**Project Documentation**

**Data Science Brainster Academy Final Project**

Team 3: Tatjana Veljkovic, Ilija Todorov, Ivana Tomovska Efremov

Project Description: the project task is to analyze health related tweets released by 16 different media outlets[[1]](#footnote-1) in the period of 2011 to 2015. Final task is to cluster the health tweets to determine various trends, such as dominating health news topic by media outlet, dominating health news trends through time periods and other related trends.

The approach used was text processing by using Natural Language Processing (NLP) methods. The dataset was processed using Python Natural Language Toolkit libraries (NLTK). Various clustering methods were used to group/cluster text patterns in order to determine the aforementioned trends.

The dataset was prepared and processed using the following steps:

* **Step 1:** Loading and joining the data - the initial data was contained in 16 text files (.txt) that needed to be cleaned, merged and prepared for further processing.
* **Step 2:** Data cleaning – cleaning and data and extraction of tweets as separate text column.
* **Step 3:** Text processing using NLTK (tokenization, POS tagging, normalization, stemming and lemmatization)
* **Step 4:** Applying Vector-space models
  + Vector-space models used: Bag of Words (BoW) and
  + Term Frequency Inverse Document Frequency (TF IDF)
  + Computing Doc2Vec similarity
* **Step 5:** Clustering
  + Applying DBSCAN algorithm
  + Applying K-means algorithm
* **Step 6**: Data Analysis
  + Exploring data by year and by source (media outlet)
  + Computing Doc2Vec similarity between two or more sources (media outlets)
  + Computing cluster similarities between two of more sources
* Step 7: Applying WordCloud visualization to confirm the findings

Findings: According to the clients' needs this analysis can be done to compare two or more sources (media outlets) and analysis can be done by year and month in the period of 2011-2015. The analysis can compare similarity between the content of tweets released by a certain media outlet and/or the month and year when it was done. Further analysis can identify top trends and top terms that were tweeted by media outlet throughout the months in 2011-2015.

Case studies were presented for 2014 and 2015 and by comparing BBC and CBC news and one more comparison between Kaiser Health News and LA Times Health.

The word clouds made for each media outlet for 2014 findinds are consistent with the type of media outlet (national service, NGO or entertainment) and the territory it covers (global vs national). 'Ebola' is still the dominating term in 2014, again because of the big Ebola outbreak in West Afrika in 2014-2016.

For more insights into the libraries, text processing and clustering algorithms that we used in this project, please check out our Jupyther Notebooks. Thank you for reading!

1. Health related tweets from the period of 2011 to 2015 were collected from the following media outlets: BBC Health, LA Times Health, CBC Health, MSN Health News, CNN Health, NBC Health, Everyday Health, NPR Health, Fox News Health, NY Times Health, GDN Healthcare, Reuters Health, Goodhealth, US News Health, Kaiser Health News, WSJ Health. [↑](#footnote-ref-1)