Metodologías para análisis político utilizando Web Scraping

Alexis Tadeo Hernández, Edy Gómez Vázquez, César Alejandro Berdejo Rincón, Jorge Montero García, Adrian Calderón Maldonado, y Rodolfo Ibarra Orozco

Universidad Politécnica de Chiapas,
Tuxtla Gutiérrez, Chiapas, México

Resumen. En este artículo se revisan distintas metodologías utilizadas para realizar un análisis político utilizando diversas fuentes de información disponibles en internet. En algunas sociedades, el uso de redes sociales tiene un impacto significativo en el ámbito político con la sociedad y se han empleado diversas metodologías para analizar diversos aspectos políticos y las estrategias a seguir. El propósito de este trabajo es entender estas metodologías para poder proporcionar información a los posibles votantes que les permitan tomar decisiones informadas. Primero, se revisa la terminología necesaria sobre web scraping, después, se presentan algunos ejemplos de proyectos para el análisis político que han empleado web scraping. Finalmente, se presentan nuestras conclusiones.

Palabras clave: Web scraping, text mining, análisis político.

1. Introducción

En Estados Unidos, las redes sociales han tomado un papel importante en el ambiente político: las usan para investigar a fondo a la oposición con equipos especializados para encontrar incoherencias en el adversario; aprovechan las diferencias y coyunturas, por ejemplo, [8], en los Estados Unidos, los republicanos en el Congreso se opusieron al recorte del impuesto sobre la nómina, que obligaría a los estadounidenses a que en cada pago de su sueldo le fueran recortados, en promedio, unos 40 dólares. Uno de los argumentos usados fue que 40 dólares no era mucho dinero.

En menos de 12 horas, la Casa Blanca reaccionó con una estrategia en la que invitaban a los ciudadanos a enviar por Twitter, Facebook, YouTube qué significaba para ellos 40 dólares y después Barack Obama tomó ejemplos de cómo 40 dólares menos al mes afectan a las familias estadounidenses logrando que el Congreso de los Estados Unidos rechazara el recorte del impuesto sobre la nómina [10].

Este es un ejemplo de cómo algunas sociedades aprovecha las redes sociales en el ámbito político. En México, por el contrario, la política no deja de ser una discusión eterna entre votados y votantes por medio de los spots que publican en fuentes convencionales (radio, televisión, incluso el cine), dando una "guerra" discursiva entre los votantes a favor de un partido o candidato y quienes no saben por quién votar [8].

Conocer a fondo a los políticos, consultar diversas fuentes y analizarlas es una labor complicada, éstas son las razones de desarrollar una herramienta que, utilizando técnicas de web scraping y text mining, permita a la población conocer de diversas fuentes (redes sociales, sitios web de periódicos, búsqueda en Google) a cualquier político, analizar los resultados de la búsqueda y mostrar indicadores de confianza en base a lo obtenido en un análisis de la información con el fin de dar al usuario una perspectiva distinta.

1.1. La política y el uso de las redes sociales

El uso de las redes sociales ha sido un factor importante para poder compartir opiniones sobre diversos temas que le interesan a la ciudadanía, conocer a candidatos que se postulan para un puesto, ya sea presidencia, senadores o demás personas que trabajen en el ámbito político.

En las elecciones del 2008, en los Estados Unidos, podemos ver un claro ejemplo acerca de cómo los candidatos a la presidencia de los Estados Unidos de América, Mitt Romney y Barack Obama, hacen uso de las redes sociales para darse a conocer con las personas, además de dar sus puntos de vista acerca de diversos temas sociales y así poder opinar junto con la sociedad.

La implementación de la tecnología en la política ha permitido participar en un nuevo nivel de conversación con los votantes, permitiendo así que una campaña donde los candidatos se dan a conocer se vuelva algo mucho más dinámico, algo más de un simple diálogo. Así Barack Obama, en las elecciones presidenciales de 2012, utilizó de forma más efectiva la web para poder establecer una campaña más insurgente y poder ganar el voto de las personas jóvenes, lo cual fue funcionando. Por ejemplo durante esta campaña (de junio 4 a junio 17), la campaña de Obama realizó 614 publicaciones mediante su plataforma, mientras que la Romney realizó sólo168, y la brecha en twitter fue aún mayor, promediando 29 mensajes diarios de Obama contra uno de Romney [7].

2. Web Sraping

También conocido como Web harvesting o Web data extraction, es el proceso de rastreo y descarga de sitios web de información y la extracción de datos no estructurados o poco estructurados a un formato estructurado. Para lograrlo, se simula la exploración humana de la World Wide Web, ya sea por implementación de bajo nivel del protocolo de transferencia de hipertexto, o la incorporación de ciertos navegadores web.

Para realizar scraping se utiliza un programa, conocido como orquestador, que organiza y ejecuta las peticiones al browser. Se deben tener bien definidos los elementos a buscar, y que se indique el estado de la búsqueda a realizar (búsqueda exitosa, errores en la búsqueda, sin resultados).

El proceso de web scraping se realiza en dos etapas, la primera es la etapa de extracción, en la cual se realiza una consulta de datos hacia un sitio y se guardan de

manera local y, después, en la segunda etapa, se realiza el análisis de estos datos para obtener información.

2.1. Extracción

Técnicas para la extracción de información

- Web bot, Spider, Crawler, Arañas y Rastreadores [11].

  Inspeccionan las páginas web de internet de forma metódica y automatizada. Se usan para rastrear la red. Lee la estructura de hipertexto y accede a todos los enlaces referidos en el sitio web. Son utilizadas la mayoría de las veces para poder crear una copia de todas las páginas web visitadas para que después puedan ser procesadas por un motor de búsqueda; esto hace que se puedan indexar las páginas, proporcionando un sistema de búsquedas rápido

- Plataformas de agregación verticales:

Existen plataformas que tienen el propósito de crear y controlar numerosos robots que están destinados para mercados verticales específicos. Mediante el uso de esta preparación técnica se realiza mediante el establecimiento de la base de conocimientos destinado a la totalidad de plataformas verticales y luego a crearla automáticamente. Medimos nuestras plataformas por la calidad de la información que se obtiene. Esto asegura que la robustez de nuestras plataformas utilizadas consiga la información de calidad y no sólo fragmentos de datos inútiles.

- Reorganización de anotación semántica.

El desarrollo de web scraping puede realizarse para páginas web que adoptan marcas y anotaciones que pueden ser destinadas a localizar fragmentos específicos semánticos o metadatos. Las anotaciones pueden ser incrustadas en las páginas y esto puede ser visto como análisis de la representación estructurada (DOM). Esto permite recuperar instrucciones de datos desde cualquier capa de páginas web.

Herramientas utilizadas en la extracción

- ScraperWiki. Es una plataforma web que permite crear scrapers de forma colaborativa entre programadores y periodistas para extraer y analizar datos públicos contenidos en la web.

- PHP. Cuenta con librerías para realizar web scraping como cURL, el cual permite la transferencia y descarga de datos, archivos y sitios completos a través de una amplia variedad de protocolos, y Crawl, que contiene varias opciones para especificar el comportamiento de la extracción como filtros Content-Type, manejo de cookies, manejo de robots y limitación de opciones.

- Guzzle: Es un framework que incluye las herramientas necesarias para crear un cliente robusto de servicios web. Incluye: descripciones de Servicio para definir las entradas y salidas de una API, iteradores para recorrer webs paginadas, procesamiento por lotes para el envío de un gran número de solicitudes de la manera más eficiente posible. Fué creado usando Symfony2 y emplea la librería cURL de PHP.

- Jsoup de Java: Es una librería para realizar web scraping. Proporciona una API muy conveniente para la extracción y manipulación de datos, utilizando lo mejor de DOM, CSS, y métodos de jQuery similares.

    - Raspa y analiza el código HTML de una URL, archivo o cadena
    - Encuentra y extrae los datos, utilizando el DOM o selectores CSS
    - Manipula los elementos HTML, atributos y texto.
    - Limpia el contenido enviado por los usuarios contra una lista blanca de seguridad, para evitar ataques XSS.
    - Salida HTML ordenada

- Beautifulsoup: Es una biblioteca de Python diseñada para proyectos de respuesta rápida como screen scraping o web scraping. Ofrece algunos métodos simples y modismos de Python para navegar, buscar y modificar un árbol de análisis: una herramienta para la disección de un documento y extraer lo que necesita, además de que no se necesita mucho código para escribir una aplicación. Beautiful Soup convierte automáticamente los documentos entrantes a Unicode y documentos salientes a UTF-8, también trabaja con analizadores de Python populares como lxml y html5lib y permite realizar el recorrido del DOM.

2.2. Análisis

Herramientas para análisis

Algunos ejemplos de herramientas utilizadas para el análisis de texto son:

- myTrama

Es un sistema web que aporta un lenguaje propio de consultas, similar a SQL. Cuenta con una interfaz visual que carga la web objetivo que permite seleccionar los datos mostrando en bloques de pantalla lo que se necesita. El proceso de selección se traduce en la construcción de una consulta en el lenguaje propio de la herramienta, al que denominan Trama-WQL (Web Query Language). Esta consulta puede ser gestionada en modo texto, incluso puede ser escrita desde cero sin tener en cuenta el seleccionador. Ambos, el editor WQL y el seleccionador están sincronizados, por lo que un cambio en uno de ellos repercute en el otro. Un sistema de recolección de la información permite que los tiempos de latencia entre myTrama y la web no afecten a las llamadas a las APIs, que devolverán muy rápido la información que hay en la caché. En caso de que el sistema detecte que los datos son obsoletos, refrescará los datos de

la caché en background. Permite trabajar con data web mining, como por ejemplo agregadores, comparadores o enlazado de datos.

- Gensim

Es una biblioteca de python que proporciona estadísticas escalables de semántica, analiza documentos de texto plano para la estructura semántica y recupera documentos semánticamente similares. Los algoritmos de Gensim, como análisis semántico latente y el de proyecciones aleatorias, descubren la estructura semántica de los documentos, mediante el examen de patrones de co-ocurrencia dentro del cuerpo de documentos de entrenamiento. Estos algoritmos son sin supervisión.

- Natural Language Toolkit (NLTK)

Es un conjunto de bibliotecas y programas para el procesamiento del leguaje natural (PNL) simbólico y estadístico para el lenguaje Python. Proporciona interfaces fáciles de usar para más de 50 cuerpos y recursos léxicos, como WordNet, junto con un conjunto de bibliotecas de procesamiento de texto para la clasificación, análisis y razonamiento semántico.

3. Trabajo relacionado

Proyecto de ley

Proyectosdeley.pe [9] es una aplicación web que muestra, en forma ordenada y accesible, los proyectos de ley presentados en el Congreso peruano. Es su primer intento de abrir la información estatal usando creativamente la tecnología para promover la transparencia. Este proyecto trata de presentar la información sobre los proyectos de ley producidos por el Congreso en una interfaz amigable e intuitiva. En "ProyectosDeLey" almacenan los datos que se extraigan con BeautifulSoup, principalmente los datos que se buscan almacenar son los títulos, autores, fechas de publicación, entre otros datos.

El software de Proyectosdeley se activa automáticamente cada 3 horas y empieza a buscar proyectos nuevos que hayan sido colgados en la web del Congreso. Si los hay, los descarga, parsea y los guarda indexados en la base de datos local. Cuando ya no hay más proyectos por descargar o procesar, empieza a generar los archivos HTML que puedes ver si visitas el sitio web. También genera las páginas web para cada congresista que haya sido autor de al menos un proyecto de ley.

6news Lawrence

6news Lawrence es un noticiero que mostraba en su programa las estadísticas de las elecciones del estado de Kansas en el 2006. Para poder tener la información antes que su competencia ellos realizaron scraping de datos. El proceso de scraping lo realizaron siguiendo los siguientes pasos:

1. Los votos contabilizados, en un principio, eran publicados en una página web privada a la que sólo podían acceder con una determinada IP.
2. Un script descarga la página de resultados cada vez que cambia y sube los resultados en una página pública.
3. Otro script raspa el html de la página publica (usando la librería de Python beautiful soup) e inserta los datos de la página en una base de datos
4. Un tercer script trae la información de la base de datos y escribe una hoja de cálculo en Excel en una url pública.
5. En 6news una ventana de Windows corre un archivo batch, el cual se encarga de descargar el archivo de Excel
6. Finalmente, el sistema on air-graphics (el cual se encarga de mostrar estadísticas y graficas en tiempo real) lee el archivo de Excel para posteriormente mostrarlo en el noticiero.

Con este procedimiento, 6news logró obtener los resultados inclusive hasta 30 minutos más rápido que su competencia.

Análisis de las tendencias políticas basadas en los patrones de web linking: el caso de Grupos políticos en el Parlamento Europeo

Con el fin de conocer la situación política de la Unión Europea (UE), en este proyecto [4] se recogieron diversos tipos de datos sobre enlaces web a sitios web de los 96 partidos que conforman la UE con el fin de encontrar patrones para su estudio. Se utilizaron 2 tipos de enlaces: los in-link que son hipervínculos incrustados en una página que apuntan a otra página; y los co-link que son enlaces incrustados en dos o más sitios que re-direccionan a una misma página.

Los datos Web co-link se visualizaron utilizando escalamiento multidimensional (MDS), mientras que los datos in-link se analizaron con un análisis de dos vías de varianza. Los resultados mostraron que los datos web de hipervínculo reflejaban algunos patrones políticos en la Unión Europea (UE). Los mapas MDS mostraron grupos de partidos políticos a lo largo de líneas ideológicas, históricas, lingüísticas y sociales.

El análisis estadístico basado en in-link confirmó además que había una diferencia significativa a lo largo de la línea de la historia política de un país, de manera que los partidos de izquierda en los antiguos países comunistas recibieron un número considerablemente menor de in-links a sus sitios web que los partidos de izquierda en los países sin una historia de comunismo.

Extracción de posiciones políticas desde textos políticos utilizando palabras como datos

Este artículo, [2], presenta una nueva manera de extraer posiciones políticas de textos políticos que, a los textos, no los ve como discursos sino como datos en forma de palabras. Se comparó este enfoque a los anteriores métodos de análisis de texto y se usó para hacer una replicación de las estimaciones publicadas sobre las posiciones políticas de los partidos en Gran Bretaña e Irlanda, en ambas dimensiones políticas, económicas y sociales.

A continuación se presentaran los pasos a seguir para la extracción y análisis de los textos.

- Paso 1: Se obtienen los textos de referencia con posiciones conocidas a priori.
- Paso 2: Se generan puntajes de palabras de textos de referencia (puntuación de palabras)
- Paso 3: Se obtiene la puntuación de cada texto virgen utilizando puntajes de palabras (textos básicos)
- Paso 4: (opcional) Se transforman las puntuaciones de texto vírgenes para una métrica original.

Para el proyecto se usaron técnicas del algoritmo "Word scoring" las cuales replican con éxito las publicaciones estimadas de política sin los costos sustanciales de tiempo y mano de obra que éstos requieren. Este algoritmo lee archivos de texto y calcula una puntuación en base a un sentido de palabras a partir de una intersección de ese conjunto de palabras y elige el sentido con las mejores puntuaciones.

El algoritmo toma una palabra de referencia y ve cuántas veces coincide en el conjunto de documentos o documento y le da una puntuación según sea la coincidencia, entre menor sea la coincidencia más alto es el puntaje que le dará y entre mayor sea la coincidencia menor puntaje dará. Si por la palabra evaluada hay varios significados en el documento, se considera la elección más cercana a ésta y es tomada como la mejor.

Midiendo opiniones políticas en blogs

En este proyecto, [3], se obtuvieron publicaciones de personas que están muy involucradas en la política, así como también de estadounidenses que normalmente bloguean cosas sobre otros temas, pero que por algún motivo deciden unirse a una conversación política en 1 o más publicaciones.

Se descargaron y analizaron todas las nuevas publicaciones de un blog cada día. La meta especifica es categorizar las publicaciones en 7 categorías únicas: extremadamente negativo (−2), negativo (−1), neutral (0), positivo (1), extremadamente positivo (2), no opinion (NA), y not a blog (NB). La metodología propuesta en este proyecto es:

Primero, se ignoran todas las publicaciones que estén en idiomas diferentes al inglés, lo mismo con las publicaciones que sean spam. Este proyecto se concentró en 4,303 publicaciones de blogs acerca del presidente Bush y 6,468 publicaciones acerca de la senadora Hillary Clinton.

Como segundo paso, se procesó el texto de cada documento, convirtiendo todo a minúscula, removiendo todo signo de puntuación, y derivando las palabras a su origen primitivo, por ejemplo, "consistir, "consistió", "consistencia", "consistiendo" se reduce a su palabra de origen primitivo que sería consistir, logrando reducir la complejidad de la información que se encuentra en el texto.

Por último se resumió el texto pre-procesado como variables dicotómicas, un tipo para la presencia o la ausencia de cada raíz de la palabra (o unigrama), un segundo tipo por cada par de palabras(o bigrama), y un tercero por cada palabra triplete (o trigrama), de esa forma hasta llegar a n-gramas, sólo se mide la presencia o ausencia

de la raíz de las palabras en vez de contarlas todas (la segunda aparición de la palabra "horroroso" en una publicación no provee tanta información como la primera aparición). Incluso así, el número de variables que restan es enorme, en el ejemplo que se tomó de 10771 publicaciones de blogs acerca del presidente Bush y la senadora Clinton incluía 201,676 unigramas únicos, 2,392,027 bigramas únicos, y 5,761,979 trigramas únicos. La forma usual para simplificar más las variables era considerar solo los unigramas dicotómicos que provengan de la raíz de las variables indicadoras.

Una investigación preliminar de análisis sentimental en el discurso político informal

Mullen [1] explica que dada la tendencia a la alza de que las publicaciones en línea se han convertido en comunicaciones estilo mensaje, el discurso político informal es ahora una característica importante dentro del internet, creando un área para la experimentación en técnicas de análisis sentimental. Menciona que, algunas de las preguntas que podríamos preguntar acerca de un texto, además de un simple juicio sobre un tópico, candidato o propuesta, son, por ejemplo:

a) Identificar la afiliación política del escritor,

b) Clasificar el punto de vista político del escritor, de acuerdo a una taxonomía, como izquierda o derecha, y

c) Evaluar el grado de confianza con el cual el escritor expresa su opinión.

En este artículo se realizó un análisis de la efectividad de métodos de clasificación estándar para predecir la afiliación política del blog evaluado. Los resultados obtenidos sugieren que los métodos tradicionales de clasificación de texto son inadecuados para la tarea de análisis sentimental político. Propone realizar un análisis de cómo un post interacciona con otro, esto con el fin de utilizar la información que se tenga de un post para ayudar a clasificar otros.

4. Conclusiones

En este trabajo se realizó una revisión del estado del arte de diferentes metodologías que se han propuesto para realizar análisis político en las redes sociales y en el internet en general. Esta es la primera etapa de un proyecto de investigación en el que se pretende obtener información de diferentes fuentes en internet mediante técnicas de web scraping y analizar lo obtenido mediante técnicas de text mining. Esta revisión nos será de gran utilidad para definir los indicadores necesarios, así como los resultados esperados de nuestro proyecto.

Nuestro objetivo es que cualquier persona pueda obtener información para poder dar una crítica fundamentada, tener una postura basada en diferentes fuentes y conocer noticias importantes con respecto a la política mexicana.

Referencias

1. Mullen, T., Malouf, R.: A Preliminary investigation into sentimental analysis of informal political discourse. AAAI Symposium on Computational Approaches of Analysing Weblogs, pp. 159–162 (2006)
2. Laver, M., Benoit, K., Garry, J.: Extracting Policy Positions from Political Texts Using Words as Data. The American Political Science Review, pp. 311–331 (2003)
3. Hopkins, D. J., King, G.: A method of automated nonparametric content analysis for social science. American Journal of Political Science, vol. 54 (1), pp. 229–247 (2010)
4. Romero-Frías, E., Liwen, V.: The analysis of political trends based on Web linking patterns: The Case of Political Groups in the European Parliament (2009)
5. Matt, T., Pang, B., Lillian, L.: Get out the vote: Determining support or opposition from Congressional floor-debate transcripts Proceedings of EMNLP, pp. 327–335 (2006)
6. Vasilevsky, D.: Parallelized web scraping using RollingCurl (2015)
7. How the Presidential Candidates Use the Web and Social Media, Pew Research Center: Journalism & Media Staff. http://www.journalism.org/2012/08/15/how-presidential-candidates-use-web-and-social-media
8. Ramos, D.: 5 usos electorales de las redes que EU hace y México desdeña (2012)
9. Organizando los proyectos de ley del congreso. http://aniversarioperu.utero.pe
10. What $40 means to Americans across the country. https://www.whitehouse.gov/40dollars
11. Schrenk, M.: Webbots, spiders, and screen scrapers, a guide to developing internet agent with PHP/CUR, 2nd edition (2012)