Práctica 1: Web scraping

Antonio Sánchez Navarro

Dataset: Examinando un artículo del New York Times

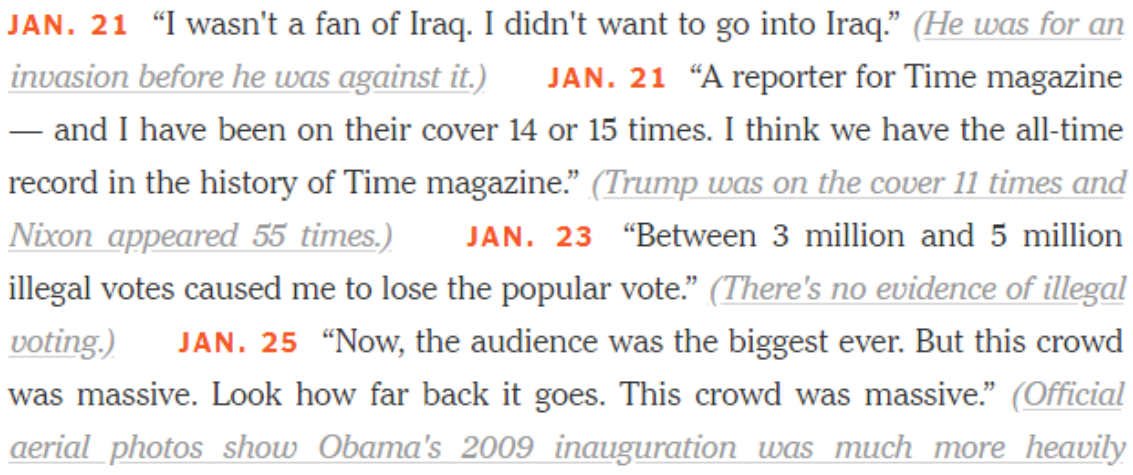
Antonio Sánchez Navarro

Octubre de 2018

Descripción

El conjunto de datos que se ha generado en esta actividad práctica sintetiza un conjunto de características procedentes de un artículo del New York Times, aparecido en diciembre de 2017. Dicho artículo es un catálogo de casi todas las mentiras que el presidente Donald Trump ha dicho públicamente desde que asumiera el cargo. Las variables que se recogen en el conjunto de datos son la fecha en que se dijo la mentira, la mentira en si misma, una explicación de por qué se puede calificar de mentira y la URL del artículo que defiende dicha justificación.

Imagen identificativa



JAN. 21 "I wasn't a fan of Iraq. I didn't want to go into Iraq." *(He was for an invasion before he was against it.)* **JAN. 21** "A reporter for Time magazine — and I have been on their cover 14 or 15 times. I think we have the all-time record in the history of Time magazine." *(Trump was on the cover 11 times and Nixon appeared 55 times.)* **JAN. 23** "Between 3 million and 5 million illegal votes caused me to lose the popular vote." *(There's no evidence of illegal voting.)* **JAN. 25** "Now, the audience was the biggest ever. But this crowd was massive. Look how far back it goes. This crowd was massive." *(Official aerial photos show Obama's 2009 inauguration was much more heavily*

Datos que me propongo extraer de la página web

Contexto

El conjunto de datos es un compendio de mentiras que el presidente Donald Trump ha dicho públicamente desde que asumiera la presidencia. Los datos recogidos cubren el año 2017 casi en su totalidad. El ejercicio realizado me ha parecido interesante desde el primer momento, ya que entiendo que no hay precedentes de presidentes norteamericanos que dedicasen tanto tiempo a decir falsedades. Ningún otro presidente, republicano o demócrata, se ha portado nunca como Trump se está comportando. Un conjunto de datos como el que he podido crear nos muestra que algunos mandatarios intentan crear una atmósfera en la que la realidad es irrelevante. Considero este ejercicio un ejemplo de caso de uso exitoso del web scraping en un contexto de ciencia sociopolítica, ya que permite vislumbrar los sentimientos de la población norteamericana y el talante político de un conocido periódico.

Contenido

Para cada noticia falsa, que se corresponde con un registro o fila completa del conjunto de datos creado, se recogen las siguientes características:

• **FECHA**: día en que se dijo públicamente la mentira, en el formato aaaa/mm/dd.
• **MENTIRA**: frase o sentencia que representa en si misma una falsedad.
• **EXPLICACION**: justificación de por qué estamos ante una sentencia falsa o mentira.
• **URL**: URL incrustada en el texto original que apoya la justificación antes comentada.

El periodo de tiempo considerado se extiende del 21 de enero al 11 de noviembre de 2017, es decir, casi la totalidad del año 2017. Los datos se recogen del artículo NYT “Trump’s lies”, cuya página web es:

https://www.nytimes.com/interactive/2017/06/23/opinion/trumps-lies.html

Agradecimientos

Los propietarios del conjunto de datos son David Leonhardt y Stuart A. Thompson, ambos periodistas del New York Times. El primero es columnista y el segundo es el director de gráficos de la sección de opiniones en el periódico. Otras fuentes de información utilizadas y que se listan en el enlace al sitio web son *Politifact; Factcheck.org*; *The Washington Post Fact Checker* y *The Toronto Star*. Les agradezco a todos el hecho de hacer público el conjunto de datos.

Inspiración

El presente conjunto de datos es de gran valor en el ámbito periodístico y de las ciencias sociopolíticas. Disponer de un dataset que recoja todas las falsedades o verdades a medias de una personalidad política relevante es disponer de un arma poderosa para prevenir a las masas y para evitar futuras selecciones erróneas de mandatarios. El conjunto de datos se centra en Donald Trump, pero se podría aplicar a muchos más protagonistas de la vida política actual. Es bastante habitual que se oigan promesas que luego se acaban desmintiendo.

Como preguntas que me gustaría responder a la comunidad, destacaría cuáles han sido los meses en los que se detecta el máximo (o mínimo) de mentiras, plantearía un debate sobre si el ascenso de Trump al poder se ha fundamentado en mentiras y revelaría cuáles son las actividades habituales de Trump los días que no proclama públicamente ninguna mentira. A más de uno le resultaría cuando menos interesante saber que, en esos periodos, Trump se ausenta de Twitter, pasa sus vacaciones en Florida o se encuentra "ocupado" jugando al golf. Otra pregunta interesante por responder es hasta qué punto la acusación de injerencias por parte de Rusia en las elecciones y su ascenso al poder ha influido en la cantidad de mentiras que ha llegado a decir, mermando dicha cantidad o aumentándola.

Licencia

La licencia escogida para la publicación de este conjunto de datos es **CC BY-SA 4.0 License**. Los motivos que han llevado a la elección de esta licencia tienen que ver con la idoneidad de las cláusulas que esta presenta en relación con el trabajo realizado:

• *Se debe proveer el nombre del creador del conjunto de datos generado, indicando los cambios que se han generado.* De esta manera, se reconoce el trabajo ajeno y en qué medida se han realizado aportaciones en relación con el trabajo original.

• *Se permite un uso comercial*. Esto haría que incrementen las probabilidades de que una empresa utilice los datos generados y se realicen trabajos de calidad que reporten cierto reconocimiento al autor original.

• *Las contribuciones realizadas a posteriori sobre el trabajo publicado bajo esta licencia deberán distribuirse bajo la misma*. Esto hace que el trabajo del autor original continúe distribuyéndose bajo los términos que él mismo planteó.

Código

Se adjunta el código R y también en Python que ha permitido la extracción de datos de la web.

Dataset

Se adjunta el dataset generado en formato CSV, con el nombre “articuloNYTanalisis.csv”. Los campos se separan con coma simple, ‘,’.

Recursos

1. Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.
2. Masip, D. El lenguaje Python. Editorial UOC.
3. Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.
4. Simon Munzert, Christian Rubba, Peter Meißner, Dominic Nyhuis. (2015). Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining. John Wiley & Sons.
5. Tutorial de Github https://guides.github.com/activities/hello-world.
6. https://www.nytimes.com/interactive/2017/06/23/opinion/trumps-lies.html