Assignment 01

Task 01: Association rules analysis

You have to perform market basket analysis using the data from 'grocery assignment.csv'.

The columns of interest in this case are 'transaction id' and 'itemDescription', and the data is in the format we have encountered before: we have several rows for each transaction, one row for each item that is part of the transaction. Transform the data into transactional format, extract frequent item-sets and create association rules as it was done in the course material.

(Note: you may have to use much smaller threshold values for support and confidence than in the lectures; experiment with different values until you get a reasonable amount of item-sets and rules).

Do the extracted rules seem to contain useful information or only trivial observations? Specifically, what would you recommend to a person (i.e. what is the consequent in association rules terminology) when you now that in the basket there is already (i) sausage, (ii) tropical fruit, (iii) sausage AND tropical fruit?

Task 02: Centrality measures

In this exercise, you will have to work with the data related to the Harry Potter series. You will need to use two files, relations.csv and characters.csv.

The relations file contains the link between characters, using an ID value. The name of the character can be then found by searching for the ID number in the characters file (this contains some additional data that will be used in the next exercise). Your task is to identify the most central characters based on the provided network structure.

Calculate the discussed centrality measures for the network (degree centrality, betweenness, closeness, PageRank). Compare the results, first by comparing the 6 most central characters based on every method and examining the rankings, secondly by calculating the correlations between the measures.

Which methods are the most similar to each other in terms of the most central characters and in terms of correlation?

Task 03: Community detection

In this exercise, you have to continue analysing the data related to the Harry Potter series. You still need to work with the two above mentioned files.

Please note that the characters.csv file contains a large amount of missing values in some columns; it is up to you to decide how to deal with this issue. After identifying the most central characters in the previous task, now in order to gain more understanding of the network, you have to perform several tasks:

• perform community detection analysis using the technique introduced in the course. Determine the optimal number of communities based on modularity, compare the resulting communities and attempt to characterize the difference between communities by focusing on the information in the characters.csv file.

• in the previous task you identified the 6 most important characters according to PageRank; now check how they relate to the resulting communities: are all of them in different communities (and each community is centered around one of those important characters) or are there some communities that include more than one of the 6 important characters.