Tarea 1

Jose Adrian Castillo Sierra

May 2022

1 Introducción

Reddit es una red social de discusión gestionado por la propia comunidad. Una plataforma social en la que los usuarios envían publicaciones que otros usuarios pueden votar según sus preferencias. Si una publicación recibe muchos votos, sube en la clasificación de Reddit y, aumentando su alacance al publico de esta manera; si recibe votos negativos, su alcance se reduce y desaparece de la vista de la mayoría de los usuarios.

Reddit se gestiona en grupos o subreddits. Cualquier usuario puede crear subreddits sobre cualquier tema, ya sea un asunto general, como tecnología, o específico, como una simple broma. Cada subreddit pasa a formar parte de la lista completa de envíos de Reddit, lo cual significa que una publicación en cualquier subreddit puede llegar a la página principal del sitio web.

Para la elaboración de esta tarea se hizo un análisis de texto sobre las palabras mas usadas en la ultima semana en los post mas populares del subreddit 'México'. Un subreddit donde se suele discutir, preguntar y opinar sobre temas relacionados a la vida de los mexicanos. Se eligió esta fuente de datos debido a la variedad de temas que se suelen hablar y discutir diariamente; desde temas de opinión política, dudas sobre como funcionan las leyes, preguntas laborales, post de sátira, entre otros.

2 Desarrollo

2.1 Obtención de la Información

La forma de obtener la información de la plataforma de Reddit ha sido muy fácil de obtener, ya que afortunadamente los desarrolladores de Reddit diseñaron la librería de Python PRAW, una biblioteca que nos permite extraer con muy pocas lineas de codigo los post y comentarios de cualquier subreddit. Como se puede apreciar en la figura 1

```python
with open('reddit.txt', 'w+') as f:
    subreddit = reddit.subreddit("mexico")
    for s in subreddit.hot(limit=200):
        submision = reddit.submission(s.id)
        for comment in submision.comments.list():
            try:
                f.write(comment.body)
            except:
                continue
```

Figure 1: Código de extracción utilizando PRAW

2.2 Limpieza de la información

Las funciones que se utilizaron para la realizacion de la limieza del texto recabado se enlistan en la figura 2

```python
def tokenize_data(text):
    tokenizer = nltk.tokenize.TreebankWordTokenizer()
    tokenized_text = tokenizer.tokenize(text)
    return tokenized_text

def remove_stop_words(tokenized_text):
    stop_words = set(stopwords.words('spanish'))
    stop_words.add(' ')
    return [token.lower() for token in tokenized_text if token not in stop_words]


def normalize(tokenized_text):
    stem = nltk.stem.SnowballStemmer('spanish')
    return " ".join([stem.stem(token) for token in tokenized_text]).strip()


def remove_garbage(text):
    garbage = "~`!@#$%^&*()_-+={[}]|\:;'<,>.?/"
    text = "".join([char for char in text if char not in garbage]).strip()
    return text

def remove_whitespace(text):
    text_list = text.split(' ')
    text_list = [word for word in text_list if word != '']
    text = ' '.join(text_list).strip()
    return text
```

Figure 2: Funciones de limpieza de texto

Cada una de las funciones realizando un paso para la limpieza de texto. Las

partes mas importantes de la limpieza de texto siendo la eliminación de las stopwords, palabras de conexión siendo normalmente en español las preposiciones; la lemmatizacion, la cual consiste en la reducción de una palabra a su raíz; la eliminación de los signos de puntuación y de los espacios en blanco.

El primer paso que se realizo fue eliminar las stopwords y posteriormente visualizar de las palabras mas utilizadas utilizando una nube de palabras.

Figure 3: Wordcloud de palabras mas usadas en r/mexico

Observando las palabras mas utilizadas en la figura 3, se puede apreciar el índole de los temas mas tratados, siendo las palabras mas utilizadas pertenecientes a partidos políticos o proyectos de gobierno; temas económicos y finalmente insultos.

Posteriormente a la eliminación de stopwords, se hizo una limpieza de los signos de puntuación, asi como realizar el stemming de cada palabra.

3 Conclusión

Gracias a la realización de esta tarea se pudo trabajar en la extracción de información utilizando un medio no tradicional como Kaggle para su uso posterior de análisis de información.

La información extraída del subreddit r/mexico puede ser utilizada en un futuro para realizar un análisis de sentimiento sobre la actitud de los mexicanos ante los temas cotidianos que nos rodean como sociedad.

References

[1] *R/mexico*. URL: https://www.reddit.com/r/mexico.

[2] Steelfenix. *Procesamientodatos/Tarea 1 at main · steelfenix/procesamientodatos*. URL: https://github.com/Steelfenix/ProcesamientoDatos/tree/main/Tarea%201.

[3] *The python reddit api wrapper*. URL: https://praw.readthedocs.io/en/stable/.