Weekly assignment 6

This weekly assignment is about Google Trends, the case study and text mining. Motivate your answers, always. The solution to this weekly assignment needs to be submitted as pdf (!) via Canvas. Expected length is about four pages (all included, R code can go to the appendix).

Question 1: Google search volumes

Collect Google search volumes for search keys related to your team assignment. Select at least three European countries of your liking, and use a relevant time window. Make a time plot of the collected search volumes, and discuss the main insights obtained from this figure.[1] [0.5 page]

Question 2: Reflections on the case study

During the lecture, we discussed a case study. For this question, you have to (i) report your initial approach to the case study and your main insights and findings, optionally supported with outcomes and graphs of analyses in R, (ii) reflect on the class discussion, and note where you agreed or disagreed, and (iii) propose the next steps if you were in charge of this project at the bank.

If you were not at the lecture, write a critical review of the referenced paper: S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. *Decision Support Systems*, `http://dx.doi.org/10.1016/j.dss.2014.03.001` [1 page]

Question 3: Stack Exchange topics and questions

Over the years, Facebook has hosted multiple competitions on Kaggle to recruit new employees. This question is based on the third challenge.[2]. The original competition tested text mining skills on a large dataset from the Stack Exchange sites. The task was to predict the tags (a.k.a. keywords, topics, summaries), given only the question text and its title. The dataset contains content from disparate stack exchange sites, containing a mix of both technical and non-technical questions. A sample of this dataset has been made available on Canvas, as *stack.csv*, and this sample provides three columns: the *Topic* of the post (one of four topics), the *Title* of the post, and the *Body* of the post which contains

[1] Search tips can be found at `https://support.google.com/trends/answer/4359582?hl=en`.

[2] See `https://www.kaggle.com/competitions/facebook-recruiting-iii-keyword-extraction/overview`

the actual question and explanation. This dataset will be used for the rest of the assignment. (*i*) Report the distribution of posts over the various topics, (*ii*) inspect a couple of question text bodies and report on your observations, and (*iii*) discuss issues that need to be solved when cleaning the text later on. [0.5-1 page]

Question 4: Preparation of the text

Choose two of the *Topic*s for use in the rest of the assignment. Prepare (cleanse) the question body texts for further analysis by means of the following steps. (*i*) Use functions from package *textclean* to remove URLs and make other desired adjustments.[3] (*ii*) Use package *tm* to build a corpus from the vector with cleansed text for each of your chosen *Topic*; implement further cleansing where necessary; do motivate your choices. (*iii*) Give a summary of the number of words and documents in the resulting corpus for each of your chosen *Topic*. [0.5 page]

Question 5: Popular associates

For each of your chosen *Topic*, make an overview of the most frequent terms in the corpus and make an overview of the terms that have the higher correlations with that *Topic*. Comment on the results. Speculate as to whether there is possible predictive value in these results. [0.5 page]

Question 6: Word clouds

Make a word cloud for each of your chosen *Topic*, and include these in your report. Compare the two graphs and discuss the result. [0.5 page]

[3] Various *textclean* functions have been demonstrated in class and in the tutorial. Further detail can be found in the package's reference manual, see `https://cran.r-project.org/web/packages/textclean/textclean.pdf`.