Aspectos lingüísticos a considerar en el trabajo con LDA para análisis discursivo

Introducción general

El auge de la sociedad de la información, apoyada en nuevas técnicas y herramientas digitales que han permitido la acumulación masiva de datos en diversos formatos, entre ellos las grandes cantidades de texto que se producen y publican diariamente en plataformas digitales, ha conllevado un creciente interés en el desarrollo de herramientas para la automatización del procesamiento, organización y clasificación de elementos del lenguaje natural digitalizado aplicables al análisis de datos publicados en la web.

Entre las herramientas que actualmente se desarrollan, con potencial uso para el procesamiento de amplios cúmulos de datos textuales digitales, podemos encontrar el modelado de tópicos (topic model), una técnica probabilística e informática que ha mostrado ser útil para el tratamiento de textos en la web, permitiendo inferir, analizar y comparar datos e información de distinta índole de manera automatizada.

En este sentido, destaca el desarrollo de la Asignación Latente de Dirichlet (LDA por sus siglas en inglés), que consiste en un modelo generativo probabilístico no supervisado para modelar grandes corpus de texto, y generar aleatoriamente los documentos que se observan en este corpus (Blei, Ng y Jordan, 2003). Este modelo, basado en conceptos de Modelos Bayesianos, permite inferir tópicos a partir de un conjunto de documentos, mediante la aplicación de una distribución a posteriori. Tales tópicos pueden ser entendidos como temas estructurantes del corpus y es factible emplearlos para organizar los documentos que constituyen el corpus, según los criterios que se definan como de interés.

Esta guía recoge elementos de distintas disciplinas (informática, estadística, matemáticas, lingüística) que buscan aportar a la comprensión del funcionamiento del LDA como modelo para el análisis textual de corpus amplios, así como su uso para el análisis de distintos tipos de discurso. Además de una perspectiva multidisciplinaria, fruto del trabajo colaborativo y abierto, este documento surge a partir de un enfoque inductivo-deductivo mediante el cual se abordó el estudio del modelado de tópicos desde sus fundamentos teóricos, lo que luego se contrastó a partir de la aplicación del LDA al análisis de tres corpus lingüísticos de distinta naturaleza discursiva.

En este documento discutimos elementos generales relativos a la técnica del modelado de tópicos, así como algunos aspectos estadísticos y otros aspectos generales de carácter lingüístico de interés para quien busca aplicar el LDA al análisis de corpus textuales. También, y a manera de aportar a la comprensión del funcionamiento de esta

herramienta detallamos tres ejercicios de análisis con corpus distintos (medios digitales, consulta pública, y discursos políticos).
El principal propósito de este documento es poner a disposición de la comunidad general elementos teóricos y resultados parciales de la investigación que actualmente se desarrolla en el Centro Nacional de Desarrollo e Investigación en Tecnologías Libres (CENDITEL), ente adscrito al MPPEUCT, y que busca aplicar herramientas digitales al análisis textual a fines de mejorar procesos en el desarrollo de tecnologías libres.

Aspectos lingüísticos a considerar en el trabajo con LDA para análisis discursivo

A fines de aproximarse al funcionamiento de la herramienta de modelado de tópicos con LDA (*latent Dirichlet allocation*) para el análisis de lenguaje natural resulta pertinente tener en cuenta algunos aspectos básicos de la lingüística, y de la construcción y el manejo de corpus lingüísticos.

I. Sobre el lenguaje y su funcionamiento

A. Entre las visiones teóricas que abordan el funcionamiento del lenguaje como sistema encontramos la formalista y la funcionalista que intentan explicar las relaciones internas de este complejo conjunto de elementos. De acuerdo con la visión formalista del lenguaje, el sistema lingüístico opera como:
1. Un todo homogéneo, cuyos elementos en consecuencia se entienden como equiprobables;

En ese sentido, dado un evento lingüístico como el siguiente[1]:
Juan llevaba aquel precioso _____ en sus manos
Los siguientes elementos para llenar la casilla vacía: niño, libro, objeto, tendrían todos las mismas posibilidades de aparición, siendo los mensajes resultantes los siguientes:
a. Juan llevaba aquel precioso niño en sus manos.
b. Juan llevaba aquel precioso libro en sus manos.
c. Juan llevaba aquel precioso objeto en sus manos.
En efecto, los tres términos tienen la misma posibilidad de aparición, por lo que son considerados sustantivos masculinos singular en español.

[1] Este ejemplo es tomado textualmente de Domínguez, C.L (2003), quien expone en detalle estas dos visiones entre otros aspectos relevantes de la lingüística y el análisis de oralidad y escritura.
(http://www.human.ula.ve/linguisticahispanica/documentos/Dominguez.pdf)

Mientras que la visión funcionalista concibe el lenguaje como:
2. Un conjunto de subconjuntos, una entidad heterogénea, cuyos elementos se realizan probabilísticamente de acuerdo con las variables (internas y externas) que operan en el momento de realización.

Si a partir del ejemplo anterior consideramos el término introducido en el caso c como un tópico determinado "objeto" un hablante venezolano dispondría de las siguientes opciones: objeto, cosa, perol, coroto, coso [masculino de cosa], bicho [inanimado], macundales, chéchere, cachivaches, entre otras, dentro del repertorio de términos asociados a tal tópico, siendo los mensajes resultantes los siguientes:
d. Juan llevaba aquel precioso objeto en sus manos.
e. Juan llevaba aquel precioso perol en sus manos.
f. Juan llevaba aquel precioso coroto en sus manos.
g. Juan llevaba aquel precioso bicho en sus manos.
h. Juan llevaba aquel precioso chéchere en sus manos.

Un hablante venezolano entendería que dadas las opciones señaladas, no estamos ante términos equivalentes en cualquier situación de habla, pues si bien pueden considerarse sinónimos, las mismas no tienen el mismo sentido respecto a su uso en contexto. Una situación de habla formal en un contexto académico podría admitir la opción a, mas no la opción g, por ejemplo.
Los términos no son entonces equiprobables, pues la selección entre las opciones estará determinada entonces por el interlocutor, la situación de enunciación y la razón de ser del mensaje.

B. Sobre las relaciones sintagmáticas y paradigmáticas
Ambas visiones concuerdan sin embargo en que el funcionamiento del sistema está dado por relaciones paradigmáticas y sintagmáticas, esto es las relaciones del eje de lo que se puede agrupar como un conjunto por su naturaleza común (paradigma), y las reglas *del orden de lo que va junto* que permiten combinar esos conjuntos en unidades compuestas (sintagmas). Son estas relaciones las que permiten que una unidad discreta se una a otra(s) para generar unidades mayores de sentido más complejo en cada nivel de articulación.

En el nivel más elemental de la lengua (Nivel fonético y fonológico), las unidades discretas carecen de sentido en sí mismas (fonemas). Cada lengua dispone de un conjunto limitado de fonemas, que se subdivide en fonemas vocálicos y consonánticos y se agrupan de acuerdo con una serie de normas internas para generar sílabas y palabras (Nivel morfológico).

La gramática de cada lengua asigna a estas unidades completas una función de

designación de elementos o entidades (sustantivos), procesos (verbos), así como la función de calificar unos u otros (adjetivos y adverbios), o servir como enlaces para construir relaciones entre estas palabras (preposiciones, conjunciones).

En el nivel semántico, el universo de palabras de una lengua se agrupa de acuerdo con su sentido en conjuntos que podemos considerar como tópicos. Estos tópicos configuran entonces subconjuntos dentro de la lengua, que se realizan de acuerdo con una serie de variables internas (cohesión semántica) y externas (variables discursivas y sociolíngüísticas). Por ejemplo, un tópico general como el de los *seres vivos* incluirá a su vez un subtópico *animales*, que luego incluye otro más específico como *mamíferos*, y así sucesivamente. En un texto sobre *biología* (un tópico más general e incluyente de los anteriores) será más probable que podamos encontrar el tópico *seres vivos*, que en un texto sobre matemática financiera[2].

En el nivel propiamente sintáctico, las palabras se unen en un orden secuencial para formar enunciados (frases u oraciones), las que a su vez se articulan en unidades mayores de sentido para generar la argumentación (el orden de las ideas y las secuencias lógicas de las mismas) y la estructura formal de un texto (párrafos, apartados, capítulos, etc).

En el nivel discursivo, estas unidades complejas de sentido se estructuran de una determinada manera para dejar saber al interlocutor la estructura discursiva del texto que se construye (introducción, desarrollo, conclusión).

En cada uno de estos niveles, que se articulan de forma simultánea al momento del enunciado, operan relaciones paradigmáticas y sintagmáticas, que permiten a los hablantes entender el funcionamiento de los subconjuntos y conjuntos de unidades a su disposición, así como las reglas mediante las que estos pueden ser combinados y recreados en cada situación de habla para denotar y connotar complejos efectos de sentido.
Retomando el primer ejemplo, diremos que los términos -niño, objeto, libro- hacen parte de un paradigma gramatical: sustantivos, masculinos en singular. Esto es lo que permite que todos ocupen la misma posición dentro del sintagma Juan llevaba aquel precioso ___ en sus manos.
Los términos -objeto, cosa, perol, chéchere- también pertenecen al paradigma sustantivos, masculinos, singular, pero a su vez a un paradigma más acotado que se determina por su sentido en tanto términos referidos a una *entidad con presencia material*, es esto lo que los hace sinónimos.
En el eje sintagmático la relación de calificación que hace el término *precioso*, y que viene dada por la formalidad del adjetivo, es la que nos permite entender que la probabilidad de aparición de “objeto” sea mayor, dado que comparten el rasgo de formalidad, mientras

[2] Si bien no sería imposible que se usara una metáfora con animales (un tópico más concreto) para explicar algún proceso financiero (un tópico más abstracto).

que su combinación con un término como *chéchere* pueda producir un efecto de sentido irónico, por ejemplo[3].

¿Qué relaciones lingüísticas permite entender el LDA?

A partir de la observación del funcionamiento del LDA entendemos que esta herramienta parece reconocer *relaciones paradigmáticas* entre los términos que se agrupan más frecuentemente para conformar un tópico dentro de un texto o un conjunto de textos (corpus).

Esto permite identificar de manera rápida en un corpus amplio **tópicos semánticos** comunes a los textos que conforman la totalidad del corpus.

Este reconocimiento también implica necesariamente un primer nivel de identificación sintagmática, al poder reconocer los términos que se asocian más frecuentemente entre sí dentro de los textos que conforman un corpus amplio.

Por ejemplo, en un corpus de artículos científicos de ciencias sociales, el LDA podrá identificar diversos tópicos, posiblemente algunos más relacionados con historia, otros con comunicación, o política. Dentro de esos tópicos, el LDA permite ver los términos que más frecuentemente se relacionan entre sí en los textos que conforman el corpus. Por ejemplo, dentro de un tópico como comunicación, podremos ver términos como lenguaje, lengua, idioma, pero también términos como social, colectivo, individual, que son frecuentemente asociados con los primeros.

II. Sobre la construcción de un corpus

Un elemento clave para abordar el funcionamiento del LDA para el análisis de lenguaje natural es la construcción de un corpus de análisis con potencialidad para el modelado de tópicos automatizado.

Al respecto diremos primer lugar de manera fundamental que un **corpus** lingüístico es un conjunto de textos que presentan características formales comunes, lo que implica que pertenecen a un mismo **género** discursivo.

Los géneros discursivos son tipos estables de formas de organización de las unidades discursivas de acuerdo con la situación y el propósito comunicacional. Estas formas de organización del discurso pueden caracterizarse de acuerdo con su modalidad de producción (oral/escrita), con su modo de transmisión, con el tipo de interacción que

[3] En esa interpretación mucho dependerá de otros elementos como el contexto específico de enunciación o los rasgos de entonación y la gestualidad del interlocutor.

permiten establecer entre los sujetos (monológicos/dialógicos), así como con el grado de formalidad (formal/informal), entre otros rasgos.

Por ejemplo, el discurso periodístico contiene diversos géneros como la entrevista (oral, dialógico, más o menos formal), el reportaje (escrito o audiovisual, monológico generalmente, y más formal), o la nota de prensa (escrito, monológico, formal).

Para considerar un conjunto de textos como pertenecientes a un corpus factible de analizar como un todo es necesario que los mismos puedan caracterizarse bajo un mismo género discursivo, además de otros posibles rasgos comunes entre los textos, por ejemplo, artículos científicos (de una misma disciplina o no), notas de prensa (de un periodo determinado, de un tema particular, o de un medio o periodista específico), poemas (de un periodo específico, de un mismo autor, o de un mismo tema).

Si bien un corpus puede estar compuesto por dos o más textos, en el caso particular de las posibilidades de análisis que brinda el LDA, esta pareciera adaptarse mejor al análisis de corpus amplios, que contengan un número considerablemente extenso de textos (AQUÍ HABRÍA QUE INDICAR SI HAY ALGÚN NÚMERO IDEAL DEL CUAL PARTIR)

*Para consultar mayores consideraciones sobre el tema de la construcción de corpus y la lingüística de corpus, ver Parodi (2008), Lingüística de corpus: una introducción al ámbito. http://www.scielo.cl/pdf/rla/v46n1/art06.pdf

Tres ejemplos de análisis de corpus lingüísticos con LDA

Con el fin de entender mejor el funcionamiento del LDA para el análisis de corpus lingüísticos, diseñamos y aplicamos un protocolo de análisis piloto para tres (3) corpus de naturaleza discursiva distinta, que nos permitieran entrever posibles diferencias en los resultados que apunten a identificar categorías discursivas que puedan ser analizadas mediante la aplicación del LDA a corpus amplios.

A continuación detallaremos cada corpus de análisis, así como los rasgos discursivos que consideramos de interés a partir de los resultados obtenidos en el análisis mediante el uso del LDA.

1. Plan de la Patria

I. Definición del corpus

A partir de la consulta pública constituyente convocada por el Presidente Hugo Chávez en el año 2012 en torno a la propuesta del Plan de la Patria (2013-2019) se constituyó un corpus de análisis conformado por XXX consultas recibidas mediante el sistema de consulta pública digital. Este sistema solicitaba al usuario (individuo o colectivo) completar una serie de campos (de identificación y relativos a la propuesta a subscribir) que le permitían desarrollar una propuesta que pudiera ser incorporada como parte del Plan Nacional de Desarrollo de la Nación, Plan de la Patria.

Las consultas recibidas en este proceso presentan una serie de características textuales y discursivas comunes que nos permiten considerarlas un conjunto de textos factibles de analizar en tanto corpus. Las mismas son muestras de habla escrita, con rasgos de formalidad, dada la situación de habla institucional en la que se enmarcan, y generalmente desarrollan uno o dos tópicos semánticos, por cuanto se solicitaba como parte de la consulta que se identificara un objetivo general del Plan de la Patria con el cual se relacionaba la propuesta a realizar mediante el sistema de consulta y esto restringe generalmente el campo semántico a desarrollar.

II. Preprocesamiento del Corpus

El procedimiento correspondiente al preprocesamiento de los corpus viene dado por un

script diseñado en *python*, que contiene una serie de configuraciones que indican cómo se debe procesar los textos antes de ser ingresados a una librería llamada *freeling*.
Esta configuración consiste en listar los elementos textuales que se deben excluir antes de ser procesado. En este apartado tenemos las categorías: verbos, adjetivos, sustantivos, adverbios, determinantes, pronombres, conjunciones, interjecciones y preposiciones, de las cuales generalmente se excluyen por razones de relevancia semántica los pronombres, conjunciones, interjecciones, preposiciones y adverbios.
Una vez hecho esto la librería se encarga de procesar y arrojar los resultados en un formato que luego será interpretado por el LDA y cuya interpretación se ve representada en la interfaz gráfica que se dispone a mostrar al usuario.
La interfaz de usuario proviene de un proyecto en *github*, perteneciente a un desarrollador de la universidad de Indiana, el cual implementa la visualización de datos del LDA en el framework VSM. En este sentido, se consideró cambiar esta implementación por un *framework* más robusto como lo es *Django* y poder así adaptarlo a las necesidades pertinentes de nuestro contexto tanto político como social.
Es importante destacar que la implementación del VSM trabajaba por defecto con una implementación de LDA basada en el muestreo de *Gibbs* (modelo con el que se estuvo trabajando en un principio y por la naturaleza de los resultados se intuye que convergía), por lo que con colaboración de algunos scripts realizados por Jamie Murdock (autor del proyecto en github), más algunos de autoría de Jorge Redondo se pudieron traer resultados del LDA-C (LDA de Blei) a la interfaz. Cabe destacar que estos resultados eran mejores que los presentados anteriormente por los del muestreo de Gibbs.
El cambio más importante a nivel de visualización con respecto al proyecto original fue la implementación que permitió ver la estructura de cada de uno de los corpus por separado, y a su vez permitir ver a través de una nube de palabras la relevancia de cada palabra dentro de los tópicos que constituyen el corpus seleccionado, es decir que entre más porcentaje (determinado por el LDA) tenga una palabra en el tópico, mayor será su tamaño en la nube de palabras.

III. Resultados

Los resultados obtenidos mediante la aplicación del LDA al procesamiento del corpus Plan de la Patria mostraron consistencia semántica en la identificación de los tópicos relevantes dentro del corpus, así como en la identificación de la relevancia de las propuestas dentro de cada tópico (por su grado de proximidad semántica con respecto al tópico), y la identificación de la relevancia de cada palabra dentro de los tópicos.
Se obtuvieron resultados que permitieron identificar desde los 10 hasta tópicos más relevantes hasta 90 tópicos, lo que resulta de gran utilidad en un corpus tan amplio y de naturaleza semántica tan diversa como un consulta pública relativa al Plan Nacional de Desarrollo, lo que contempla todas las áreas de competencia del Estado y las áreas de interés de las y los ciudadanos.

En la figura C. 1 se puede apreciar la visualización de los resultados del análisis del Corpus Plan de la Patria identificando 70 tópicos.

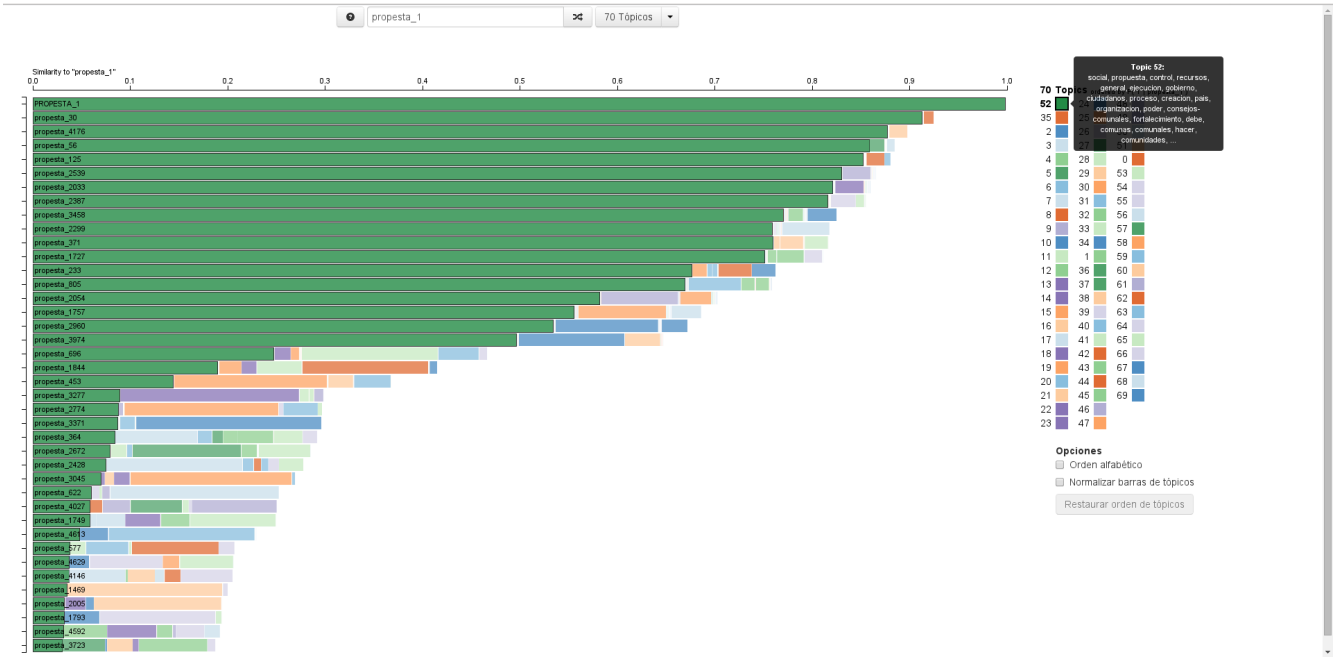

Figura C.1

Mediante la herramienta de visualización es factible seleccionar un tópico, en este caso el tópico 52 (social, propuesta, general, ejecución, gobierno, ciudadanos, proceso, creación, país, organización, poder, consejos comunales, fortalecimiento, debe, comunas, comunales, hacer, comunidades…), y ordenar los textos que componen el corpus de acuerdo con la relevancia que tenga ese tópico en cada uno de los textos. En este caso la propuesta_1 es el documento más relevante para el tópico 52.

En la figura C.2 podemos apreciar la visualización de la relevancia de cada palabra dentro de un texto, en este caso la propuesta_1.

Figura C. 2

Esta interface permite identificar rápidamente mediante el uso de colores el tópico de pertenencia de cada palabra identificada como perteneciente aun tópico dentro de la propuesta, lo que resulta útil al momento de identificar relaciones semánticas entre los textos que componen el corpus. Igualmente, el tamaño de la palabra dentro del texto nos indica la relevancia del término a lo interno del tópico al que pertenece, esto es su frecuencia de aparición dentro del tópico.

1. Medios digitales

I. Definición del corpus

A fines de constituir un corpus factible para probar el funcionamiento del modelado de tópicos mediante el uso del algoritmo LDA en el análisis de medios de comunicación digitales en Venezuela, se definió un periodo comprendido entre el 17 y 18 de febrero de 2016. Tal periodo se definió tomando en cuenta la alocución presidencial del día 17 de febrero en la que el Presidente de la República Nicolás Maduro y su gabinete ministerial anunciaron una serie de medidas económicas de alto impacto en la vida nacional, lo que se identificó como un evento comunicacional de alta repercusión en la agenda mediática del país. Este evento genera un parámetro claro, tanto para la definición del corpus de estudio, como para la evaluación de la eficacia de la herramienta para el análisis de discurso mediático, al poder comprobar en los resultados del análisis si el LDA modela los tópicos relativos a los temas presentados en tal evento comunicacional, que se espera sean los temas más recurrentes en la agenda de los medios nacionales.

El corpus está constituido enteramente por notas de prensa digitales, cuyo formato textual generalmente conserva una tipología definida por ser un tipo de texto formal, conciso (un promedio de dos párrafos por nota), en el que se desarrolla uno o dos temas generales en promedio.

II. Automatización de la compilación del corpus

Se diseñó una herramienta de *web scrapping* para la recolección automatizada de las notas de prensa identificadas como publicadas en el periodo definido. Para tal fin, se identificó las secciones de Nacionales, Políticas y Economía como las de interés para el análisis, excluyendo así las demás secciones de los medios a analizar. Se seleccionó un grupo de cinco (5) medios digitales de relevancia nacional, con el propósito de normalizar la identificación tanto de la fecha como de la sección de publicación de la nota.

Se obtuvo de esta manera un corpus de 915 notas de medios digitales publicadas entre el 17 y 18 de febrero en las secciones *nacional*, *política* y *economía* que esperábamos mostraran principalmente los temas abordados en los anuncios económicos gubernamentales.

La herramienta de *web scrapping* se desarrolló usando un *framework* de *Python* llamado Scrapy, el cual esta diseñado precisamente para esa tarea. Es importante resaltar que para poder realizar *scrapping* a una web es necesario conocer con antelación la estructura del sitio web a inspeccionar, hecho esto se procede a crear un araña (término que se le da a un programa que inspecciona una web de manera automatizada) con las configuraciones correspondientes al sitio del que se extraerá la información, por lo que es importante resaltar que debido a la diversidad de los sitios de noticias es preferible contar con una araña personalizada que se adapte a las necesidades especificas de un sitio, de modo que si el mismo cambia con el tiempo, el único código que se vería afectado es el de la araña correspondiente.

Los principales parámetros que se deben considerar son las URL's o direcciones del sitio que se desean explorar, las categorías que se desean tomar en cuenta y lo más importante y que conlleva más trabajo es conocer la estructura de los artículos para así proceder a la extracción de la información que los conforman.

Para el trabajo planteado en particular fue necesario plantearse dos parámetros en particular, la fecha de inicio y la fecha de fin, es decir el intervalo del que se extraerá la información.

Otro punto relevante con los medios digitales, es que la estructura de los sitios web se deben prestar para el *scrapping*, lo que se puede resumir para este caso en 3 aspectos: El primero es que el sitio tenga sus noticias clasificadas por categorías (es algo elemental en toda noticia, pero hay sitios que no lo hacen), segundo que tengan en sus páginas de categorías un historial (es decir, la data histórica de todas las noticias que se han publicado, por lo general en una tabla), se puede citar el ejemplo de El Universal, que no lo hace por ejemplo en ninguna categoría; el tercer y último aspecto es que en caso de que la tabla cargue de forma dinámica (Ajax por lo general) es necesario consultar a las URL que hace petición el servidor para traerse los datos, y

algunos sitios manejan autenticación para poder acceder a dichas URL'S.

Ahora entrando en materia sobre el procedimiento que se realizó para extraer el material de los medios digitales se puede resumir en los siguientes pasos:

•Crear la araña y configurarlas con URL's del sitio

•Configurar los parámetros para extraer la información (se especifica de donde se extraerá el autor, titulo, fecha, cuerpo de la noticia, etc)

•Se corre por consola la araña pasando por parámetro el intervalo de las fechas que se desea buscar

•Al finalizar el scrapping la araña crea un archivo en formato .json con los resultados de todos los medios

•*Nota: Como la araña en si busca por la tabla que se encuentra en la sección especificada, a modo de reducir los tiempos de espera se puede configurar dentro de la araña desde que página a que página se debe buscar (obviamente conociendo dicho intervalo a priori)*

Una vez realizado el *scrapping* es necesario transformar los archivos .json que arroja como salida en archivos de texto plano que puedan ser tratados por el preprocesamiento, tarea que se realizó con un script en *python*.

Es importante que los tiempos de espera son cortos, pero a su vez van relacionados con los servidores en los que estén alojados los sitios, como ejemplo de guía: Si se establecen a priori las páginas, un *scrapping* con una conexión promedio a un sitio con una velocidad promedio puede tardar de 2min a 5min, ahora sin conocer a priori las páginas y tomando en cuenta unas fechas como las analizadas (febrero), digamos una noticia de unos 3-4 meses de anterioridad, dependiendo del flujo de noticias que tenga el sitio, se puede estimar que el tiempo de espero podría ser de 15-30min. Ahora el tiempo que tarda el script en convertir .json en texto plano, son milésimas de segundos, si son muchos datos a procesar tal vez unos pocos segundos, en general nada de que preocuparse.

III. Preprocesamiento del corpus

A partir de una primera corrida de los textos compilados se pudo identificar una serie de términos de frecuente aparición a lo largo de todos los tópicos y que son característicos del tipo de género discursivo periodístico. Estas palabras se identificaron y seleccionaros para ser excluidas junto con el preprocesamiento estandar del texto que excluye palabras de bajo interés para el análisis por su naturaleza gramatica (preposiciones, artículos, adverbios).

Preprocesamiento de discurso periodístico

Sustantivos
País
Venezuela
Año
Día

Caracas
Ayer

Adjetivos
Venezolano

Verbos
Haber
Decir
Hablar
Explicar
Indicar
Asegurar
Aseverar
Anunciar
Realizar
Informar
Calificar
Poner
Querer
Presentar
Seguir
Llevar
Expresar
Manifestar
Considerar
Afirmar
Destacar
Señalar
Referir
Llamar
Agregar
Publicar
Poder

IV. Resultados

A partir de este piloto de análisis automatizado mediante el uso de LDA del corpus constituido por cerca de mil notas digitales de cinco medios venezolanos se obtuvo resultados de interés que nos permiten entrever la pertinencia del uso de esta herramienta para la automatización de procesos de análisis de medios de comunicación digital.

El análisis del corpus arrojó los siguientes datos para la visualización de diez tópicos:

10 Tópicos
Tópico 3 – Comisión de contraloría de la AN investiga por corrupción a altos funcionarios del gobierno
Tópico 1 – Aumento precio gasolina, precio petróleo, sistema cambiario
Tópico 2 – Guyana, Ginebra, Asamblea
Tópico 4 – Leopoldo López
Tópico 5 – Visita premio Nóbel de la Paz – Leopoldo López
Tópico 6 – Ley de Amnistia
Tópico 7 – Medidas económicas / modelo económico
Tópico 8 – No identificado claramente / relativo a medidas económicas
Tópico 9 - No identificado claramente / relativo a medidas económicas
Tópico 0 - No identificado claramente / relativo a medidas económicas

Estos resultados muestran el tópico principal de las medidas anunciadas por el Presidente Nicolás Maduro en el periodo seleccionado, desplegado en dos subtópicos (Tópicos 1 y 7), y además muestran tópicos políticos de la agenda mediática de la oposición venezolana recogidos por los medios digitales (tópicos 2, 4, 5, y 6).
Tales resultados permiten entrever que la herramienta de modelado de tópicos resulta pertinente para la automatización de análisis discursivo de medios de comunicación digital, cuyo formato textual y temático se comporta adecuadamente con el modelo del LDA.

1. *Aló, Presidente*

I. Definición del Corpus

A partir de la publicación del portal digital www.todochavezenlaweb.gob.ve del Instituto de Altos Estudios del Pensamiento de Hugo Chávez, se pudo tener acceso a las 378 emisiones del espacio audiovisual *Aló, Presidente* como corpus discursivo constituido característico del discurso presidencial

del Comandante Hugo Chávez.
La compilación de este corpus se hizo mediante la aplicación de una herramienta automatizada de *web scrapping* que permitió la captación de las 378 emisiones y su paso al formato de texto adecuado para su análisis mediante el modelado de tópicos con LDA.
El formato textual del *Aló, Presidente* ha sido caracterizado como un nuevo género político-mediático complejo (Bolívar, 2003; Elrich, 2005), en el que no sólo se da una imbricación de otros géneros discursivos simples, sino que además se entretejen distintos tópicos semánticos (temas centrales y temas satelitales) (Gualda, 2010)[4]. Tal complejidad se figurativiza en un formato textual extenso (un promedio de duración de 5 horas por sesión), en el que participan diversos actores sociales (Estado, Gobierno, Poder Popular, FANB, entre otros) junto al Presidente Hugo Chávez como un ejercicio de representación, deliberación y participación política mediante el uso de las tecnologías de información y comunicación en la esfera pública medida.

II. Preprocesamiento

Dado el carácter político institucional del género discursivo analizado en este corpus se decidió aplicar el mismo preprocesamiento definido para el corpus Plan de la Patria, esperando encontrar en los resultados incidencia de unidades lexicales que se evidenciaran como características del corpus de análisis, tal y como se definió a partir de los primeros resultados con el corpus de medios digitales.

III. Resultados

Los resultados obtenidos del análisis automatizado mediante modelado de tópicos con LDA en el corpus constituido por las 378 emisiones del *Aló, Presidente* no permitieron entrever relaciones paradigmáticas o sintagmáticas entre las unidades que componen los conjuntos propuestos por la herramienta como posibles tópicos lexicales del corpus.
Consideramos que tales resultados responden justamente a la complejidad discursiva y temática del género discursivo en cuestión, que no se ajusta adecuadamente a la herramienta de modelado de tópicos con LDA. También puede tratarse de un corpus que no alcanza la extensión requerida por el modelo estadístico para arrojar resultados consistentes, pues si bien cada emisión es significativamente extensa en sí misma, el corpus en total está compuesto sólo por 378 emisiones, a diferencia de los dos corpus anteriores que planteaban un número mayor de unidades textuales a analizar.

---

[4] BOLIVAR, A. (2003). Nuevos géneros discursivos en la política: El caso de Aló Presidente. En L. Berardi (Comp.), *Análisis crítico del discurso. Perspectivas latinoamericanas* (pp. 101-130). Santiago de Chile: FRASIS editores.
ELRICH, F. (2005) La relación interpersonal con la audiencia: el caso del discurso del presidente venezolano Hugo Chávez, en *Revista Signos*, v.38, n.59, Valparaíso: Pontificia Universidad Católica de Valparaíso. pp.287-302. Consultado el 06/02/2009, Disponible en: http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-09342005000300002&nrm=iso&lng=es. ISSN 0718-0934
GUALDA, R. (2012) *The Discourse of Hugo Chávez in "Aló Presidente": Establishing the Bolivarian Revolution through Television Performance*. Faculty of the Graduate School of The University of Texas at Austin.