Social Media Text Processing and Semantic Analysis for Smart Cities

João Filipe Figueiredo Pereira

Master's thesis project supervised by *Prof. Rosaldo Rossetti* and *Pedro Saleiro* at *LIACC*

1. Motivation

With the rise of Social Media, people obtain and share information almost instantly on a 24/7 basis. Many research areas have tried to gain valuable insights into these large volumes of freely available user generated content. The research areas of intelligent transportation systems and smart cities are no exception. However, extracting meaningful and actionable knowledge from user generated content is a complex endeavor. First, each social media service has its own data collection specificities and constraints, second the volume of messages/posts produced can be overwhelming for automatic processing and mining, and last but not the least, social media texts are usually short, informal, with a lot of abbreviations, jargon, slang and idioms.

2. Problem Description

3. Goals

In this dissertation, we try to tackle some of the aforementioned challenges with the goal of extracting knowledge from social media streams that might be useful in the context of intelligent transportation systems and smart cities. We designed and developed a framework for collection, processing and mining of geo-located Tweets. More specifically, it provides functionalities for parallel collection of geo-located tweets from multiple pre-defined bounding boxes (cities or regions), including filtering of non complying tweets, text pre-processing for Portuguese and English language, topic modeling, and transportation-specific text classifiers, as well as, aggregation and data visualization. We address nine different goals in this dissertation to achieve our aim:

1. Continuous collection of geo-located tweets from multiple bounding boxes in parallel and in compliance with Twitter API usage limits
2. Tackling Twitter Geo API inconsistencies and filtering noisy tweets
3. Implement standard text pre-processing methods for social media texts
4. Content analysis using topic modeling and comparative characterization among different bounding boxes (e.g. cities)
5. Travel-related classification of tweets using supervised learning
6. Train word embeddings from geo-located tweets
7. Study the impact of word embeddings in travel-related classification
8. Creation of gold-standard data for travel-related supervised learning
9. Aggregation and visualization of results

4. Proposed Solution

In this section it is described the problem to be tackled in this dissertation as well as the designed and implemented framework that we proposed to solve it and the core modules composing it.

4.1. Architecture Overview

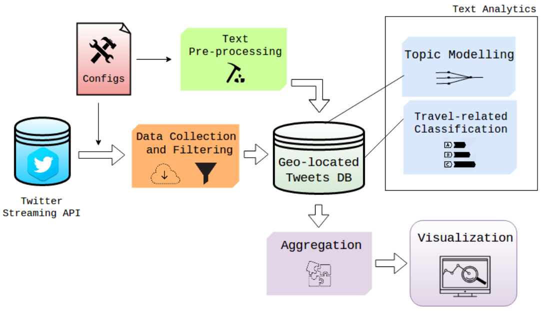

Fig. 1 – Framework Architecture Overview

4.2. Data Collection and Filtering

The data collection module was built using Tweepy, an open-source Python library to access the Twitter APIs. We explore the Twitter Streaming API using the locations heuristic that allows the retrieving of geo-located tweets through a bounding-box matching. The filtering task is due to the large amount of tweets in different languages comparative to the one spoken in the target city and to tweets retrieved by the API that are actually outside of the searching bounding-box.

4.3. Text Preprocessing

We apply a considerable group of text preprocessing operations to the messages such as lowercasing, lemmatization, tokenization, transformation of repeated characters, punctuation removal, cleaning of *metadata*, numerical symbols in the text as well as removal of stop and short words.

4.4. Topic Modelling

4.5. Travel-related Classification

5. Conclusion

Referências