, estos eran guardados por el *Cabildo de la Catedral de Plasencia*. Las actas están comprendidas en los años 1399 – 1527. 4, 7, 10, 11, 13–20, 22, 26, 66, 75–77, 92, 100, 104

C

Concept extraction Agrupación de palabras y frases en grupos semánticamente similares [11]. 24, 26

D

datos estructurados Son todos aquellos datos que se encuentran organizados y almacenados de tal manera que cumplen una estructura especifica,es decir, son datos bien definidos, por ejemplo, las fechas, los números de teléfono, entre otros. 24

datos no estructurados Son todos aquellos datos que no se encuentran organizados de tal manera que no hagan uso de una estructura especifica, por ejemplo, los nombres de las personas, documentos de texto, comentarios, entre otros. 24

diccionario de lugares Repertorio ordenado alfabéticamente con todos los lugares nombrados en los documentos. 23, 77

diccionario de nombres Repertorio ordenado por familias con todos los nombres de las personas encontradas en los documentos. 23, 77

diccionario de roles Repertorio ordenado por importancia, el más demandado, con todos los roles encontrados en los documentos, esto junto con los nombres de las personas que tenían dichos roles. 23, 77

Document classification Agrupación y clasificación de los documentos y/o fragmentos mediante el uso de algoritmos de clasificación de la minería de datos basados en modelos de pruebas . 24, 26

Document clustering Agrupación y categorización de documentos y/o fragmentos haciendo uso de diferentes algoritmos de clasificación por agrupamiento de la minería de datos . 24, 26

E

Electronic Text Encoding and Interchange Estandar internacional de facto que tiene como objetivo hacer más sencilla la representación en formato digital de textos y literatura cuando se trata de aquella con fines investigativos o para la educación en línea. TEI es lanzado por la Association for Computers and the Humanities en el año 1987. TEI P4 es la versión actual desarrollada para un correcto funcionamiento y soporte sobre lenguaje XML . 31

Extensible Markup Language Es un meta-lenguaje usado para la especificación de lenguajes de marcado, este permite definir etiquetas para la descripción de un conjunto de datos . 48, 49

I

Information Extraction Identificación y extracción de los diferentes hechos y relaciones encontradas dentro de los documentos de texto, el proceso consiste en generar datos estructurados a partir de los no estructurados . 24, 26

K

KNIME Plataforma de código abierto de fácil acceso orientada a posibilitar la Minería de Datos. Fuente de ayuda que dentro de sus principales funcionalidades permite encontrar información oculta en grandes volúmenes de datos además de hallar patrones de comportamiento de datos, hacer predicciones futuras y visualización . 5, 22, 37, 57, 58, 72

L

linaje Línea de descendientes o antepasados de una persona, muy importante para el continente Europeo debido a que en dicho contexto el linaje se refería además a la sucesión de propiedades, títulos, derechos, entre otros. Para la investigación, corresponde entonces también a la delegación de tareas o trabajos de una persona dentro de una sociedad, según fuera el de la línea de sus antepasados. 16

N

Natural Language Processing También conocido como linguistica computacional, es un campo del aprendizaje automatizado o aprendizaje de máquina con foco en las relaciones entre el lenguaje humano y el lenguaje de máquina . 24, 26

O

ontología En una ontología (informática) se definen formalmente relaciones, propiedades y tipos entre entidades que existen en un "corpus". Se representa el conocimiento, se aplica sobre el mismo un método para representar formalmente los datos y finalmente se obtienen relaciones entre conceptos. Como principales objetivos está el limitar la complejidad y organizar información mas eficazmente . 32

S

Search and Information Retrieval Almacenamiento y recuperación de documentos de texto, proceso utilizado dentro de los motores de búsqueda, búsqueda de palabras claves. También es aplicado en los procesos de extracción de resúmenes de libros . 24, 26, 42, 43

U

UTF-8 Se entiende como un sistema de codificación de caracteres, la longitud de dicha codificación es variable. Es posible representar cualquier caracter Unicode siendo compatible con ASCII. UTF-8 hace uso de grupos de bytes para dicha compatibilidad pueda ser posible y aplicable en la mayoría de lenguajes del mundo . 43

W

Web mining Aplicación de las diferentes técnicas de minería de datos, minería de texto y análisis de texto en diferentes páginas de Internet . 24, 26

Bibliografía

[1] Roger Louis Martínez-Dávila. Revealing cooperation and conflict. http://revealingcooperationandconflict.com/the-project/.

[2] Dr. Roger Louis Martínez-Dávila. Explorando las humanidades digitales. http://www.rogerlouismartinez.com/?page_id=3133&lang=es.

[3] Martin Baumeister y Bernardo Teuber. La obra de américo castro y la españa de las tres culturas. pages 82–95, 2010.

[4] Roger Louis Martínez-Dávila. Historical background on medieval spain and plasencia. http://revealingcooperationandconflict.com/historical-background-on-medieval-spain-and-plasencia/.

[5] Lucía Brage Martínez. ENTREVISTA AL DR. ROGER LOUIS MARTÍNEZ-DÁVILA. *ArtyHum, Revista de Artes y Humanidades*, pages 178–185, 2014.

[6] Archivo Municipal de Plasencia. Serie 001 - libros de actas capitulares. http://archivo.plasencia.es/index.php/libros-de-actas-capitulares.

[7] Roger Louis Martínez-Dávila. The cathedral of plasencia's actas capitulares (the chapter acts). http://revealingcooperationandconflict.com/about-the-actas-capitulares-the-chapter-acts/.

[8] The Deciphering Secrets website. Welcome to deciphering secrets. http://decipheringsecrets.net/.

[9] Roger L. Martínez-Dávila. Ds: Unlocking the manuscripts of medieval spain. https://www.decipheringsecrets.com/portfolio/ds-unlocking-the-manuscripts-of-medieval-spain/?lang=es.

[10] Ah-Hwee Tan et al. Text mining: The state of the art and the challenges. In *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, volume 8, pages 65–70, 1999.

[11] Gary Miner, Dursun Delen, John Elder, Bob NIsbet, Thomas Hill, and Andrew Fast. *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press, 2012.

[12] University of New Brunswick. How google thinks. http://media.lib.unb.ca/research/video/2Google@unb.pdf, 2010.

[13] Sašo Džeroski and Tomaž Erjavec. Learning to lemmatise slovene words. In *Cussens and S. Džeroski, Learning Language in Logic, Number 1925 in Lecture notes in artificial intelligence*, pages 69–88. Springer-Verlag, 2000.

[14] Andrea Gesmundo and Tanja Samardžić. Lemmatisation as a tagging task.

[15] Arianna Ciula, Paul Spence, and José Miguel Vieira. Expressing complex associations in medieval historical documents: the henry iii fine rolls project. In *Literary and Linguistic Computing*, volume 23, pages 311–325, 2008.

[16] Fuminori Kimura, Takahiko Osaki, Taro Tezuka, and Akira Maeda. Visualization of relationships among historical persons from japanese historical documents. In *Literary and Linguistic Computing*, volume 28, pages 271–278, 2013.

[17] The JUNG Framework Development Team. JUNG Java Universal Network/Graph Framework - Overview. http://jung.sourceforge.net/.

[18] Andreas Hauser, Markus Heller, Elisabeth Leiss, Klaus U. Schulz, and Christiane Wanzeck. Information access to historical documents from the early new high german period, 2007.

[19] PROEL. Lengua alemana. http://www.proel.org/index.php?pagina=mundo/indoeuro/germanico/germanooc/aleman.

[20] Loes Braun, Floris Wiesman, and Ida Sprinkhuizen-Kuyper. Information retrieval from historical corpora, 2002.

[21] Rada Mihalcea Tze-I Yang, Andrew J. Torget. Topic modeling on historical newspapers, 2011.

[22] Andrew Kachites McCallum. Mallet: A machine learning for language toolkit, 2002. http://mallet.cs.umass.edu.

[23] KNIME ORG. Knime analytics platform. https://www.knime.org/knime-analytics-platform.

[24] RapidMiner. Rapidminer, open sourcez data science platform. https://rapidminer.com/products.

[25] David Norris. Rapidminer - a potential game changer. http://www.bloorresearch.com/analysis/rapidminer-a-potential-game-changer, 2013.

[26] The R Foundation. What is r? https://www.r-project.org/about.html.

[27] The University of Waikato. Weka 3: Data mining software in java. http://www.cs.waikato.ac.nz/ ml/weka/index.html.

[28] Ching man Au Yeung. Studying how the past is remembered: Towards computational history through large scale text mining. In *In Proc. CIKM*, pages 1231–1240, 2011.

[29] Daniel McDonald. A text mining analysis of religious texts. In *The Journal of Business Inquiry*, volume 13, pages 27–47, 2014.

[30] Microsoft Developer Network. Conceptos de minería de datos. https://msdn.microsoft.com/es-es/library/ms174949(v=sql.120).aspx.

[31] Un Yong Nahm. Text mining with information extraction. In *AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases*, pages 5–6, 60–67, 2002.

[32] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. ACM press, 1999.

[33] Barry de Ville. *Integrated Bussines Intelligence for e-Comerce and Knowledge Management*. Digital Press, 2001.

[34] Universidad Carlos III de Madrid. EXPERIENCED FELLOW - Roger Louis Martínez-Dávila. http://www.uc3m.es/ss/Satellite/UC3MInstitucional/es/FormularioTextoDosColumnas/1371215847808/.

[35] Inc. RStudio. Rstudio ide features. https://www.rstudio.com/products/rstudio/features/.

[36] The Text Encoding Initiative Consortium. Tei: Text encoding initiative. http://www.tei-c.org/index.xml.

[37] J. Tomás Nogales Flores. Tei: Text encoding initiative. http://www.bib.uc3m.es/ nogales/cursos/tei.html.

[38] Margaret Rouse. Xml (extensible markup language). http://searchsoa.techtarget.com/definition/XML.

[39] Asunción Gómez-Pérez PhD, MSc, MBA; Mariano Fernández-López PhD, MSc; Oscar Corcho MSc. *Ontological Engineering with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. Springer, first edition, 2004.

[40] Inc. Unicode. General questions, relating to utf or encoding form - what is the definition of utf-8? http://unicode.org/faq/utf_bom.html#utf8-1.

[41] Moodle. Utf-8. https://docs.moodle.org/all/es/UTF-8.

[42] BIBLIOTECA DIGITAL GREENSTONE. Biblioteca digital greenstone del papel a la colecciÓn: Chapter 3 ocr: reconocimiento óptico de caracteres. http://www.greenstone.org/manuals/gsdl2/es/html/Chapter_ocr.htm.