ANÁLISIS AUTOMÁTICO DE POLARIDAD DE MEDIOS DE PRENSA

Santiago Barboza – Virginia Sabando

HERRAMIENTAS PARA EL ANÁLISIS DE LA WEB – MINERÍA DE LA WEB | Prof. Dra. Ana Maguitman

INTRODUCCIÓN

En el presente informe se brinda una breve descripción del problema elegido y se presenta una solución al mismo; se enumeran los objetivos perseguidos, así como también los lineamientos y estrategias adoptados en la resolución.
Por otra parte, se detallan los métodos de prueba de los códigos fuente entregados, así como también se brinda una breve explicación sobre el contenido de los restantes documentos adjuntos.
Seguidamente se dan detalles de los resultados obtenidos: se realiza un análisis de los resultados de la ejecución para un conjunto de noticias de ejemplo, y se elabora una conclusión pertinente.
Por último, se comentan brevemente las dificultades encontradas en el desarrollo de la solución propuesta, y se enumeran propuestas de trabajo futuro.

El Proyecto: Análisis Automático de Polaridad de medios de prensa

Motivación

Ante la propuesta de trabajar con herramientas vistas y estudiadas en el transcurso de la materia, y dada la posibilidad de encarar dominios de aplicación diversos y diferentes, se decidió encarar una idea de proyecto que permitiera trabajar con algoritmos de clustering o clasificación (no aplicados en el desarrollo de los trabajos prácticos previos), combinando dichas técnicas con algoritmos de minería de texto, análisis de sentimientos y recuperación automática de contenidos. Además, se optó por no trabajar con Twitter, con el objeto de experimentar con textos de mayor extensión y contrastar los resultados con el comportamiento previamente observado en los trabajos prácticos.

La idea central del proyecto consiste en realizar un análisis de los perfiles de diferentes noticieros argentinos, aplicando algunas de las técnicas vistas en el transcurso de la materia. Esta idea surge del interés de ambos miembros de la comisión por temas relacionados a política y comunicación; fue propuesta con el objeto de brindar un análisis automatizado de contenidos de diferentes medios, intentando siempre mantener la independencia y la diversidad.

Objetivos propuestos:

- para una cierta temática, estudiar qué palabras o términos son utilizados por cada medio (da una idea de la postura del mismo sobre el tema) y su polaridad.
- comparar los perfiles de diversos medios en función de la polaridad observada
- estudiar qué palabras o términos son utilizados por cada medio en general en redacción (da una idea del grado de neutralidad).
- realizar tareas de agrupamiento entre diferentes portales por medio de técnicas de clustering sobre noticias.

- Comparar y contrastar texto de noticias de diversos medios en función de los términos utilizados.

Para el desarrollo del proyecto, algunos de los recursos utilizados son:

- Canales RSS de diversos medios de comunicación argentinos para realizar la recuperación automática de noticias:
  - TN , Clarín, La Nación, Página 12, La Izquierda
- Diccionario de análisis de sentimiento en español: listas de palabras positivas y negativas (provistas por la cátedra)
- Diccionario de stopwords en español (provisto por la cátedra)
- Intérprete Python 3.5 + librerías
  - requests, feedparser, numpy, re, etc
- WEKA
- Bibliografía aportada por la cátedra

Los códigos fuentes fueron desarrollados en Python 3.5, y todas las pruebas fueron realizadas en consola SO Debian 8 (Jessie).

LINEAMIENTOS Y ESTRATEGIAS

Recuperación de la información

Se trabajó con portales de noticias en español, todos de origen argentino. Se eligió en particular aquellos que contaran con canales RSS, y en función de esto fue posible realizar la recuperación automática de los enlaces a noticias. Se trabajó con canales RSS de política.
RSS son las siglas para Really Simple Syndication, consiste en un formato XML para compartir contenido en la web. Por medio de los canales RSS es posible difundir información actualizada frecuentemente, por medio de una modalidad de suscripción: un medio que posea un canal de difusión RSS provee una URL al mismo, de forma tal que suscriptores accedan al mismo y obtengan en tiempo real actualizaciones de contenido.
RSS permite distribuir contenidos sin navegador, utilizando agregadores de noticias, diseñados para leer contenidos RSS; la mayoría de los medios masivos de comunicación con portales electrónicos cuenta con canales RSS.
Python provee una librería especial para la recuperación automática de información por medio de canales RSS, denominada feedparser; esta fue importada y utilizada en un script de recuperación de enlaces, a partir de los cuales se obtuvieron las noticias con cuyo texto se desarrolló el trabajo.

Estrategia de Análisis de sentimientos – enfoque por etapas (Pang & Lee - 2002)

La estrategia de resolución del problema planteado consiste en una adaptación del enfoque propuesto por Pang & Lee (ver referencias). Está constituido básicamente de tres etapas principales:

- **Tokenization**: eliminación de stopwords, filtrado sintáctico, eliminación de URLs, signos de puntuación, tildes y acentos, caracteres especiales (ñ) mayúsculas - minúsculas, extracción de bigramas y trigramas,

sustantivos propios. Para el análisis de la negación, especialmente importante en análisis de sentimientos, aplicación del algoritmo de Das & Chen (ver referencias): consiste en agregar el prefijo "NO_" a todas las palabras que hay entre la negación y próximo signo de puntuación. Este método amplía el corpus de términos pero simplifica la tarea de clasificación, y permite realizar una correcta evaluación de sentimiento. La política adoptada consiste en considerar una palabra positiva negada como negativa, y una palabra negativa negada como neutral. Ejemplo:

No me gustó esta película, pero yo...
||
\/
No NO_me NO_gustó NO_esta NO_película pero yo

- **extracción de características:** elaboración de corpus de términos a distintos niveles, identificación de palabras o frases claves.
  Para la confección del corpus de términos se consideraron 3 niveles diferentes de abstracción:
  - Ring-2: denominado corpus global, contiene todas las palabras de todas las noticias de todos los portales. Se genera un solo corpus Ring-2; este corpus es utilizado para realizar una comparación entre distintos portales.
  - Ring-1: conformado por todas las palabras de todas las noticias de cada portal por separado. Se generan 5 corpus Ring-1 , uno para cada portal; estos corpus son utilizados para hacer análisis de polaridad sobre tópicos particulares y de neutralidad en la redacción
  - Ring-0: un corpus Ring-0 está formado por todas las palabras encontradas en una noticia. Se generan 5 * n corpus Ring-0, donde n es la cantidad de noticias recolectadas de cada portal; estos corpus son utilizados para realizar un estudio individual por noticia de polaridad sobre términos, y en la posterior evaluación de similitud entre portales por medio de sus noticias (en conjunto con corpus Ring-2).

  La confección de los corpus de palabras fue realizada estratégicamente, con el objeto de optimizar los resultados obtenidos y mejorar el análisis automático y el clustering; algunas de decisiones adoptadas en la construcción de los corpus de términos son:
  - Para todos los niveles: incluir sólo términos que aparezcan más de una vez en el scope analizado (por ejemplo, para corpus ring-0, términos que aparezcan al menos dos veces en la noticia en cuestión).
  - No incluir stopwords
  - Respetar las mayúsculas y minúsculas: permite obtener una mejor vista respecto a posibles tópicos de interés, considerando que en los medios de comunicación se utilizan numerosas veces nombres de persona como eje temático de las noticias.

- Si bien se realizó una extracción minuciosa de bigramas y trigramas, por cuestiones de tiempo y performance éstos no fueron contemplados como tales en los corpus (ver Trabajos futuros).
- Los términos modificados por la aplicación de Das & Chen fueron almacenados en los corpus luego de la modificación, y no en su estado original.

- **Elección de términos particulares**, que constituyen el objeto de estudio de polaridad: ***Términos X.*** El conjunto denominado "Términos X" fue confeccionado a mano, no automáticamente, considerando los tópicos de relevancia en el momento de la recuperación automática. Dichos términos son recuperados de un archivo anexo al programa principal.

- **Estudio de neutralidad discursiva:** se analiza el corpus Ring-1 de cada uno de los medios de comunicación seleccionados y se observa en particular la presencia de palabras catalogadas como negativas y positivas en el mismo. La presencia/ausencia de este tipo de términos constituye un indicador de la neutralidad de un medio de comunicación en su discurso.
  - Como política general en el desarrollo de la solución, a cada ocurrencia de una palabra denominada positiva se le asignó un peso individual de +0.5, mientras que a cada ocurrencia de palabras negativas se le asignó un peso individual de -0.7. Esta política se desprende de la consideración personal de que el uso de terminología negativa resulta en un mayor impacto en el lector.

- **Estudio polaridad sobre términos X:** se analiza cada una de las noticias por separado. Se calcula un peso global para cada uno de los términos incluidos en el conjunto Términos X, denominado Polaridad_noticia. Este peso será positivo, si hay más connotaciones positivas asociadas al término; negativo, si hay más connotaciones negativas asociadas al mismo, o cero si no se habla del término en la noticia, o bien si en ella se habla equitativamente con calificativos positivos y negativos del mismo.
  - En el cálculo de la polaridad de una noticia sobre un término particular se decidió considerar la frecuencia de aparición de dicho término en el texto de la noticia.

  Para cada término **t** en **Términos-X**:
  - Para cada noticia que contenga a **t**:
    - Para cada oración en la noticia que contenga a **t**:
      - Por cada palabra con carga POSITIVA en la oración:

        Polaridad_noticia = Polaridad_noticia + 0.5 * TF (**t**)

      - Por cada palabra con carga NEGATIVA en la oración:

        Polaridad_noticia = Polaridad_noticia - 0.7 * TF (**t**)

  En función de la polaridad observada individualmente en cada noticia, se elabora un peso promedio para cada término para cada portal,

considerando para esto únicamente las noticias de dicho portal que no son neutrales con respecto al término en cuestión.

- **Estudio de similitud entre portales:** se analiza el corpus Ring-0 de cada una de las noticias por separado, y de forma paralela el corpus Ring-2, elaborado en la etapa de recolección del cuerpo de texto de cada noticia.
    - Con los corpus Ring-0 y el conjunto de términos global dado por el corpus Ring-2 se elabora una matriz TF de términos / noticias. En esta tarea se consideran absolutamente todas las noticias recolectadas, provenientes de todos los portales electrónicos considerados; el objeto de esta etapa es el de encontrar similitud de estilos de redacción entre los distintos medios.
    - Por medio de la matriz TF elaborada, se construye una matriz de similitud por coseno noticias/noticias. Dicha matriz se genera como archivo externo en formato ARFF. Esta matriz se construye para su posterior análisis por medio de la herramienta WEKA.

- **Clustering:** se toma la matriz de similitud por coseno entre noticias generada en la etapa anterior y se la procesa por medio de WEKA; el algoritmo de clustering elegido para la tarea es Expectation Maximization. Se realizaron experimentos con K-means, con diversidad de números de clusters, y los resultados obtenidos fueron demasiado variables para ser considerados significativos. Utilizando Expectation Maximization se logró una relativa consistencia en los resultados obtenidos a lo largo de extensiva experimentación. De esta etapa se generan dos archivos en formato .ARFF, ambos como resultado de realizar clustering sobre la matriz antes mencionada; uno que separa las observaciones en 3 clusters, y otro que lo hace en 5 clusters.

- **Análisis de clusters:** a partir de los archivos generados en la etapa anterior con la herramienta WEKA, se analiza a qué cluster pertenece cada una de las noticias involucradas (es decir, se realiza un análisis exhaustivo de los resultados de clusterización para todas las noticias recolectadas en el experimento). Los enlaces web recolectados para cada noticia son volcados en forma de resultado, agrupados por clusters.
    - La idea de esta etapa consiste básicamente en buscar evidencias de similitud entre medios; por ejemplo, se esperaría encontrar noticias provenientes de distintos medios diferentes que hablan sobre tópicos similares. Por otra parte, estudiando la matriz de similitud por coseno generada se podría eventualmente evaluar el grado de similitud entre noticias de distintas fuentes.

IMPLEMENTACIÓN

La implementación realizada consta de siete partes implementada cada una de ellas en un script Python diferente. Luego, en los programas principales se realiza oportunamente la invocación a cada uno de dichos scripts.
El programa principal consta de dos partes: en Programa_A.py se realiza la ejecución en orden de 1, 2, 3, 4 y 5. Por otra parte, en Programa_B.py se ejecuta

6. En medio de la ejecución de ambas partes se realiza la tarea de clustering con WEKA.

1- ObtenerEnlaces.py

En esta primer parte se realiza la recolección de los enlaces de noticias de cada portal. Se establece una conexión con el canal RSS de cada portal en cuestión, y se recuperan los enlaces de las noticias provistos por el feed. Los enlaces son almacenados en archivos separados para cada uno de los portales. La ubicación de los archivos generados es ScriptsAuxiliares/Archivos. No se almacenan títulos ni descripción de las noticias.

2- ObtenerCuerpoNoticias.py

En esta segunda parte se realizan numerosas tareas. En primer lugar, para cada portal electrónico, para cada enlace a una noticia se recupera pertinentemente la página web indicada. Seguidamente, se procede a realizar un filtrado del cuerpo de texto de la noticia, eliminando todo rastro de lenguaje html; esta tarea se realizó por medio de parsers html específicamente diseñados para cada uno de los medios con los que se trabajó.
Previo al almacenamiento de los contenidos, el texto es filtrado sintácticamente: se aplica algoritmo de Das & Chen, se eliminan stopwords, se eliminan tildes, caracteres especiales y signos de puntuación.
Se separa el texto en oraciones, donde se considera una oración a cualquier fracción de texto separada por signos de puntuación predefinidos en el código fuente (punto, dos puntos, punto y coma, interrogativos, de exclamación, etc).
El texto recolectado de cada noticia, una vez filtrado y separado en oraciones es almacenado en un archivo particular a la misma, en una carpeta que indica el medio de comunicación fuente. Además se almacena los bigramas y trigramas identificados, junto con su frecuencia de aparición. Esta información no obstante no es utilizada (ver Trabajo futuro). La ubicación de los archivos generados es ScriptsAuxiliares/Noticias. No se almacenan títulos ni descripción de las noticias. Además en esta etapa se generan los corpus Ring-1 y Ring-2, siguiendo las políticas antes descriptas. Los archivos generados correspondientes a los corpus se pueden visualizar en ScriptsAuxiliares/Corpus.

3- AnalisisRing1.py

En esta parte se realiza el análisis de los corpus Ring-1 de cada portal, con el objeto de brindar información sobre neutralidad discursiva de cada uno de los medios. Se recorre el conjunto de términos de cada medio y se calcula un peso global considerando la cantidad de ocurrencias de palabras positivas y negativas, contemplando también los pesos asignados a cada palabra positiva y negativa en particular. Los resultados son volcados en un archivo de resultados finales, denominado Resultados.txt.

4- AnalisisRing0-A.py y AnalisisRing0_B.py

En esta parte se recupera el conjunto de término denominado Términos-x y se realiza un análisis de la polaridad de cada una de las noticias por separado sobre cada término. Para ello se calculan, con los criterios antes mencionados, los pesos por noticia para cada uno de los términos en Términos-X. Los resultados de esta evaluación se vuelcan en Archivos separados para cada uno de los medios, los cuales pueden visualizarse en /ScriptsAuxiliares/Corpus/Ring0. Por otra parte, se calcula un peso por término promedio para cada portal, en función de los criterios antes establecidos. Los resultados de este proceso se vuelcan en Resultados.txt.

5- AnalisisRing2.py

En esta parte se genera la matriz tf de términos/noticias; se recupera cada una de las noticias de todos los portales, se genera un corpus Ring-0 para cada una de ellas, y con el corpus Ring-2 previamente construido (disponible en ScriptsAuxiliares/Corpus/Ring2) se genera la matriz en cuestión. Con esta matriz, a su vez, se construye una matriz de similitud por coseno entre todas las noticas de todos los portales. Esta matriz es exportada en un archivo denominado Simcos-ARFF-Dense.arff. Este archivo se encuentra en ScriptsAuxiliares.
Se trabajó con matriz tf y no con tf-idf para evitar penalizar a términos que apareciesen en múltiples noticas, pues justamente en la ocurrencia de dichas coincidencias se basa el análisis de similitud entre medios.

WEKA

Por medio de la herramienta WEKA se analiza el archivo .ARFF previamente generado, conteniendo la matriz de similitud por coseno entre noticias. Se realiza Clustering seleccionando el algoritmo EM (Expectation Maximization); en turnos se realizó el experimento para 3 clusters y para 5 clusters. Es importante recalcar que los resultados deben ser descargados en formato .ARFF a la carpeta ScriptsAuxiliares/ResultadosWeka, para su posterior análisis. El nombre de dichos archivos es resultados-EM3.arff y resultados-EM5.arff, respectivamente.

6- Analisis-Clustering_EM3.py y Analisis-Clustering_EM3.py

En esta última parte se realiza un estudio del contenido de los archivos generados por WEKA, identificando qué noticias de qué portales pertenecen a qué clusters, y seleccionando pertinentemente los enlaces a las mismas. Los resultados producto de este proceso son volcados en Resultados.txt

PRUEBA Y EJECUCIÓN

0) Es necesario realizar las pruebas situado en la carpeta de origen de los archivos fuente y de los archivos "stopwords.txt", “positivas.txt” y "negativas.txt"

1) Ejecutar Programa_A.py

$ python Programa_A.py

2) Realizar Clustering con WEKA:

- Abrir la herramienta
- Seleccionar el archivo ScriptsAuxiliares/Simcos_ARFF_dense.arff
- Aplicar Expectation Maximization para num_clusters=3 y num_clusters=5
- Descargar los archivos generados: ScriptsAuxiliares/ResultadosWeka/resultados-EM3.arff y ScriptsAuxiliares/ResultadosWeka/resultados-EM5.arff

3) Ejecutar Programa_B.py

$ python Programa_B.py

Producto de la ejecución se generan carpetas y archivos auxiliares. De todos ellos, el archivo Resultados.txt contiene los resultados relevantes al análisis.

IMPORTANTE: ejecutar separadamente alguno de los otros archivos fuente puede generar inconsistencias en los resultados.

ANÁLISIS DE RESULTADOS Y CONLUSIONES

A continuación se brinda un breve detalle de los resultados arrojados por la ejecución, y un análisis de los mismos.
Resulta importante destacar el uso de listados de palabras positivas y negativas en español, los cuales fueron modificados para contemplar términos Chen-modificados, y para omitir la categorización de palabras de uso corriente en el dominio de trabajo (por ejemplo, se eliminó la familia de palabras de “trabajador” del diccionario de palabras positivas, considerando que La Izquierda es un diario escrito por “trabajadores”).

Experiencia 1: Estudio de neutralidad discursiva de los distintos medios (Corpus Ring-1):

Medio	Puntaje Positivo	Puntaje Negativo	Total
TN	17.1152233956	-17.6601730833	-0.544949687771
Clarín	31.0017929989	-38.8662599783	-7.86446697933
La Nación	35.1536311941	-49.3420781199	-14.1884469258
La Izquierda	54.2635391606	-84.6819642165	-30.4184250559
Página 12	48.1883028412	-63.4380840021	-15.2497811609

De los resultados observados se desprende que Todo Noticias utiliza menos palabras con connotación, mostrando un discurso más neutral. Por su parte, el resto de los medios exhibe una tendencia al uso más asiduo de palabras positivas y negativas, con una clara inclinación a los términos considerados negativos. En particular, siendo Pagina Doce y La Izquierda dos medios de

prensa con un estilo orientado al editorial, los valores observados sorprenden menos en el análisis que aquellos vistos para La Nación, por ejemplo. Todos los medios muestran una tendencia al uso de vocabulario negativo.

Experiencia 2: Análisis de polaridad de cada medio sobre términos seleccionados (Corpus Ring-0):

Términos-X: ("Vido",”trabajadores", "Kirchner”, "Macri", "Massa", "Carrio", "fueros", "Randazzo", "Boudou", "Cristina").

Medio	“Vido”	“trabajadores”	“Kirchner”
TN	-0.7	0.5	-2.1
Clarín	-7.47142857143	0.0	-1.52
La Nación	-2.5875	0.0	-0.533333333333
La Izquierda	-7.76666666667	-0.815	-1.7
Página 12	-11.2	-2.21666666667	-6.9

Medio	“Macri”	“Massa”	“Carrio”
TN	4.48571428571	-1.2	-3.6
Clarín	1.85	-1.2	0.633333333333
La Nación	-1.73333333333	2.4	1.0
La Izquierda	-0.7125	0.5	-4.4
Página 12	-3.10833333333	-5.4	-6.7

Medio	“fueros”	“Randazzo”	“Boudou”
TN	-7.76	-4.05	0.0
Clarín	-20.42	-0.4	0.0
La Nación	-13.45	2.22044604925	0.0
La Izquierda	-3.0	6.5	0.0
Página 12	-22.9	-5.0	0.0

Medio	“Cristina”
TN	-0.666666666667
Clarín	-0.2
La Nación	-8.86666666667
La Izquierda	-3.22
Página 12	-0.466666666667

Algunas conclusiones que se desprenden de los resultados observados:

- La prensa de TN y Clarín (Grupo Clarín) ofrecen una imagen predominantemente positiva del actual presidente Mauricio Macri.
- Todos los medios ofrecen una imagen negativa de la ex presidente Cristina Fernandez; en particular La Nación (Grupo Clarín) presenta la tendencia más marcada.
- En todos los medios es más factible encontrar el término “Cristina” que el término “Kirchner” ligado a la ex presidente.
- En ningún medio se menciona al ex vicepresidente Amado Boudou

- El término más recurrente es “Vido”, asociado a Julio De Vido; en todos los medios se lo expone asociado a vocabulario predominantemente negativo. La tendencia es más clara en Clarín (Grupo Clarín).
- Clarín y La Nación (Grupo Clarín) son las únicas prensas que exhiben resultados positivos para “Carrio”, asociado a Elisa Carrió.
- Los resultados exhibidos por La Izquierda (Red Internacional La Izquierda Diario) y por Página Doce (Grupo Octubre) son principalmente negativos, con una leve tendencia a disminuir para personalidades de la política argentina (con excepción de Cristina para P12 y de Randazzo para La Izquierda). No resulta sencillo identificar tendencias para estos dos medios.
- El tópico “fueros”, referido a los desafueros de los legisladores, es junto con “Vido” el tópico que más resultados negativos exhibe.

Experiencia 3: Expresión de similitud entre noticias de distintos portales
Clustering: Expectation Maximization (3 clusters)

Cluster 0: 68 observaciones

- Los tópicos primordiales entre las noticias pertenecientes al cluster son: Florencio Randazzo, la inhibición de Carlos Menem, diversos asuntos asociados a Cristina Kirchner y a Santa Cruz, la detención de Mariano Bruera. Todos estos tópicos tienen en común que refieren a las próximas elecciones legislativas.
- Cabe destacar que casi todas las noticias de La Izquierda y de Página Doce se fueron situadas en este cluster, sin distinción alguna de tópico. La información sobre temática es aportada por los medios de Grupo Clarín.

Cluster 1: 32 observaciones

- Los tópicos primordiales entre las noticias pertenecientes al cluster son: fueros parlamentarios, desafueros, mina Río Turbio, desafuero de De Vido, diversos candidatos y sus renuncias a fueros. Tópicos en común: De Vido, fueros.
- Escasas observaciones pertenecientes a diarios La Izquierda y Página doce, que sin embargo son semánticamente relevantes a los tópicos identificados.

Cluster 2: 28 observaciones

- Los tópicos primordiales entre las noticias pertenecientes al cluster son: cumbre G20, Mauricio Macri en el exterior, campaña de Mauricio Macri, Dólar. Tópicos en común: G20, Macri
- Escasas observaciones pertenecientes a diarios La Izquierda y Página doce, que sin embargo son semánticamente relevantes a los tópicos identificados.

Experiencia 4: Expresión de similitud entre noticias de distintos portales
Clustering: Expectation Maximization (5 clusters)

Cluster 0: 50 observaciones

- Los tópicos primordiales entre las noticias pertenecientes al cluster son: diversos asuntos asociados a Cristina Kirchner y a Santa Cruz, PASO, candidatos. Todos estos tópicos tienen en común que refieren a las próximas elecciones legislativas.
- Cabe destacar que casi todas las noticias de La Izquierda y de Página Doce se fueron situadas en este cluster, sin distinción alguna de tópico. La información sobre temática es aportada por los medios de Grupo Clarín.

Cluster 1: 20 observaciones

- Los tópicos primordiales entre las noticias pertenecientes al cluster son: Florencio Randazzo, la inhibición de Carlos Menem, la detención de Mariano Bruera, Milagro Sala.
- Escasas observaciones pertenecientes a diarios La Izquierda y Página doce, que sin embargo son semánticamente relevantes a los tópicos identificados.

Cluster 2: 15 observaciones

- Los tópicos primordiales entre las noticias pertenecientes al cluster son: fueros parlamentarios, desafueros, diversos candidatos y sus renuncias a fueros.
- Escasas observaciones pertenecientes a diarios La Izquierda y Página doce, que sin embargo son semánticamente relevantes a los tópicos identificados.

Cluster 3: 15 observaciones

- Los tópicos primordiales entre las noticias pertenecientes al cluster son: fueros parlamentarios, Julio De Vido, mina Rio Turbio.
- Escasas observaciones pertenecientes a diarios La Izquierda y Página doce, semánticamente irrelevantes a los tópicos identificados.

Cluster 4: 27 observaciones

- Los tópicos primordiales entre las noticias pertenecientes al cluster son: cumbre G20, Mauricio Macri en el exterior, campaña de Mauricio Macri, Dólar. Tópicos en común: G20, Macri
- Una sola observación perteneciente a diario La Izquierda, semánticamente irrelevante.

Algunas conclusiones sobre los resultados observados:

- Los medios Pagina Doce y La Izquierda no se integran exitosamente en la clusterización, para ninguno de los dos experimentos. Si bien se observaron mejores resultados en el caso de EM 5 clusters, la mayor parte de las observaciones pertenecientes a estos dos portales se concentren en un único cluster, cuando no guardan relación semántica entre ellas ni con los restantes miembros del cluster. Esto puede ser a causa de los estilos de redacción, menos neutrales, más orientados al estilo editorial o de opinión característico de ambos medios. También puede inferirse que, ante la falta de relación entre estos medios y los de Grupo Clarín, no guardan entre sí relación alguna en el tratamiento de los tópicos.
- Los tres medios de Grupo Clarín (Todo Noticias, La Nación y Clarín), en cambio, mostraron en todas las experimentacio0nes guardar una estrecha relación; se pone en evidencia por medio de los resultados de clustering, tanto para EM3 como para EM5, que ante tópicos similares las noticias de los respectivos portales tendieron a agruparse. Este hecho cobra mayor relevancia en el contexto de los denominados clusters “caóticos”, donde se incluyeron la mayoría de las observaciones de los dos medios restantes (lo que conllevó una gran diversidad de tópicos); sin considerar los nuevos tópicos aportados al clúster por éstos últimos, la cohesión temática entre las noticias agrupadas de los tres medios hegemónicos es notable.

DIFICULTADES ENCONTRADAS

- Resultó inviable la utilización de dos portales de noticias inicialmente considerados: Télam y Ámbito Financiero; esto ocurrió por ser necesario construir un parser html específico para cada portal electrónico, dada la variabilidad de formatos utilizados por dichos portales en la construcción de sus páginas web.
- Se ha observado la ocurrencia de un error aislado en la recuperación del cuerpo de una noticia que no contiene texto (por ejemplo, sólo imágenes). El error se produce en instancias de ejecución del script ObtenerCuerpoNoticias.py. El enlace fue removido.
- En el transcurso de la ejecución de las distintas instancias del programa se genera una gran cantidad de archivos auxiliares y carpetas; el número de archivos escala según la cantidad de enlaces recolectados. Por otro lado, esto resta flexibilidad y portabilidad al sistema, ya que de eliminarse algún archivo o script auxiliar en medio de la ejecución ésta terminaría anormalmente.
- El análisis automático de resultados resultó ser demasiado complejo como para ser implementado en el marco de este proyecto. Se brindan los resultados de la ejecución, y éstos deben ser estudiados por el usuario.

TRABAJO FUTURO

- Incorporar nuevos portales y nuevos canales RSS
- Mejorar la interacción con el usuario: brindar medios de analítica visual para representar los resultados de forma más amigable
- Reducir la cantidad de archivos auxiliares generados en ejecución
- Automatizar el proceso de clustering, posiblemente por medio de la utilización de librerías de Python (Scikit Learn o similares) en lugar de WEKA
- Construir un parser html unificado, que contemple potencialmente cualquier formato de página web
- Incorporar bigramas y trigramas en los corpus y considerarlos en el análisis de polaridad

REFERENCIAS

1. WEKA – Expectation Maximization - http://weka.sourceforge.net/doc.dev/weka/clusterers/EM.html
2. Clarín RSS - https://www.clarin.com/rss.html?ns_campaign=prueba&ns_channel=prueba&ns_source=ageamkt_google-dynamic_pago&gclid=Cj0KEQjwy4zLBRCOg6-4h6vs3cUBEiQAN-yzfrvzFS22D4xpR6RpvlWwe2u_BKLLJ3dcYuc7o-RbaEMaApXS8P8HAQ
3. La Nación RSS - http://servicios.lanacion.com.ar/herramientas/rss/ayuda
4. TN RSS - http://tn.com.ar/rss
5. Página 12 RSS - https://www.pagina12.com.ar/pagina/rss
6. La Izquierda RSS- http://www.laizquierdadiario.com/RSS
7. Feedparser documentation for Python - https://pypi.python.org/pypi/feedparser
8. Feedparser documentation for Python - https://pythonhosted.org/feedparser/
9. Scikit-Learn for Python - http://scikit-learn.org/stable/
10. Análisis de sentimiento en español - https://pybonacci.es/2015/11/24/como-hacer-analisis-de-sentimiento-en-espanol-2/
11. Pang & Lee aplicado a análisis de sentimiento - http://pdln.blogspot.com.ar/2014/01/analisis-de-sentimientos-un-algoritmo.html
12. Opinion mining and sentiment analysis- Bo Pang & Lillian Lee http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf