Project 1: The Big Bang Theory Scripts

Text Mining (G00C8A)

Martial Luyts

Description of the dataset

The Big Bang Theory is an American television sitcom created by Chuck Lorre and Bill Prady, both of whom served as executive producers and head writers on the series, along with Steven Molaro. It aired on CBS from September 24, 2007, to May 16, 2019, running for 12 seasons and 279 episodes.

Here, text scripts of the first 10 seasons are collected and stored as .csv file, consisting of 3 attributes:

- episode_name: Season no., episode no. and full name of the episode
- dialogue: Text script
- person_scene: Person, scene, etc. that is related to the text script

To import the data, the following python code can be considered:

```
#import dataset
import pandas as pd

path = 'C:/.../Scripts TBBT.csv'

df = pd.read_csv(path, sep=',', skipinitialspace=True, engine='python')
```

Instructions of this project

To prepare the text data for model building, preprocessing is crucial and need to perform with caution and precision. In this project, you are expected to perform several preprocessing steps from the text mining framework, in order to get valuable information from the scripts:

- Removing punctuations such as '!"#$%&()*+,-.:;?@[]\^_‘{|}~'
- Lower case the text
- Perform sentence and word tockenization on every script
- Remove stopwords. Remark: You can use the NLTK library for this task
- Use the Porter stemmer to perform stemming on the used words in the scripts
- Perform POS tagging & NER. Remark: You can use the Spacy library for this task

Based on these steps, please give an answer to the following questions. Remark: You need to choose which preprocessing step(s) are important to address the questions asked. Every group will answer these questions based on the assigned character (e.g., Leonard, Sheldon, Penny, etc.).

1. On average, how many sentences and words does your character have to speak per episode? Does this deviate across seasons?
2. Globally, over all episodes within the first 10 seasons, how many times does your character mention nouns, and person names? Make a Wordcloud of this tag/entity to have a clear visualization which nouns/person names are mostly used by your character.
3. What are the most important words mentioned by your character? Do this analysis per episode, per season and overall over the first 10 seasons. To achieve this task, please first make a bag-of-words and/or use the TF-IDF statistical principle. Remark: You can try to make a Wordcloud for visualization, based on the given bag-of-words.
4. Examine the co-occurence of words for your character by using the Positive Pointwise Mutual Information measurement. Which words are commonly used together in his/her dialogues? Remark: You can try to make a Word-Word co-occurence matrix.

You are expected to make this project in Jupyter Notebook, consisting of clear comments and codes such that the steps and conclusions made by you can easily be followed and verified. You need to upload this file (in .ipynb extension) **before December 13th, 2024 (23h59 BE time)**, on **Toledo**.

This project will count for **4 points of the total grade**.

Good luck!