Práctica 1

Por José Andrés Navarro Yepes y Andreu Fiol Bibiloni

1. Contexto.

Tras analizar diversas páginas web, hemos elegido el sitio ‘electocracia.com’, donde se recoge información sobre sondeos de intención de voto a nivel estatal desde 2017. Esta web fue elegida por lo interesante de sus datos y porque tanto la inexistencia de una API como la permisividad de su archivo ‘robots.txt’ la hacían encontrarse en una ventana ideal para la práctica de *web scraping*.

La información recolectada es sobre los sondeos de intención de voto a nivel estatal, con la estimación porcentual de cada partido. Se recogen sondeos de la práctica totalidad de fuentes y medios gracias a que el equipo que hay detrás de ‘electocracia.com’ persigue un ideal difícil de observar en política como es la objetividad, la integridad, la independencia y la ecuanimidad. En dicha página no hay únicamente información electoral y política, sino también información empresarial (indicadores, publicidad, consultoría de marketing, etc).

2. Definir un título para el data set.

Evolución histórica y estimaciones porcentuales de voto de los principales partidos políticos en las elecciones generales de España según los sondeos (2017-2019).

3. Descripción de la data set.

En el dataset nos encontramos con todos los sondeos de intención de voto, detallando los 5 partidos políticos con mayor influencia de España con sus respectivas estimaciones porcentuales, desde los más recientes (15/04/19) hasta el primero del 2017 (08/01/27) y ofreciendo entre otros datos el medio de comunicación donde se publicó, su fuente y el tamaño de la encuesta.

De igual forma, se encuentran datos relevantes que favorecen un análisis objetivo y claro sobre la credibilidad de todos y cada uno de los sondeos electorales especificando su fuente, fecha de publicación y el medio de comunicación por el cual se ha transmitido.

Cabe decirse también que se ha tenido que realizar data cleaning porque en según qué caso había comas en vez de puntos como separador decimal. También hubo que dar formato a las columnas de fecha.

En cuanto a limitaciones del dataset, por un lado hay factores que limitan el objetivo del proyecto y no se han podido solucionar tales como la incongruencia/inconsistencia en los datos. Al realizar una gráfica sobre las estimaciones de voto se observan diferencias notables con la gráfica de otro dataset también mostrado por 'electocracia.com'. Los datos sobre estimaciones del dataset histórico de esta web parece que en muchos casos no se hallan registrados en la columna correcta. Por ejemplo, según los datos correctos, PSOE empezó a superar al PP en junio de 2018 y el dataset completo muestra un PSOE superior antes de esa fecha. Por esta razón decidimos contactar con los responsables de la web. Les enviamos el siguiente correo electrónico que de momento no ha recibido respuesta:

Histórico de sondeos

José Andrés Navarro Yepes <navarroyepes@gmail.com>
Sáb 13/04/2019 13:36
Usted; info@electocracia.com

Buenas.

Somos José Andrés Navarro y Andreu Fiols. Estamos realizando una práctica de "web scraping" para una asignatura del Máster en Data Science de la Universitat Oberta de Catalunya.

Hemos decidido realizar la práctica basándonos en electocracia.com porque nos parece una web que ofrece datos interesantes y necesarios en la sociedad actual. Por ello, en primer lugar nos gustaría darles las gracias y felicitarles por la iniciativa.

También queríamos comentar una serie de problemas que hemos encontrado en la información del histórico de sondeos. Queremos compartir con ustedes estos problemas con el fin de que la página pueda dar un servicio todavía mejor.

Por ejemplo, al intentar ordenar por fechas no se obtiene un orden adecuado, debido a la falta de formato constante en las fechas. Lo mismo puede decirse de los valores (la estimación porcentual), que varían dependiendo de los decimales contemplados y de los separadores decimales (hay alguna 'coma' en vez de 'punto').

Todo esto puede, sin embargo, solucionarse rápidamente con un proceso de limpieza de datos que hemos llevado a cabo para construir un DataFrame con la información histórica de sondeos de su página.

Sin embargo, hay un problema todavía mayor, y que no puede solucionarse con limpieza de datos. Se trata de que se muestran valores de intención de voto en columnas de partidos que no les corresponden, cambiándolos de sitio. Por ejemplo, y guiándonos por el dataset reflejado en 'cotizaciones', Ciudadanos empezó a superar a Podemos en octubre de 2017. Los datos del 'histórico' no lo reflejan así. Asimismo, PSOE empezó a superar al PP en junio de 2018, pero los datos del 'histórico' muestran un PSOE superior también antes.

Rogamos que puedan solucionar este último punto para que los investigadores y los ciudadanos podamos acceder a unos datos abiertos de calidad.

Muchas gracias por su atención, y enhorabuena de nuevo por la web.

José Andrés Navarro

Por otro lado, los datos también presentan limitaciones si nos alejamos del objetivo principal del proyecto. Por ejemplo, si quisiéramos comparar los resultados reales de las elecciones con nuestro dataset de sondeos, deberíamos recurrir a otra fuente de información adicional, ya que los resultados no se encuentran en él.

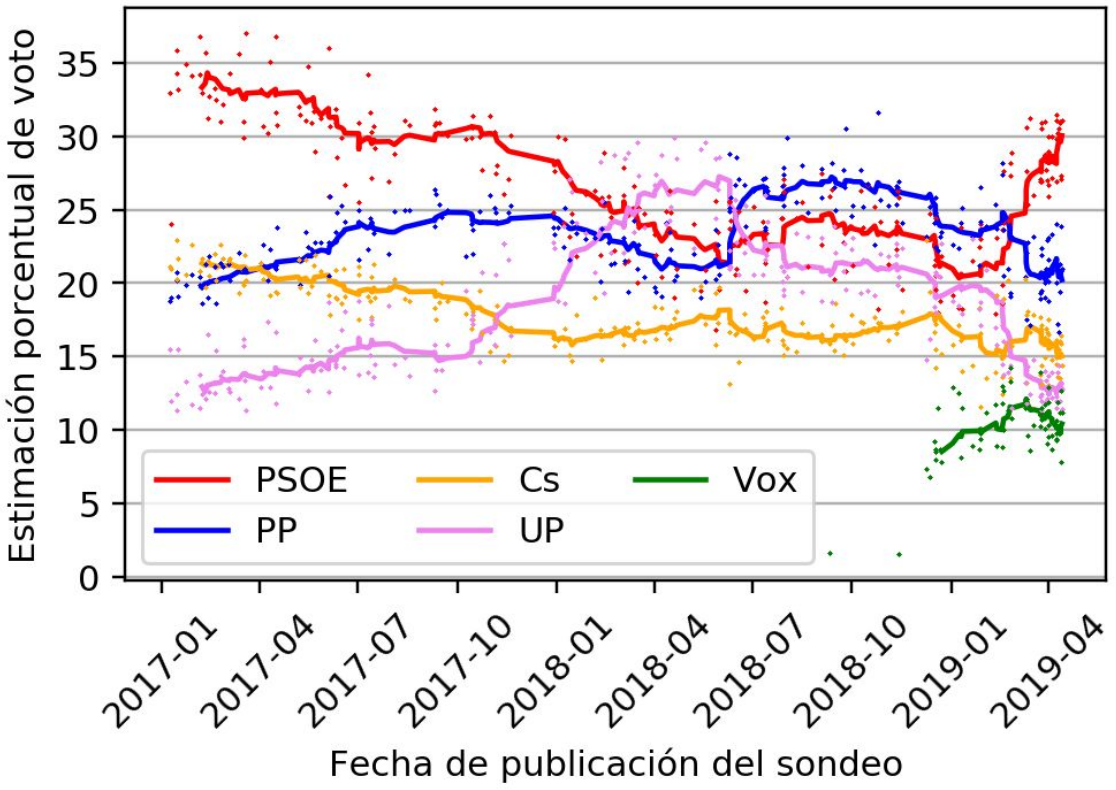
En definitiva, en cuanto a calidad de los datos se refiere, tiene una calidad media-baja porque:

- El dataset es completo en lo que respecta al tema encuesta/sondeo, con muy pocos datos vacíos
- Son datos únicos también, porque ninguna otra web recoge toda esta información sobre los sondeos
- Son puntuales porque representan datos en fechas muy concretas
- Son válidos
- No siempre son exactos (por lo comentado anteriormente sobre la confusión de columnas)
- Son consistentes

4. Representación gráfica.

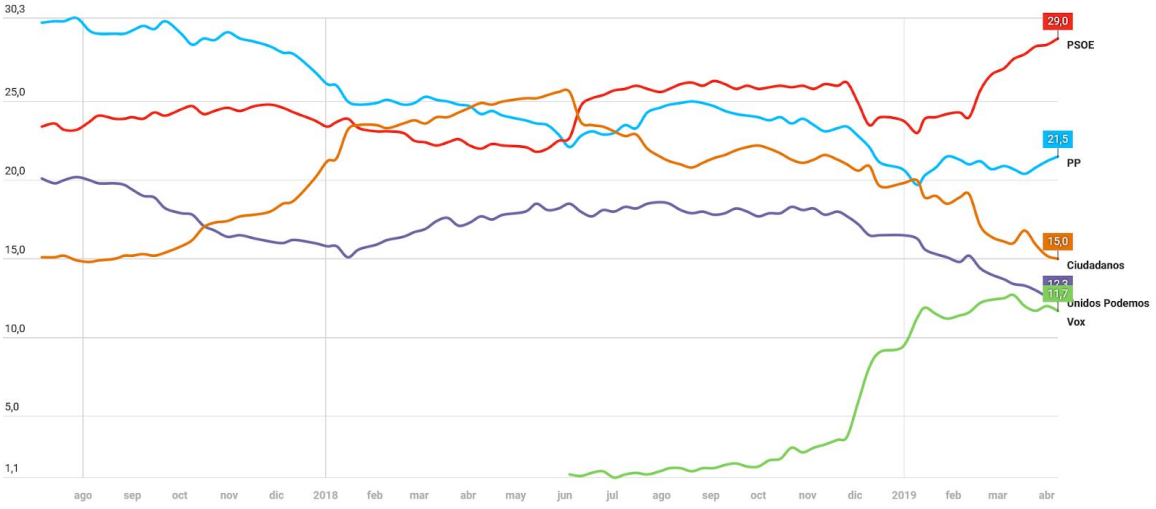
La idea de la representación gráfica es mostrar cronológicamente todas las estimaciones porcentuales de los 5 partidos principales (scatter), al que añadimos una línea con la media móvil de 10 estimaciones. En el eje horizontal se encuentra la fecha de publicación de los sondeos y en el eje vertical el rango de estimaciones porcentuales. Cada línea representa un partido y cada punto una estimación en un sondeo determinado.

Representación gráfica en base a nuestro dataset



El gráfico siguiente es una captura de pantalla del gráfico generado por electocracia.com con DataWrapper en base a un dataset distinto, en el cual se ve que la evolución es muy diferente en comparación con nuestro dataset.

Histórico de cotizaciones por Electocracia



5. Contenido.

Se recogen las siguientes características:

- MEDIO: el medio de comunicación por el que se ha informado del sondeo
- FUENTE: la fuente demoscópica que ha realizado el sondeo
- PUBLICACIÓN: fecha de publicación del sondeo
- FIN CAMPO: última fecha en la que se cogieron encuestas para el sondeo
- TAMAÑO: número de encuestas realizadas en cada sondeo
- PSOE: Estimación porcentual de votos del Partido Socialista Obrero Español
- PP: Estimación porcentual de votos del Partido Popular
- Cs: Estimación porcentual de votos del partido Ciudadanos
- UP: Estimación porcentual de votos del partido Unidos Podemos
- Vox: Estimación porcentual de votos del partido político Vox
- Otros/Blancos: Estimación porcentual de votos del resto de partidos políticos y votos en blanco
- MUESTRA, FECHA, FIABILIDAD y VALOR (1-5): valores que el equipo de electocracia.com establece basándose en las características de cada sondeo, con el objeto de medir el peso de los mismos en sus propias estimaciones.

Los datos han sido recolectados mediante técnicas de web scraping desarrolladas en el fichero de código para coger los datos de la página web HTML.

Antes de empezar, el propio código comprueba, con la herramienta 'robotparser', si el archivo 'robots.txt' nos permite realizar el scraping deseado. De no ser así, el código hará que se muestre un mensaje informando de ello.
Al empezar el web scraping, se han realizado los primeros pasos propuestos por Subirats & Calvo (2019) en la cuestión de envío de peticiones HTTP. Luego, se ha procedido a descargar la página web mediante la librería *requests,* para así analizar la página web con la librería 'BeautifulSoup'. Observando la información extraída con esta librería pudimos comprobar que los datos deseados se encuentran en una tabla html con el id 'tablepress-2', que recogemos. Acto seguido, se arregla la cuestión de los separadores decimales y pasamos de la tabla html a un dataframe operable aprovechando una función de la librería 'pandas'. Seguidamente cambiamos los nombres de las columnas de los partidos porque en la web aparecen con una imagen de su símbolo. Damos entonces fomato a las columnas de fechas y por último generamos el archivo CSV solicitado.
Una vez hecho esto, también hemos realizado una representación gráfica de los principales datos del set con pyplot.

Las librerías utilizadas son :
robotparser, requests, BeautifulSoup, html5lib, lxml para web scraping, pandas para tratar los datos, y matplotlib para la representación gráfica.

6. Agradecimientos.

Agradecemos al equipo de Electocracia (José Miguel Silva, José Manuel San Millán, Rubén Rodríguez y Miguel Ángel Ramos) por ofrecer información política de calidad e independiente. En cuanto a análisis similares, no se ha encontrado ningún análisis previo con el mismo objetivo y/o concebido en el mismo espacio temporal. Se menciona lo último porque lo poco que más se acerca al proyecto son los siguientes trabajos académicos correspondientes a los años 2013 y 2015:

- https://dialnet.unirioja.es/servlet/articulo?codigo=4314470
- http://www.fes-sociologia.com/files/congress/12/papers/3677.pdf

7. Inspiración.

En primer lugar hemos considerado el interés de este dataset por ser un tema de extrema actualidad (el último sondeo es de hoy mismo 15/04/2019) y por el interés que despierta la situación política actual. El dataset posee la información de 233 sondeos desde el 2017 hasta hoy, lo cual da para contestar muchas preguntas de evolución e históricas, como observar la evolución de la intención de voto de los distintos partidos nacionales. Al haber bastantes registros, el dataset ofrece bastante información.

Este dataset puede ayudar a partidos políticos, a ciudadanos curiosos, a analistas políticos, a medios de comunicación o incluso a las fuentes demoscópicas de los mismos sondeos para realizar autocrítica (si fuera necesario) y tener una visión completa, objetiva e integrada de los datos proporcionados por todos los sondeos electorales desde 2017.

También nuestro dataset puede ser usada por los mismos que realizaron los análisis anteriores para comparar con los resultados reales de las elecciones generales y comprobar en mayor medida la credibilidad y el grado de acierto real de las fuentes realizadoras de los sondeos.

Asimismo, al poder realizar análisis controlando por las fuentes demoscópicas, se puede extraer una mejor información sobre las tendencias reales en intención de voto.

Otra idea también sería que podría ser usado por analistas en el cual monitorizan las principales cuentas de las redes sociales de los partidos y establecen comparaciones con nuestro dataset para estudiar la causalidad entre el resultado en los sondeos según lo posteado en Twitter, dicho en una noticia, etc.

8. Licencia.

Hemos escogido la licencia CC BY-SA 4.0 por las siguientes razones:

- Consideramos importante reconocer el trabajo del equipo de electocracia.com y con esta licencia se reconoce el trabajo original y si se han hecho aportaciones o cambios.
- Se puede hacer un uso comercial, así es más probable que se dé a conocer más este sitio web que cede datos de los sondeos de forma gratuita e independiente.
- Se debe compartir igual, lo que significa que si se cambia o transforma el proyecto deberá seguirse la misma licencia, lo que promueve este trabajo del autor original y darse más a conocer.

9. Dataset y Código.

El proyecto ha sido desarrollado en Python y tanto el código como el CSV se encuentra en el siguiente enlace al repostorio Git Hub (https://github.com/navarroyepes/TCVDPRAC1).

Contribuciones	Firma
Investigación previa	AFB, JANY
Redacción de las respuestas	AFB, JANY
Desarrollo código	AFB, JANY

Recursos

- Electocracia.com [Consulta: 15 de abril de 2019].
- Masip, D. (2012). El lenguaje Python. Barcelona: Editorial UOC.
- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.
- Subirats, L. y Calvo, M. (2019). Web Scraping. Barcelona: Editorial UOC.