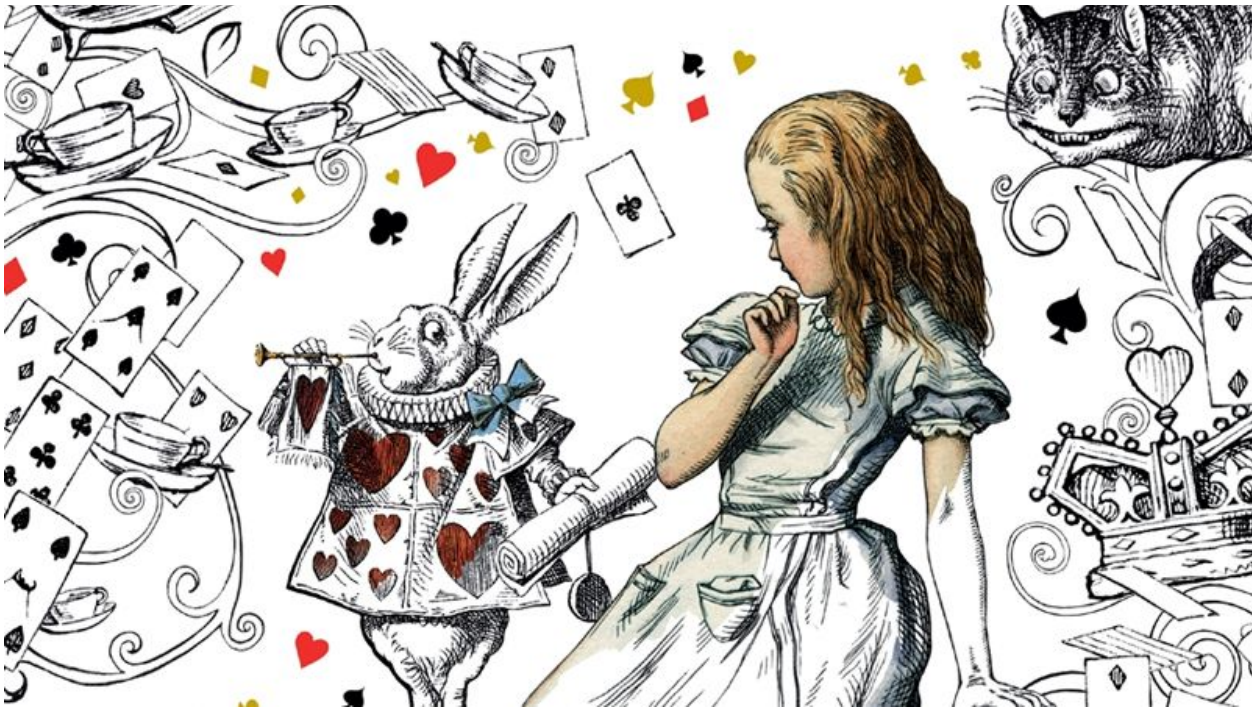
1.Título del dataset. Poned un título que sea descriptivo.

Alice's Adventures in Wonderland most frequent words

2. Subtítulo del dataset. Agregad una descripción ágil de vuestro conjunto de datos por vuestro subtítulo.

List of the most frequent words in the Lewis Carroll's book.

3. Imagen. Agregad una imagen que identifique vuestro dataset visualmente



4. Contexto. ¿Cuál es la materia del conjunto de datos?

La materia del conjunto de datos es una novela clásica de la literatura inglesa contenida en la web de Project Gutenberg, donde se recogen la mayoría de clásicos de la literatura.
En concreto, el dataset contiene una lista de las palabras que salen en el libro eliminando los Stopwords (NLTK), concepto que hace referencia a las palabras más comunes de un lenguaje (en este caso inglés) que no aportan valor al análisis.

5. Contenido. ¿Qué campos incluye? ¿Cuál es el periodo de tiempo de los datos y cómo se ha recogido?

Los campos que contiene son dos variables que recogen un índice de las filas y cada una de las palabras que han salido en el scraping. El periodo de tiempo es inmediato, no hemos

utilizado ningún listener para captar los datos que se iban generando en la web sino que ya estaban presentes desde el inicio. La captación se ha realizado con la libreria request haciendo una petición al servidor web y procesando los datos con BeautifulSoup con un hmtl_parser

6. Agradecimientos. ¿Quién es propietario del conjunto de datos? Incluid citas de investigación o análisis anteriores.

El dataset previene de la web www.gutenberg.org

7. Inspiración. ¿Por qué es interesante este conjunto de datos? ¿Qué preguntas le gustaría responder la comunidad?

Me ha inspirado este proyecto los numerosos artículos que llevo leyendo sobre text mining y natural language processing y quería practicar con esas librerías para desarrollar un análisis de las palabras eliminando stopwords. Un primer paso para continuar con otros puntos en prácticas posteriores.

8. Licencia. Seleccionad una de estas licencias y decid por qué la habéis seleccionado:
- **Released Under CC0: Public Domain License**
- **Released Under CC BY-NC-SA 4.0 License**
- **Released Under CC BY-SA 4.0 License**
- **Database released under Open Database License, individual contents under Database Contents License**
- **Other (specified above)**
- **Unknown License**

La licencia que he escogido es CC0: Public Domain License porque es un dataset que ha sido creado por mi pero los valores no son originales sino que provienen de una fuente pública por lo tanto considero que debería quedarse público de igual modo.

9. Código: Hay que adjuntar el código con el que habéis generado el dataset, preferiblemente con R o Python, que os ha ayudado a generar el dataset

El código está colgado junto con el dataset .csv() en el enlace

https://github.com/NowelBcn/WordFrequencyAnalysis