Social data gathering

Álvaro Domínguez Calvo
adomingue599@alumno.uned.es

Abstract

1 Introduction

2 Data retrieval process: gather and formatting social media data

In this work it is proposed three datasets with social data from three social networks that are Twitter, Reddit and post blogs. Each dataset is build by using three different APIs: the Twitter API for the Twitter dataset, the Twingly API for the blog posts dataset and the Social Searcher API for the Reddit dataset. For each dataset, it is considered a set of keywords that will retrieve social information by exact matching in the corpus of each document. Since the retrieval must be performed for different languages, the keywords corresponds to named entities or specific lingo used in the social media in question. As it will be seen, some tweets were retrieved by using the special character *$* which denotes information related to the stock market.

The next sections are intended to explain the gathering process, how the data is formatted in order to persist it according to legal concerns of the social media source information and some limitations related to the use of these type of APIs and the information they provide.

2.1 Gathering process

Typically, in the gathering process of any type of social data is implied the time variable which, for example, can be used to cluster documents by fixed window times. The building process of the different datasets provided comprise periods of time in terms of its retrieval. The table 1 shows the dates when the retrieval process starts and ends per dataset.

Dataset	Start date	End date
Twitter	2021-12-04	2021-12-31
Blog posts	2021-11-08	2021-12-06
Reddit	2021-11-08	2021-12-06

Table 1: Start and end dates of the retrieval data

Each dataset is built according to a fixed set of keywords that corresponds to named entities. Each keyword belongs to the domain of music, stock market and news related to a natural disaster. The table **??** summarizes this information.
TODO: Table summarizing the keywords used per dataset

4 Experiments

5 Conclusions