Lab IV - Clustering

module IV, Unsupervised Learning

DT4031, *Applied Machine Learning*
Jens Lundström

Introduction & Assignment

In module IV you have been introduced to clustering which is a useful tool for discovering natural groups in data. In this lab it is hypothesized that you are working for a medical company as a data scientist and one of your responsibilities is to keep track of health trends. One way to do this is to investigate tweets from other companies and try to discover trends, this is also the task of this lab. The general assignment is to load 3199 tweets from the MSN Health channel, clean and transform the data into useful features and then apply k-means to find natural groups of news (possible health trends). All necessary code is supplied in the lab and therefore this lab is more of a data and algorithm investigation where you will be answering a series of questions given the theory from the literature, slides, and lectures.

Lab report instructions

First task, download the health tweets (see link in Notebook) and run the whole Notebook. Go through each cell and result (and try to understand both, start by reading the code comments). You may have to install necessary Python modules before running the notebook.

Second task, given the three measures of cluster quality (Calinski-Harabasz, Silhouette and SSE) and the 3D-plot of the data answered the question: *does the initial guess of k=2 make sense? Motivate your answer!*

Third task, set k=8 and re-run the Notebook, and study especially the output of cell 21 (mean feature vectors of all the clusters). It seems that most of words are not related to health, or are they? *Answer this question and motivate your answer.*

Task four, remove (clean) some non-health related words from the data by copy and paste the below lines into the end of cell 6. Re-run the notebook and study the results now with k=8. Answer: *Did you get a better result? Motivate your answer.*

```
df_tweets['tweet_text'] = df_tweets['tweet_text'].str.replace('study', '')

df_tweets['tweet_text'] = df_tweets['tweet_text'].str.replace('report', '')

df_tweets['tweet_text'] = df_tweets['tweet_text'].str.replace('says', '')

df_tweets['tweet_text'] = df_tweets['tweet_text'].str.replace(':', '')

df_tweets['tweet_text'] = df_tweets['tweet_text'].str.replace('suggests', '')

df_tweets['tweet_text'] = df_tweets['tweet_text'].str.replace('confirms', '')

df_tweets['tweet_text'] = df_tweets['tweet_text'].str.replace('finds', '')

df_tweets['tweet_text'] = df_tweets['tweet_text'].str.replace('experts', '')

df_tweets['tweet_text'] = df_tweets['tweet_text'].str.replace('raise', '')

df_tweets['tweet_text'] = df_tweets['tweet_text'].str.replace('risk', '')

df_tweets['tweet_text'] = df_tweets['tweet_text'].str.replace('linked', '')

df_tweets['tweet_text'] = df_tweets['tweet_text'].str.replace('new', '')

df_tweets['tweet_text'] = df_tweets['tweet_text'].str.replace('help', '')
```

Task five, given the three measures of cluster quality (Calinski-Harabasz, Silhouette and SSE) with the added cleaning code and the 3D-plot of the data: *choose a better suited k, motivate your answer and re-run the Notebook.*

Task six: *now, study the output of the function* print_b_number_of_sentences_from_the_computed_clusters *and try to manually find similarity among the sample tweets* *within* *each cluster. Are you able to find similarities? What would you say that the main health trend topics from MSN Health are? Please motivate your answers.*

The lab solution shall be documented in the form of a written report. As a data scientist it is your task to explain what you have done in a detail such that the solution is reproducible, the reasoning behind the decisions, the outcome (results) and the conclusions you have drawn from such results. It is suggested to have the following lab report structure: Introduction & Motivation; Methods & Solutions; Results; Conclusions & Discussion; Appendix. Keep the number of pages in the lab report limited, no more than five pages (excluding *Appendix* which should contain Notebook code and figures not directly linked to the result).

Lab assistance

Feel free to reach out to Guojun or Zeinab if you have questions regarding the lab.