IRWA Project: Part 1 - Procesamiento de texto y análisis exploratorio de los datos

Git URL: https://github.com/raquel-sb/IRWA_2023.git

1. Pre-Process the Code

1.1. Introducción

La primera parte del proyecto consta de limpiar y procesar el conjunto de datos proporcionados. Este consta de una serie de tweets, donde primero realizaremos un preprocesamiento tratando principalmente la eliminación de signos de puntuación, eliminación de textos alfanuméricos, tokenización de texto, etc.

A partir de este pre procesamiento de los datos, obtendremos un conjunto nuevo ya limpio y práctico que utilizaremos en el resto del proyecto.

1.2. Formato de los datos originales

Inicialmente tenemos dos ficheros: un json con los distintos tweets y un csv que contiene los ids de los tweets y la nomenclatura doc_xxx.

Los campos para cada tweet en el fichero json son los siguientes:

```
created_at, id, id_str, full_text, truncated, display_text_range, entities, metadata, source, in_reply_to_status_id, in_reply_to_status_id_str, in_reply_to_user_id, in_reply_to_user_id_str, in_reply_to_screen_name, user, geo, coordinates, place, contributors, is_quote_status, retweet_count, favorite_count, favorited, retweeted, lang
```

1.3. Formato deseado de los datos

Para obtener el formato deseado de los datos, en nuestro caso, nos quedaremos únicamente con los siguientes campos:

- **ID:** el identificador del tweet.
- **Tweet:** el campo full_text, el cual es el texto completo del tweet.
- **PreProcessed_Tweet:** el campo full_text una vez aplicadas las transformaciones.
- **Username:** el campo user, que corresponde al nombre de usuario que ha publicado el tweet.
- **Date:** el campo created_at, que es la fecha en la que se publicó el tweet.
- **Hashtags:** el campo hashtags que se encuentra dentro de entities. Hemos creado una función para extraer esta información (getHashtagsFromTweet), la cual muestra los hashtags con '#'.
- **Processed_Hashtags:** los hashtags una vez aplicadas las transformaciones a la lista original.
- **Likes:** el campo favourite_count, que representa el número de likes que ha recibido el tweet.

- **Retweets:** el campo retweet_count, que equivale al número de veces que ese tweet ha sido retuiteado.
- **Url:** este campo debemos construirlo siguiendo el siguiente esquema: https://twitter.com/screen_name/status/id_str.

2. Funciones desarrolladas

2.1. Función clean(text)

Primero transformamos el texto a lower case (para eliminar mayúsculas), seguidamente quitamos las urls y hashtags y finalmente los carácteres no alfanuméricos.

2.2. Función build_terms(text)

Después de aplicar la función "clean" para depurar el tweet, procedemos a la tokenización con el fin de obtener una lista de palabras. Posteriormente, eliminamos los "stopwords," tomando en consideración y asumiendo que todos los tweets están redactados en inglés. Por último, llevamos a cabo un proceso de "stemming" con el objetivo de conservar la raíz de las palabras "compuestas".

2.3. Función getHashtagsFromTweet_Original(tweet)

En primer lugar, determinamos la cantidad de hashtags empleados en el tweet en cuestión. A continuación, creamos una lista vacía destinada a almacenar estos hashtags. Para lograrlo, implementamos un bucle for que itera a través de cada hashtag, guardándolo junto con el carácter #. En última instancia, devolvemos la lista completa de hashtags que acabamos de conformar.

2.4. Función getHashtagsFromTweet(tweet)

Contrario a la función previa, en esta etapa llevamos a cabo un preprocesamiento de los hashtags con el propósito de desglosarlos en palabras clave. Esto nos permite facilitar un análisis futuro, en caso de que sea necesario.

2.5. Función getTweetInfo(tweet)

Esta función toma como entrada un solo tweet y devuelve el tweet modificado. Recopilamos los campos mencionados anteriormente, y es aquí donde empleamos las funciones de preprocesamiento tanto para los hashtags como para el texto del tweet.

En resumen, la lista de campos que almacenamos en el `tweet_dict`, que luego devolvemos, se presenta en la imagen adjunta.

2.6. Función string_concat(stringList)

Empleamos esta función para combinar una lista de cadenas en una sola. Específicamente, la utilizamos para unir el "PreProcessed_Tweet", ya que almacena el tweet preprocesado como una lista de palabras, lo cual consideramos útil de conservar. Al mismo tiempo, queremos la capacidad de imprimir esta lista como una única cadena para realizar comparativas entre el tweet original y el preprocesado.

2.7. Función tweets_dict(json_doc, csv_doc)

Esta función posibilita la iteración a través de cada uno de los tweets presentes en el archivo JSON. Realiza la correspondencia entre el ID del tweet y el documento "doc_xxx" del archivo CSV, extrae la información deseada mediante la función "getTweetInfo(t)", y almacena los resultados en un diccionario donde la clave corresponde a "doc_xxx". Finalmente, devuelve este nuevo diccionario de tweets.

El resultado obtenido es:

```
get_tweets = tweets_dict(data_json, data_csv)
get_tweets['doc_3904']
```

```
{'ID': 1575164742859378689,
 'Tweet': 'Whether you are visiting Nigeria or you living in Nigeria, we understand the importance of information; we know that a lot of our customers sometimes are looking for
ideas of where to go and spend their leisure.\n#WelcomeToIndonesia_NCTDREAM #logistics #usa #UkraineRussiaWar #uk https://t.co/T3I9gNVpne',
 'PreProcesed_Tweet': ['whether',
  'visit',
  'nigeria',
  'live',
  'nigeria',
  'understand',
  'import',
  'inform',
  'know',
  'lot',
  'custom',
  'sometim',
  'look',
  'idea',
  'go',
  'spend',
  'leisur'],
 'Username': '@simpreslogistis',
 'Date': '28/09/2022 16:45:14',
 'Hashtags': ['#WelcomeToIndonesia_NCTDREAM',
  '#logistics',
  '#usa',
  '#UkraineRussiaWar',
  '#uk'],
 'Processed_Hashtags': ['Welcome To Indonesia N C T D R E A M',
  'logistics',
  'usa',
  'Ukraine Russia War',
  'uk'],
 'Likes': 3,
 'Retweets': 0,
 'URL': 'https://twitter.com/simpreslogistis/status/1575164742859378689'}
```

2.8. Función separate_by_words(input_string)

Esta función la utilizamos dentro de la función getHashtagsFromTweet(tweet) para separar cada hashtag por palabras.

3. Tratamiento de los hashtags

En primer lugar, optamos por incorporar un campo denominado "hashtags", que consiste en una lista de los hashtags empleados en el tweet. Esto nos proporciona un registro del conjunto original de hashtags utilizado en cada mensaje.

Además, tomamos la decisión de crear otra lista de hashtags, denominada "processed_hashtags", la cual separa los hashtags en palabras individuales. De esta manera, en lugar de conservar el hashtag original, mantenemos uno procesado para facilitar futuras búsquedas y análisis.

Por último, decidimos preservar los hashtags en el cuerpo de los tweets (los mensajes), manteniendo el símbolo #.

Original hashtag

```
['#WelcomeToIndonesia_NCTDREAM', '#logistics', '#usa', '#UkraineRussiaWar', '#uk']
```

Processed hashtag

```
['Welcome To Indonesia N C T D R E A M', 'logistics', 'usa', 'Ukraine Russia War',
'uk']
```

4. Tweet original vs Pre-procesado

Después de procesar el contenido de cada tweet, obtenemos una lista de palabras utilizadas en él, siguiendo los pasos de limpieza descritos anteriormente. A continuación, presentamos un ejemplo que ilustra la transformación de un tweet, mostrando tanto la versión original como la versión procesada.

Original tweet

```
Whether you are visiting Nigeria or you living in Nigeria, we understand the importance of information; we know that a lot of our customers sometimes are looking for ideas of where to go and spend their leisure. #WelcomeToIndonesia_NCTDREAM #logistics #usa #UkraineRussiaWar #uk https://t.co/T3I9gNVpne
```

Pre-processed tweet

```
whether visit nigeria live nigeria understand import inform know lot custom sometim look idea go spend leisur welcometoindonesianctdream logist usa ukrainerussiawar uk
```

5. Exploratory data analysis

5.1. Average Sentence Length

Resultado obtenido:

```
The average sentence length of a tweet is: 11 words.
```

Análisis:

Los tweets tienen una longitud promedio de 11 palabras por oración. Esto indica que los usuarios de Twitter que participan en debates relacionados con este tema prefieren expresarse de manera concisa y directa. Además, esta brevedad es común en plataformas como Twitter, donde los mensajes son limitados en caracteres; así pues, los usuarios intentan comunicarse de manera efectiva en un espacio limitado.

5.2. TOP 10 Words (with their frequencies)

Resultado obtenido:

```
--Top 10 words and their frequencies--

Word: ukrain - Frequency: 1088
Word: russian - Frequency: 1022
Word: russia - Frequency: 609
Word: the - Frequency: 563
Word: putin - Frequency: 510
Word: ukrainian - Frequency: 469
Word: war - Frequency: 467
Word: forc - Frequency: 277
Word: i - Frequency: 267
Word: region - Frequency: 253
```

Análisis de las palabras más frecuentes:

Las palabras más utilizadas en los tweets relacionados con el conflicto son "Ucrania", "ruso", "Rusia", "Putin", "ucraniano", "guerra", "fuerza", "yo", y "región". Estas palabras clave resaltan los temas principales de la conversación sobre la guerra.

Es evidente que los términos "Ucrania" y "Rusia" son cruciales debido a que se refieren a los países que están involucrados en el conflicto.
Debido a su influencia en la toma de decisiones en Rusia, "Putin" es una figura importante en las conversaciones.
Las palabras "guerra" y "fuerza" sugieren una conversación constante sobre los aspectos militares del conflicto.
"Región" podría significar que se discuten áreas específicas del conflicto.

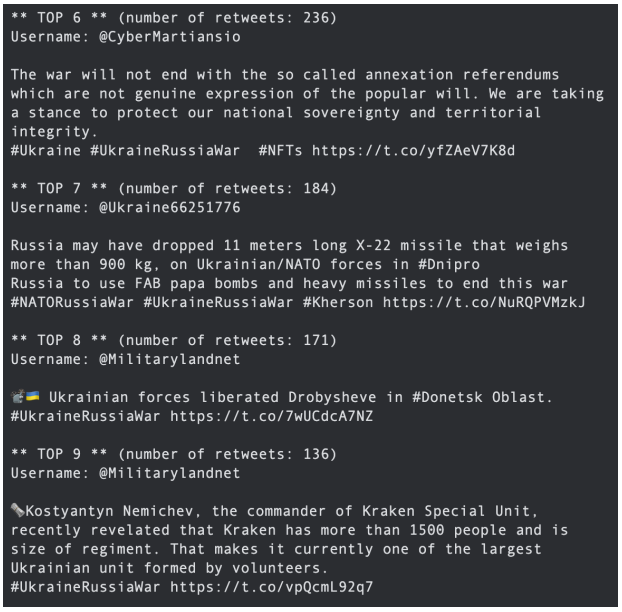
Estas palabras clave dan una idea básica de los temas más discutidos.

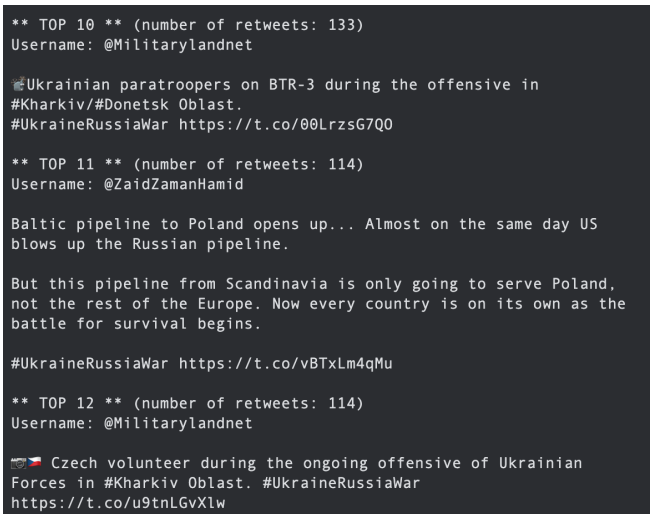
5.3. Most Retweeted Tweets

Resultado obtenido:

```
--We consider the most retweeted tweets the tweets that have more
than 100 retweets.--
So the most retweeted tweets are:

** TOP 1 ** (number of retweets: 646)
Username: @Militarylandnet

🗺️Situation around Lyman - Sep 30 11:00:
- UA forces liberated Yampil and advancing north
- RU troops are reportedly abandoning its positions in Drobysheve
- The only exit route from Lyman is within the firing range of UA
forces
#UkraineRussiaWar https://t.co/jGJUhXcr1y

** TOP 2 ** (number of retweets: 338)
Username: @Militarylandnet

📷Unique and rare photos of Ukrainian forward command post during
the offensive in #Kharkiv Oblast. News reporters aren't usually
invited to such places, but here seems to be an exception.
#UkraineRussiaWar https://t.co/AmSijyM59c

** TOP 3 ** (number of retweets: 283)
Username: @Militarylandnet

💣Operation Interflex: Ukrainian recruits continue to master their
skills under the guidance of British and Canadian instructors in
the UK.
#UkraineRussiaWar https://t.co/oYWThs8qNe
```

```
** TOP 4 ** (number of retweets: 251)
Username: @OSINTschizo

The following countries have urged their citizens to leave 🇷🇺 will
update if other governments make similar statements.
#UkraineRussiaWar #AnnexationofUkraine
#NAFO

Poland 🇵🇱
Estonia 🇪🇪
Latvia 🇱🇻
Italy 🇮🇹
United States 🇺🇸
Bulgaria 🇧🇬
Romania 🇷🇴
Taiwan 🇹🇼
Canada 🇨🇦
Portugal 🇵🇹

** TOP 5 ** (number of retweets: 247)
Username: @Militarylandnet

💣Russians shelled the outskirts of #Zaporizhzhia and hit a civilian
humanitarian convoy heading towards the occupied parts. 23 people
were killed, a dozen more wounded.
#UkraineRussiaWar https://t.co/365j43jy51
```





Análisis:

En resumen, los tweets más retuiteados incluyen actualizaciones en tiempo real, contenido visual impactante y declaraciones sobre la situación internacional. El alto número de retweets en estos mensajes demuestra la importancia de mantenerse informado y aumentar la conciencia sobre los conflictos en la comunidad en línea.

Es interesante remarcar que la mayoría de los tweets más retuiteados provienen del mismo usuario, @Militarylandnet, que también opera un sitio web con el mismo nombre, "MilitaryLand." Esta observación sugiere varias conclusiones significativas como:

- La concentración de retweets en un único usuario, @Militarylandnet, resalta su papel central en la difusión de información sobre el conflicto entre Rusia y Ucrania en Twitter.
- El usuario demuestra un enfoque en la cooperación internacional y utiliza hashtags para facilitar la participación en la conversación en línea.
- La presencia de un sitio web sugiere un compromiso más profundo con la cobertura del conflicto y la promoción de recursos adicionales para la comunidad en línea interesada en este tema.

5.4. Most Frequent Hashtags (WordCloud)

Para llevar a cabo el análisis de los hashtags más frecuentes, nos hemos ayudado de la representación gráfica de WordCloud.

Un "Wordcloud" o "nube de palabras" es una representación visual de un conjunto de palabras donde el tamaño de cada palabra representa su frecuencia de aparición. En el contexto de los hashtags en redes sociales, un wordcloud de hashtags mostraría los hashtags más utilizados y su popularidad relativa. Es decir, contra mayor sea el tamaño del hashtag en el wordcloud, mayor frecuencia tendrá.

Resultado obtenido:

Análisis de los hashtags más frecuentes:

- **Hashtags relacionados con el conflicto en general**:

#UkraineRussiaWar (Frecuencia: 3851)

#Ukrainerussiawar (Frecuencia: 98)

#UkraineWar (Frecuencia: 1101)

#RussianUkrainianWar (Frecuencia: 103)

#RussiaUkraineWar (Frecuencia: 146)

Estos hashtags se centran en el conflicto en su conjunto y son utilizados para englobar conversaciones generales sobre la guerra entre Ucrania y Rusia. La frecuencia de estos hashtags se debe a que proporcionan una manera efectiva de agrupar discusiones relacionadas con el conflicto y mantener a los usuarios informados sobre sus desarrollos.

- **Hashtags que mencionan actores clave**:

#Putin (Frecuencia: 527)

#Biden (Frecuencia: 146)

#Zelensky (Frecuencia: 215)

Estos hashtags se refieren a figuras clave involucradas en el conflicto. #Putin hace referencia a Vladimir Putin, el presidente de Rusia; #Zelensky a Volodymyr Zelensky, el presidente de Ucrania; y #Biden a Joe Biden, el presidente de los Estados Unidos. La frecuencia de estos hashtags demuestra la importancia de estas figuras en la toma de decisiones y la diplomacia relacionada con el conflicto.

- **Hashtags relacionados con la geografía**:

#Kherson (Frecuencia: 434)

#Donetsk (Frecuencia: 96)

#Kharkiv (Frecuencia: 318)

#Kyiv (Frecuencia: 88)

#Lyman (Frecuencia: 194)

Estos hashtags hacen referencia a ubicaciones geográficas clave en el conflicto. La alta frecuencia de estos hashtags se debe al interés de los usuarios en conocer los desarrollos en estas áreas específicas y cómo están siendo afectadas por el conflicto.

- **Hashtags que expresan apoyo y solidaridad**:

#StandWithUkraine (Frecuencia: 250) #UkraineWillWin (Frecuencia: 190)

Estos hashtags son utilizados para expresar apoyo y solidaridad hacia Ucrania en medio del conflicto. La frecuencia de estos hashtags refleja la necesidad de mostrar apoyo en línea y mantener viva la esperanza de una resolución pacífica.

- **Hashtags que critican a Rusia**:

#RussiaIsATerroristState (Frecuencia: 257) #RussiaRussiaWar (Frecuencia: 74)

#Russiaisateroriststate (Frecuencia: 67)

#RussiaInvadedUkraine (Frecuencia: 172)

#RussiaUkraineWar (Frecuencia: 146)

Estos hashtags expresan críticas y condenas hacia Rusia en el contexto del conflicto. La alta frecuencia de estos hashtags refleja la intensidad de las opiniones y emociones de quienes los utilizan, así como el enfoque en los actos percibidos como agresivos por parte de Rusia.

- **Hashtags relacionados con la OTAN**:

#NATO (Frecuencia: 439) #NATORussiaWar (Frecuencia: 114)

Estos hashtags se refieren a la OTAN y su papel en relación con el conflicto. La frecuencia de #NATO puede deberse al interés en la respuesta de la OTAN al conflicto y su implicación en la región, lo que puede influir en la estabilidad y seguridad.

- **Hashtags sobre el Nord Stream**:

#NordStream2 (Frecuencia: 167) #Nordstream (Frecuencia: 74)

#Nordstream1 (Frecuencia: 68)

Estos hashtags se centran en el proyecto de gasoducto Nord Stream y sus implicaciones en el conflicto. La frecuencia de estos hashtags sugiere un interés en los aspectos económicos y energéticos del conflicto, ya que el suministro de energía es un factor crítico en el contexto geopolítico.

<u>Conclusión:</u>

En resumen, la variedad de hashtags que se utilizan en la conversación en Twitter sobre el conflicto entre Rusia y Ucrania refleja la complejidad y la variedad de temas involucrados. Los hashtags más utilizados capturan los temas más relevantes y polémicos del conflicto, como sus protagonistas, geografía, opiniones sobre Rusia, solidaridad con Ucrania y asuntos económicos y energéticos. Además, expresan las opiniones y emociones de los participantes en estas conversaciones en línea.