Minería de Texto histórica, colaboración al proyecto "Revealing Cooperation and Conflict Project"

Diana María del Pilar Socha Díaz, *Escuela Colombiana de Ingeniería Julio Garavito*,
Juan Sebastián Martínez Serna, *Escuela Colombiana de Ingeniería Julio Garavito*,
Cristian Camilo Medina Mosquera, *Escuela Colombiana de Ingeniería Julio Garavito*

***Abstract*—El creciente interés por los documentos históricos, su análisis y el poder compartirlos en formatos digitales, ha estado en la mira tanto de las humanidades como de la ciencia y específicamente en la informática, desde algunas ramas como la minería de texto, minería de datos o el machine learning. Es importante preservar dichos documentos históricos tanto para la investigación como para su uso en ambientes académicos para que así sean realmente provechosos y trasciendan más allá de una biblioteca.**

Para éste proyecto, titulado *"Minería de Texto histórica, colaboración al proyecto 'Revealing Cooperation and Conflict Project'"*, se analizaron varios problemas incluyendo como uno de los principales, el hecho de que los documentos base de investigación, se encuentran escritos en lenguaje histórico y no moderno por lo que se dificulta el poner en práctica algúnas técnicas de Minería de Texto.

Siendo el proyecto *Revealing Cooperation and Conflict Project* de carácter abierto y participativo se pretende proponer una metodología de trabajo basada en técnicas de minería de texto con el fin de descubrir información no evidente y relevante para el desarrollo de la investigación que se lleva a cabo.

***Keywords*—colaborativo, corpus, diccionarios, documentos antiguos, MOOC, minería de texto, transcripciones, minería de datos**

I. Revealing Cooperation and Conflict Project

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 Plasencia, 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 (de ahora en adelante como Actas) [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, lingistas 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]

A. Actas Capitulares de la Catedral de Plasencia

Las Actas *"son los documentos más representativos de la gestión y administración de los concejos medievales castellanos"*[3] que son pertenecientes a los siglos desde el **siglo XIV hasta el siglo XVI** [4], 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

Dr. Ignacio Prez Vlez Director del proyecto y profesor investigador, *Escuela Colombiana de Ingeniería Julio Garavito*

Dr. Roger Martínez-Dávila co-director del proyecto

información sobre la economía y gestión de la ciudad.

La periodicidad que presentaban las Actas era semanal [3], 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"* [3]

Estos documentos tienen el aspecto más crítico del proyecto dirigido por el Dr. Roger Louis Martínez-Dávila puesto que *"son una fuente inestimable para la reconstrucción de los diversos aspectos de la vida de la ciudad de Plasencia"*[3]. En la imagen de la figura 1 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*

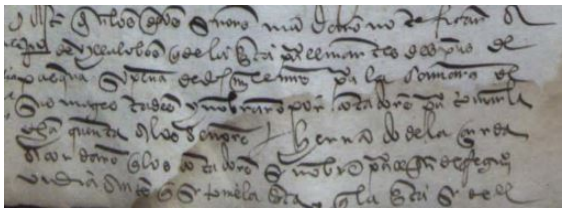

Fig. 1. Fragmento de las Actas Capitulares de la Catedral de Plasencia

II. Deciphering Secrets: Unlocking the Manuscripts of Medieval Spain

Deciphering Secrets, 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.

Deciphering Secrets comienza en el año 2014, como un curso en línea, masivo y abierto (MOOC), impartido por el Dr. Roger Louis 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 RCCP. [5]

Además de lo anterior y, *"lo que es más importante, los estudiantes contribuyen a un esfuerzo académico internacional ayudando a transcribir manuscritos"*. [5]

Los objetivos especficos del curso, descritos por los profesores son [6]:

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

III. Proceso del trabajo

La minería de texto también es conocida como minería de datos de texto [7] 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 [7].

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 [8], si bien estos algoritmos pueden llegar a ser completamente diferentes, en el fondo tienen el mismo objetivo, *"realizar una transformación de texto a números"* [8], esto con la finalidad de poderlos aplicar completamente a todo el documento.

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

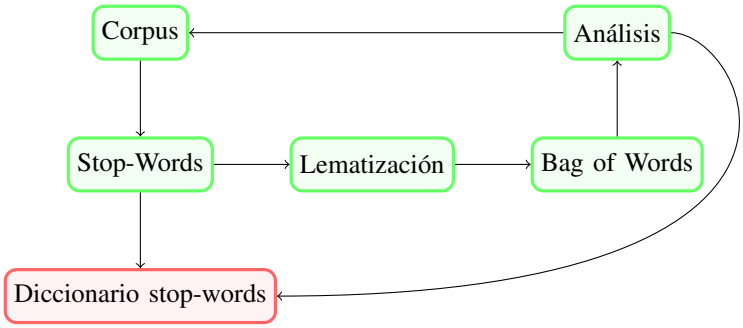

Fig. 2. Diagrama del ciclo de trabajo

IV. El conjunto de datos

La base de datos contenía todos los datos correspondientes al MOOC, ya antes mencionado, 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.

A. *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 3, en su respectiva carpeta con la identificación del mismo dentro del curso, como se ve en la imagen 4, dentro de esta se encuentran las calificaciones dadas por los otros dos o tres estudiantes del MOOC, 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.

Fig. 3. Carpetas con todas las actividades del curso

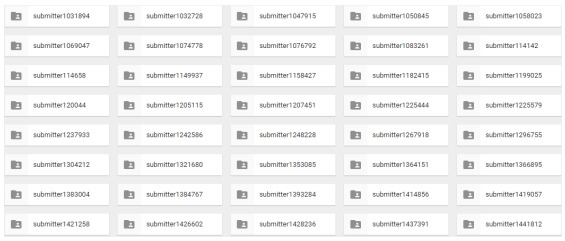

Fig. 4. Carpetas de los estudiantes de cada actividad

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 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 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.

Fig. 5. Transcripciones digitalizadas "sucias"

Para poder realizar una limpieza del documento y filtrar cada transcripción es necesario comprender primero la estructura de las transcripciones, 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 de realizar la transcripción. Dichas transcripciones cuentan con el mismo problema antes mostrado en la imagen 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.

V. Trabajo

A. 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:

```
def deleteTag(s):
    startsTag = s.find('<'); endsTag = s.find('>')
    if startsTag == -1 or endsTag == -1: return None
    elif len(s) == endsTag: return s[:startsTag]
    else: return s[:startsTag] + ' ' + s[endsTag+1:]

def deleteHTMLTags(s):
    ok = True
    while ok:
        aux = deleteTag(s)
        if aux is None: ok = False
        else: s = aux
    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 = ''
    strtCode = s.find(startCode)

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

        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)
```

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:

```
t = []
...
aux = extractTranscriptions(dataClean(s, e2del))
t += [e for e in aux if e not in t]
...
```

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 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 agora corre, ó dela moneda que corriere al tiempo delas paga aquí 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 opi
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 resciviria 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 Berdug

Fig. 6. Documento obtenido del proceso de la limpieza

B. *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 y en la imagen 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.

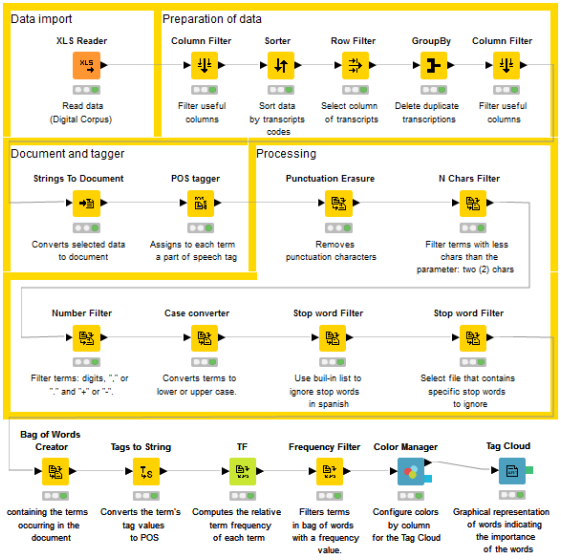

Fig. 7. Flujo en KNIME de Tag Cloud para limpieza y extracción

C. *Correlaciones*

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

Fig. 8. Tag Cloud para limpieza y extracción

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

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

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 transcripcin que al menos un elemento de A es mediante el siguiente cdigo.

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

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

VI. 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. Roger Louis 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.

REFERENCES

[1] R. L. Martnez-Dvila, "Revealing cooperation and conflict project," http://revealingcooperationandconflict.com/the-project/.

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

[3] A. M. de Plasencia, "Serie 001 - libros de actas capitulares," http://archivo.plasencia.es/index.php/libros-de-actas-capitulares.

[4] R. L. Martnez-Dvila, "The cathedral of plasencias actas capitulares (the chapter acts)," http://revealingcooperationandconflict.com/about-the-actas-capitulares-the-chapter-acts/.

[5] T. D. S. website, "Welcome to deciphering secrets," http://decipheringsecrets.net/.

[6] R. L. Martnez-Dvila, "Ds: Unlocking the manuscripts of medieval spain," https://www.decipheringsecrets.com/portfolio/ds-unlocking-the-manuscripts-of-medieval-spain/?lang=es.

[7] A.-H. 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*, vol. 8, 1999, pp. 65–70.

[8] G. Miner, D. Delen, J. Elder, B. NIsbet, T. Hill, and A. Fast, *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press, 2012.