Assignment 1: Studying 2016 US Elections Through Analysing Twitter Data

António Mendes
17amendes@gmail.com
11925051

Judit Győrfi
judit.gyorfi@student.uva.nl
13209647

Orlando Scarpa
orlando.scarpa@student.uva.nl
13266918

Horváth Ádam
email address
UVA-net ID

Miklos Kosarszky
miklos.kosarszky@student.uva.nl
13242857

Contents

1 Data Preparation

After extracting all the json objects and flattening all the fields the DataFrame had 35 columns, Of these, 10 were selected as valuable insights into the data.

Using the columns relating to the language and country code, the DataFrame was reduced to only tweets in english and posted from the United States. After this, the column containing the full name of the location of origin of the tweet was used to extract the two letter code of the state of origin, keeping in mind to exclude territories whose inhabitants don't vote for the president like Puerto Rico and Guam. Of the 517,724 tweets in english from the United States, less than 5000 tweets were excluded in this manner. To clean the text and prepare it to be tokenized, all text in the tweets was stripped of punctuation, special characters, new lines, hyperlinks and trailing spaces. Furthermore, stop words were removed and select hashtags and mentions were counted to give a rudimentary estimate of the candidate being talked about in each tweet.

2 Data Splitting

3 Sentiment Analysis

4 Topic Modelling

5 Results Understanding

6 Limitations

7 Context