**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 compare various media outlets and determine various trends, such as dominating health news topic by media outlet, dominating health news trends through time period 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.

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.

**Result of Step 1 and 2:** separate csv file for each year named PartOne\_’year’ to be used as input file for further processing.

* **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

**Result of Step 3, 4 and 5:** separate csv file for each year named PartTwo\_’year’ to be used as input file for further processing.

* **Step 6**: Data Analysis – Tweet Analysis on data for year 2014
  + Exploring data for year 2014 being the year with highest number of tweets, highest number of sources (media outlets), and highest number of sources with monthly distribution of tweets to be compared.
  + Selecting two pairs of sources to be compared, based on monthly tweet distribution and source origin.
  + Computing similarity of tweets between the selected two pairs of sources (media outlets) using: Jaccard similarity, Cosine similarity based on doc2vec, and similarity of clusters
* **Step 7**: Applying Word Cloud visualization to confirm the findings
  + Word Cloud visualizations are confirming the findings for the year 2014
  + Due to the lack of data to be compared for the years 2011, 2012, 2013 and 2015 Word Cloud is the only tool used for presentation purposes.

**Result of Step 6 and 7:**  Core Tweets Analysis notebook referring to the year 2014, and separate Word Cloud notebook for each of the years 2011, 2012, 2013 and 2015.

**Findings:**

Case studies were presented for 2014 as Core Tweets Analysis 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 findings are consistent with the type of media outlet (national service, NGO or entertainment) and the territory it covers (global vs national). 'Ebola' is 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)