IR Project

Part 1 Report

Pol Ayala - (u173483)
Toni Carbonell - (u172950)
Nil Distefano - (u172948)

Context

The aim of this report is to explain the functions and decisions we made on the first part of our IR project.

Functions

The purpose of this delivery was to pre-process the tweets. The core functions that we created are the following:

1. clean: Preprocess any text by lower-casing all characters, tokenizing, removing non-alphanumerics and stop-words and stemming.

2. feature extraction: given a JSON, it creates and returns dictionary with the core features of the tweet. This are: id, username, likes, mentions, retweet… It uses clean() function to clean the tweet text.

3. build tweets df: given all JSON tweets, it goes through them creating a panda where each line is a tweet and each column is a tweet feature. It uses feature extraction for every JSON file and append all the returned dictionaries to a combined tweet dict, which we will create the pandas from.

Further Information

To deal with URLs, we had to manually create them. This can be seen in line: “tweet[’Url’]=’https://twitter.com/’+tweet[’Username’]+’/status/’+tweet line[’id str’]”

In addition to that, some tweet texts finished with another URL which we also had to deal with. This can be seen in line: “tweet[’Clean text’] = clean(re.sub(r’http§+’, ”, tweet line[’full text’]))”

Besides what was demanded on the lab, we decided to also include the “mentions” of the tweets. This can be seen in line “tweet[’Mentions’] = mentions if len(mentions) != 0 else NaN”

On the .ipynb file we also run “feature extractions” and display instances of the pandas, so that it can be seen that they work correctly.

Finally, to merge the existing pandas to the “map path”, we performed a merge using “pd.merge” function on the attribute ‘Id’. Input and output folders can be seen on the Readme file.

Conclusion

All in all, we performed well on the first part of our project. We build the necessary functions to construct a pandas that had each tweet main features, processed and cleaned.