TOKENIZACIÓN Y ANÁLISIS MORFOLÓGICO DE MENSAJES DE TEXTO DE TWITTER EN ESPANOL

MARÍA CRISTINA PORTILLA CORTÉS
maria.portilla.cortescorreounivalle.edu.co

JOSE HERIBERTO TORRES CORTÉS
jose.heriberto.torres@correounivalle.edu.co

Escuela de Ingeniería de Sistemas y Computación
Facultad de Ingeniería
Universidad del Valle
Cali, Colombia

Octubre 2016

PROYECTO

ÍNDICE GENERAL

1 INTRODUCCIÓN

1.1 DESCRIPCIÓN DEL PROBLEMA

La normalización de texto es una de las principales tareas en el procesamiento de lenguaje natural (PLN) con varias fases entre ellas la segmentación de sentencias y palabras (tokenización) y el análisis morfológico de las palabras. Lenguajes como el Español guardan una rica información de los aspectos de forma de la palabra como el género, el número, la conjugación, etc. Por otra parte el PLN está incursionando con mucha fuerza en la era de Big Data y las redes sociales, disciplinas como el análisis de sentimientos y la mineria de opinión están siendo utilizadas en tiempo real y con resultados soprendentes en la toma de decisiones. Una de las tareas importantes en la normalización de texto tiene que ver con la tokenización y en la actualidad la seg- mentación en tweets que son más que un textese son una mezcla de jergas, abreviauras, nicknames, urls, hashtags y emoticones. Las reglas léxicas, morfológicas y sintácticas de este tipo de mensajes no obeden a la lingüistica tradicional y por lo tanto la normal- ización debe ser especializada. Por ejemplo consideremos el siguiente twitter:

SantosNobeldePaz Malala, premio Nobel 2014, alaba la "valentıa" de@JuanManSantos en las negociacione s de paz http://ow.ly/qdLo304XXtd [u'SantosNobelPaz', u'Malala', u',', u'premio', u'Nobel', u'2014', u',', u'alaba', u'la', u"'", u'valent', u", u"'", u'de', u'@JuanManSantos',u'en', u'las', u'negociaciones', u'de', u'paz', u'http://ow.ly/qdLo304XXtd']

En este caso las palabras Malala, premio, Nobel, alaba, valentıa, de, en, las, ne- gociaciones, de, paz tienen información morfológica basada en caracterısticas léxicas, mientras que SantosNobeldePaz, @JuanNobelPaz, 2014, http://ow.ly/qdLo304XXtd se etiquetan con una categoria especial sin ninguna caracterización léxica. Muchas de las palabras tokenizadas tienen información morfológica y otras hacen parte del lenguaje propio de twitter. En Español son pocas las herramientas que procesan este tipo de información, sin embargo, encontramos entre ellas a Freeling una plataforma que per- mite realizar muchas de las tareas de PLN, en especial el análisis morfológico de las palabras que hacen parte del Español. En tal sentido, el analizador morfológico de Freeling utiliza un conjunto de etiquetas para representar la información morfológica de las palabras. Este conjunto de etiquetas se basa en las etiquetas propuestas por el grupo EAGLES (Expert Advisory Group on Language Engineering Standards) para la anotación morfosintáctica. Freeling es entrenado sobre el corpus AnCora a través del alizador MORPHO de dos niveles como PC-Kimmo en el cual el análisis consiste en extraer la información de forma de la palabra dada una palabra en su estado fuente.

Tomado del enunciado del proyecto.

2 INSTALACIÓN

La aplicación fue desarrollada sobre Docker, por lo tanto para ponerla en ejecución es necesario instalar esta herramienta. La información completa para s instalación se encuentra en:

https://docs.docker.com/engine/installation/

Una vez instalada, debe iniciar el servicio mediante el comando:

systemctl start docker

Es posible que le pida iniciarse sesión como root o administrador.

Para crear la imagen del contenedor, ubicarse en la carpeta donde está el archivo llamado *Dockerfile* y ejecutar el siguiente comando:

docker build -t p9 .

Para ejecutar el contenedor con la anterior imagen digite lo siguente:

docker run –rm -it -p 5000:5000 -v $(pwd):/root/app p9 python3 /root/app/hello_flask.py

Para ver la interfaz de la aplicación use la siguiente direccion:

http://127.0.0.1:5000/analize

3 DESARROLLO DEL PROYECTO

La idea general para desarrollar el proyecto fue recibir un texto y tokenizarlo.Guardar esos tokens en una lista. Mediante un ciclo *for* recorrer la lista e ir clasificando las palabras, para finalmente costruir el código *html* que será mostrado en la interfaz.

3.1 TOKENIZACIÓN

Ha sido realizada con la herramienta NLTK importando el módulo TweetTokenizer.

3.2 FREELING

Se usa la libreria *Pyfreeling*. La cual realiza un llamado a la terminal de freeling.

`https://github.com/malev/pyfreeling`

3.3 CLASIFICACIÓN DE TOKENS

Se coloca en un ciclo *for* que recorre todos los tokens entregados por el módulo TweetTokenizer. Se clasifican en el siguente ordén:

- Emoticones
- Flechas
- Hashtags
- Nicknames
- Urls
- Palabras

Con el objetivo que si no está en las primeras categorias, es muy probable que sea una palabra.

3.4 IMPLEMENTACIÓN

El proyecto está implementado en el lenguaje de programación Python versión 3.4 usando el microframework Flask para la interfaz.