Tendencias de Tecnología en Twitter y periódico El Tiempo

Nefrety Sanchez
Ingeniería de Sistemas
Universidad Tecnológica de Bolívar
Email: nefretysanchez@gmail.com

Dayana Rodriguez
Ingeniería de Sistemas
Universidad Tecnológica de Bolívar
Email: lorenarodriquez@gmail.com

Yulissa Restrepo
Ingeniería de Sistemas
Universidad Tecnológica de Bolívar
Email: yulissatatiana@gmail.com

Valentina Correa
Ingeniería de Sistemas
Universidad Tecnológica de Bolívar
Email: valencorreabarco@hotmail.com

***Abstract*—In this document, we explain the analysis of trends in the social network Twitter and in the newspaper el universal, where trends in the area of technology are searched. For this, an extraction of the data was carried out, for a subsequent creation of the corpus, then a pre-processing of the texts, transformation and finally some results and / or graphs and an analysis were produced. In addition to the above, the description of the problem, the objective with this analysis, its conclusions and the address of the repository on GitHub were made.**

Keywords: News, twitter, newspaper, trends, technology, Corpus Creation, Data Preprocessing, Transformation, Unigrams, Bigramas, N-Grams

I. Descripción del problema

La información que se encuentra en páginas web junto con las redes sociales es muy diversa provocando que a partir de un comentario sobre algún tema se haga tan viral que genere una gran tendencia e incluso recomendarte temas de los cuales se habla como lo es twitter en su #tendencia en el cual te muestra de lo que recientemente se habla en los tweets de las personas alrededor del mundo y en páginas web el top de noticias más recientes y frecuentadas.

El objetivo de la predicción tanto en twitter como en la página web "El tiempo" es encontrar aquellos temas tendencias en el área de la tecnología para su posterior análisis.

II. Dirección del repositorio

El código del proyecto a tratar a continuación se encuentran en el siguiente repositorio de Github:

https://github.com/YuliRestrepo/ProyectoBigData

III. Extracción de datos (Web)

Para la extracción de datos fue utilizada la técnica conocida como Web scraping es una forma de copia al texto fuente de las páginas web, está escrito en el lenguaje de HTML (Hypertext Markup Language), estos códigos fuentes en HTML son información legibles para los humanos y para las máquinas, también llamados tags o etiquetas el cual nos sirve para el escarbado de información, para definir y extraer, los datos que son recopilados y copiados de la web, generalmente para recogerlos en una base de datos local central, para su posterior recuperación o análisis.

A. Extracción/Creación del corpus

Se llama corpus a un conjunto de documentos sin clasificar son textos de dominios pero no se encuentran determinados a que van hacer utilizados, los datos son provenientes de una fuente de datos HTML y están organizados con unos identificadores: texto, fuente,identificador correspondiente a valor único que diferencie a un registro de los demás.

Para extraer el texto usamos una técnica llamada web scraping dado que nuestro conjunto de documentos son artículos de una página de internet. Existen muchas librerías para aplicar esta herramienta, en esta caso las librerías utilizadas para el desarrollo del scraping fueron:

- Urllib.request: esta librería nos ayuda a abrir la página web para leer el contenido en html.
- Beautifulsoup: es una librería de python que permite analizar documentos HTML y extraer datos de ellos, compensando imperfecciones que puedan existir,por ejemplo, permite extraer los atributos href de las etiquetas de anclaje donde esta ayuda a evaluar los tag, attribute del link
- Selenium.webdriver: esta librería nos ayuda a mostrar el código fuente en el formato de html, la variable es un objeto que extrae las las etiquetas de anclaje, se guarda en la variable tags para que contenga todas las etiquetas que contengan hipervínculos y para cada una de estas se recorren por un bucle for y se imprime el contenido del atributo href o puede ser los ítem, etc.

Al final el top de noticias con su respectivo título, descripción, categoría, link y fecha de publicación serán guardados en un documento excel.

B. Preprocesamiento de los textos

Pre-procesar el texto es limpiar nuestro texto de ruido y estandarizar nuestro texto para que sea fácil de comparar. Algunos de estos pasos incluyen quitar las puntuaciones, transformar el texto a minúscula, o eliminar palabras frecuentes usadas en el lenguaje como preposiciones, artículos, etc. que por sí mismas no aportan mucha información, además de usar la raíz de las palabras.

C. Transformación

Para el análisis, se le agregó la librería pandas encargada de enviar a un dataframe para que se almacene y se pueda hacer un mejor uso de la información obtenida.

- import pandas: Pandas es una herramienta de manipulación de datos
- import numpy: Trae integradas muchas funciones de cálculo matricial de N dimensiones, así como la transformada de Fourier, múltiples funciones de álgebra lineal y varias funciones de aleatoriedad.
- import matplotlib.pyplot as plt: es una librería para generar gráficas a partir de datos contenidos en listas, vectores, en el lenguaje de programación Python y en su extensión matemática NumPy.

Para el análisis de noticias nuestra fuente primaria fue el tiempo, se definen las variables que se quieren extraer que son las subpáginas de la página del tiempo que es un periodico de libre uso, toda la información que se depositada es accesible en cualquier momento, la técnica utilizada fue RSS (Really Simple Syndication) es un emisor de información que se utiliza y fue definido para masificar información de manera estándar la cual se emite por medio de los formatos rss o xml, este formato nos permite ver el top de noticias de forma plana, la forma de recolectar toda la información es conocida como corpus, este se define como la colección de documentos de un dominio particular, en este caso nuestro corpus es específicamente de noticias, porque solo estamos extrayendo fuentes de un periodico en la subpágina de tecnologia, para esto se definió una clase rss donde nos recibe la url y nos retorna un conjunto de artículos que se convierte en nuestro corpus, importamos las librerías y creamos las funciones para tener una comunicación vía web, se buscan los artículos que son las fuentes y los enlaces de donde se va a traer la información, que busque titulo, link, descripción y la fecha de publicación, dentro de esa búsqueda luego se busca los otros ítems que son como los hijos de esa noticia del articulo principal eso se almacena en un diccionario específico para la extracción, se almacena en una lista, para esto también es necesario definir los links de las páginas que queremos extraer en el formato xml, para obtener las páginas hija de esta.

D. Resultados y análisis

Para la extracción de noticias de la página web “El Tiempo”, se tomó el top 10 de los artículos presentes en la gráfica con sus respectivas categorías, ver **Gráfica 1**.

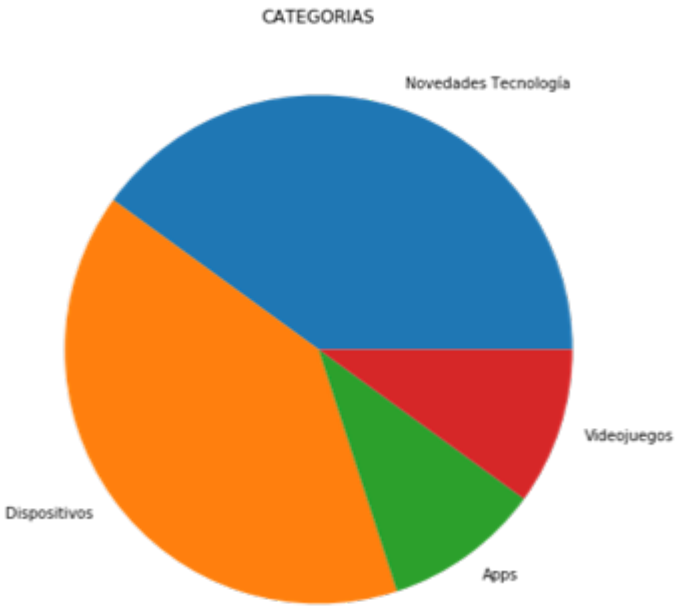

Fig. 1. Gráfica de frecuencias de las categorías. Fuente: Propia

En la **Gráfica 2**, no se observa una correlación entre las variables.

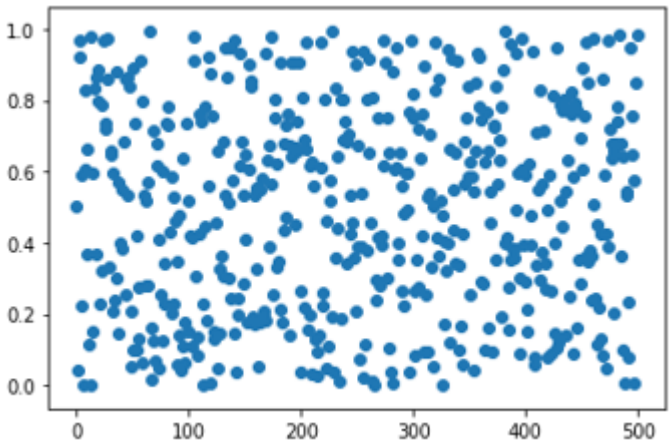

Fig. 2. Gráfica de dispersión de las categorías. Fuente: Propia

IV. Extracción de datos(Twitter)

A. Extracción/Creación del corpus

Dentro de la extracción de los tweets en el #tecnologia se usa en tweepy, es una libreria de la API de twitter para python. Para esto se necesita tener unas keys y tokens correspondientes, autenticando estas con la cuenta de desarrollador para tener el acceso a los tweets con OAuthHandler, usando tweepy.Cursor(), la cual lleva unos parámetros específicos

para hacer su uso como api.search, el buscador específico del hashtag, el idioma y la ubicación para que no agarre tweets de diferentes lugares del mundo. Los tweets extraídos se guardan en un csv. Además de librerías como:

- import sys: este es un módulo que se encarga de proveer variables y funcionalidades, directamente relacionadas con el intérprete.
- import pandas: es una librería destinada al análisis de datos, que proporciona unas estructuras de datos flexibles y permiten trabajar con ellos de forma eficiente.
- import unicodedata: Este módulo proporciona acceso a la base de datos de caracteres Unicode (UCD), que define las propiedades de todos los caracteres Unicode. Los datos contenidos en esta base de datos se compilan a partir del UCD versión 13.0.
- import re: Este módulo proporciona operaciones de coincidencia de expresiones regulares similares a las encontradas en Perl.Tanto los patrones como las cadenas de texto a buscar pueden ser cadenas de Unicode (str) así como cadenas de 8 bits (bytes).
- import preprocessor: es una biblioteca de preprocesamiento para datos de tweets escritos en Python.
- import nltk: módulo de Python que contiene muchas funciones diseñadas para su uso en el análisis lingüístico de documentos y en el procesamiento de lenguaje natural

B. Preprocesamiento de los textos

Los diferentes tweets llegan con muchos aspectos que obstaculizan el análisis limpio del texto, por ello hacemos uso de algunas funciones que nos ayudan a limpiar todo el texto como las etiquetas, condiciones con "utf-8" pero aun asi las etiquetas también se eliminan para que esté más limpio el texto final.

C. Transformación

En el análisis se usó las librerías:

- from collections import Counter: contador es un contenedor que almacena elementos como claves de diccionario y sus recuentos se almacenan como valores de diccionario.
- import matplotlib.pyplot as plt: es una librería para generar gráficas a partir de datos contenidos en listas, vectores, en el lenguaje de programación Python y en su extensión matemática NumPy.

Para el análisis, el texto que fue mejorado en el preprocesamiento se le agregó la librería pandas encargada enviar a un dataframe para que se almacene y se pueda hacer un mejor uso de la información de los tweets suministrados por tweepy. Para poder encontrar y agrupar las palabras frecuentes necesitadas, se tuvo que usar unigrams, bigrams y trigrams.

A este modelo de agrupación de palabras anteriormente dichas se les llama unigrams, bigrams y trigrams dependiendo de cuantas palabras se agrupe, por como su nombre lo indica una solo abarca una palabra o letra, la siguiente dos palabras o letras y la tercera usa tres palabras o letras. Dentro del proyecto se hizo uso de estas para la organización de las palabras del texto usando la librería nltk y algunas funciones como el counter() para el conteo de las más frecuentes mostrando el número de veces que se usó dicha palabra, generando un valor que se usa para hacer la estimación de cada gráfica de los unigrams, bigrams y trigrams.

D. Resultados y análisis

Para los resultados de twitter, se obtuvo el top 10 de los unigrams, bigrams y trigrams, encontrados en el dataset.

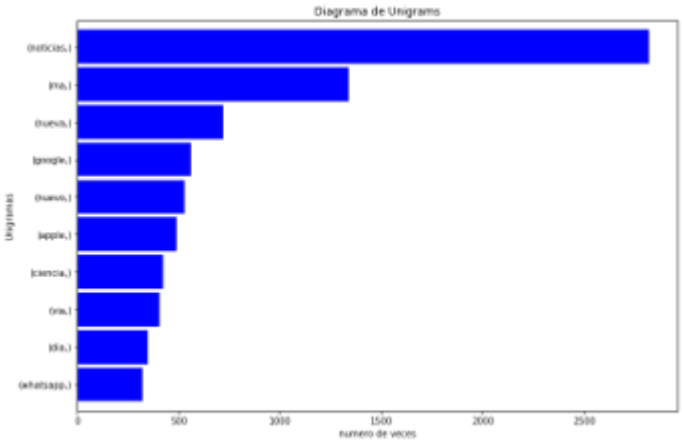

Fig. 3. Diagrama de unigramas. Fuente: Propia

Dentro del unigrams (Ver **Gráfica 3**) podemos notar, hay una gran tendencia hablar de noticias, por tanto se observa un número de veces que esta palabra es mencionada, aun así hay otros temas tendencias como apple, whatsapp, google que aunque no tienen gran incidencia también están en el top 10 son temas que tienen relevancia.

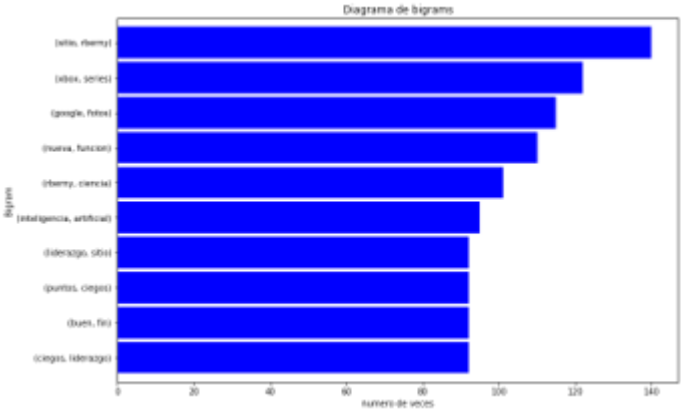

Fig. 4. Diagrama de bigramas. Fuente: Propia

Dentro de la **Gráfica 4** se puede observar, muchos temas tienen tendencia e incluso tiene la misma frecuencia de veces que se menciona dentro de la captura de tweets pero podemos ver que los más hablados son el sitio web rberny, xbox, google,

inteligencia artificial que aun asi que varían en su frecuencia están dentro de top.

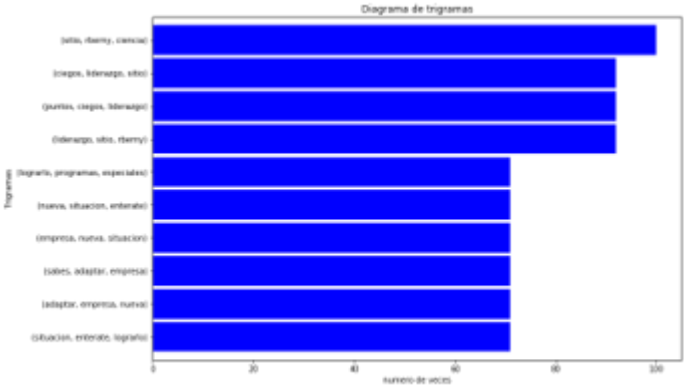

Fig. 5. Diagrama de trigramas. Fuente: Propia

La **Gráfica 5** tiene algunos aspectos, muestra la frecuencia de veces que se habla del sitio web rberny que también tiene una gran tendencia, la cual nota en las gráficas anteriores, se puede ver que hay otras 3 y 6 barras que tienen la misma frecuencia.

V. Conclusiones

Se puede observar que los procesos para la extracción de la información es muy diferente debido a que usan librerías específicas, usa funciones de limpieza para que el texto final este bien y su análisis sea el más eficiente. Dentro de twitter cabe destacar que hay muchas palabras que están dentro de los tweets que por más filtros se le aplique para eliminar las stopwords y las etiquetas hay cosas que agarra de manera repetida como verbos, adverbios entre otros, siendo mostrados con frecuencia en las gráficas.

References

[1] Clase 8: PLN con Python. Instituto de Ingeniería UNAM. Recuperado de: http://www.corpus.unam.mx/cursopln/plnPython/clase8.pdf

[2] Moya, R. (2015). Pandas en Python, con ejemplos -Parte I- Introducción. Jarroba. Recuperado de: https://jarroba.com/pandas-python-ejemplos-parte-i-introduccion/

[3] Python. Documentación: re — Operaciones con expresiones regulares. Recuperado de: https://docs.python.org/es/3/library/re.html

[4] Python. Documentación: unicodedata — Base de datos Unicode. Recuperado de: https://docs.python.org/es/3.9/library/unicodedata.html

[5] The Python Package Index. Project description: tweet-preprocessor. Recuperado de: https://pypi.org/project/tweet-preprocessor/

[6] Unipython. Matplotlib: Funciones principales. Recuperado de: https://unipython.com/matplotlib-funciones-principales/

[7] Uniwebsidad. (2013). Python: 10.1. Módulos de sistema. Recuperado de: https://uniwebsidad.com/libros/python/capitulo-10/modulos-de-sistema