**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:** Exploratory Data Analysis (EDA) – making some preliminary analysis and conclusion regarding the data (frequency of tweets per day of the week, per media outlet (source), etc.)
* **Step 4:** Text processing using NLTK (tokenization, POS tagging, normalization, stemming and lemmatization)
* **Step 5:** Applying Vector-space models and clustering
  + Vector-space models used: Bag of Words (BoW) and
  + Term Frequency Inverse Document Frequency (TF IDF)

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)