Business Data Analytics

Homework 6

The result of your homework should be .pdf / .r / .rmd file that should contain questions and answers. Use course webpage to submit the report and don’t forget to check it before submitting to ensure that vital information is readable.

The homework should contain:

- The report, which has answers to the questions along with the plots and code snippets.
- In order to submit these two files, they need to be zipped first.

You are always welcome to ask us any questions in Piazza! With the gained knowledge during the lecture and lab session you should be able to solve the following tasks:

Part 1. (Brand Value Monitoring)

Input: “apple1.csv” and “apple2.csv”

In apple1.csv and apple2.csv you have comments about apple before (apple1.csv) and after (apple2.csv) announcing the quarterly profits.

1.1. (1 point) Clean both datasets:
   - change all characters to lowercase;
   - remove URLs;
   - remove words starting with ‘@’;
   - remove words starting with ‘$’ (use ‘\$’ in regular expression);
   - remove punctuation;
   - remove stopwords;
   - remove numbers;
   - remove whitespaces;

   Don’t forget to inspect the text after each cleaning to see if the outputs are desirable.

   Tip: Think about the best order in which you would like to do the cleaning for the above cleanings.

1.2. (1 point) Create word clouds (for each dataset) with words that occur at least 8 times for each dataset. Which are the top 2 frequent words in both word clouds.

1.3. (2 points) Build bar plot with the total scores of the all sentiments and emotions from both datasets. Which tweets (before or after) are more positive?

1.4. (2 points) Combine datasets (apple1 and apple2). Predict sentiments of the tweets using Random Forest. Build confusion matrix and calculate accuracy, recall and precision of the model.
Part 2. (Networks)

Input: “Hi-tech-Nodes.csv” and “Hi-tech-Edges.csv”

“Hi-tech-Nodes.csv” contains information on users who are working in Hi-tech: name, gender and department to which they belong. Edges in “Hi-tech-Edges.csv” represent who converses with whom during the weekdays, and weight represents the frequency of conversation. Given this data build a directed graph and do following tasks:

2.1. (2 points) Calculate statistical information:
   a. Density of the network;
   b. Clustering coefficient;
   c. Reciprocity of the network;
   d. Average path length;
   e. Diameter (by considering weights). Please do not report the non-weighted diameter. Interpret these metrics' output (that is what you Infer about the network)?

2.2. (2 points) Plot the graph with following rules:
   - Each node has a name associated with it;
   - Each node has different shape, depending on gender of the user;
   - Each node has different color, depending on department to which user belongs;
   - Each node has different size depending on the hub size of the nodes;
   - Each edge has different size depending on the weight;
   - Find the path of the diameter on the graph and colorize its edges only.

2.3. (2 point) Find the total communities in the network using Walktrap algorithm. How much communities are there? Check some samples from different communities. Does it make sense to put these samples in different groups? (Discuss which samples you took and your reasoning about why these samples fall in different groups).