Data Analytics – Exercises

(Week 11)

In these exercises, you will learn:

- to preprocess text as part of Natural Language Processing (NLP).
- to use NLP and a classification model to predict sentiments from hotel reviews.

In the data analytics process model, these exercises cover part of the steps "Statistical data analysis and/or Modeling" and "Evaluation & Interpretation" (see figure 1). Results of the exercises must be uploaded as separate files (**no .zip files!**) by each student on Moodle. Details on how to submit the results can be found in the tasks below.

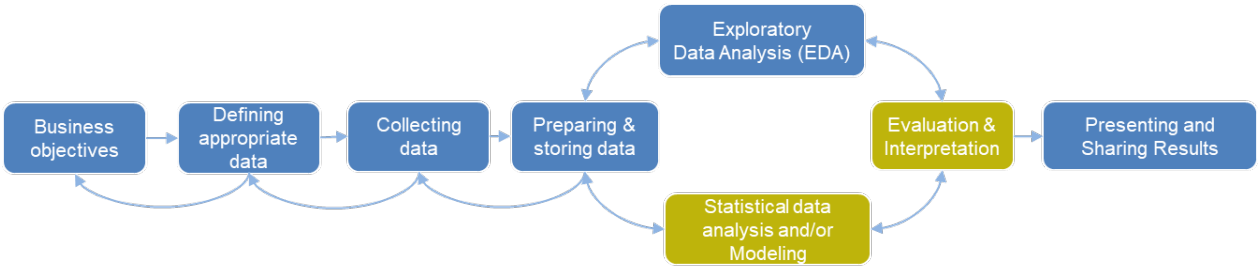

Figure 1: Data analytics process model (see slides of week 01)

Task 1

In this exercise, you will learn to use NLP methods to preprocess text. You will also learn to perform Part-of-Speech (POS) tagging. The tasks are:

a) Run the Jupyter notebook 'NLP_text_preprocessing.ipynb' step by step and try to understand, what the Python code does.
b) In the section 'Defining documents', three documents (d1, d2 and d3) are provided in the example. Define your own documents and perform all text pre-processing steps as shown in the section 'Text preprocessing'.
c) In the section 'Redefine the text corpus (pre-processed)' adapt the object named 'corpus' by hand. For this, use the pre-processed and lemmatized words from your documents.
d) Based on the 'corpus' from c), create a document-term matrix with ngram_range=(1,1) and a document-term matrix with ngram_range=(2,2).
e) Based on 'corpus' from c), create a:
    - Term Frequency (TF) matrix
    - Inverse Document Frequency (IDF) matrix
    - Term Frequency - Inverse Document Frequency (TF-IDF) matrix

f) In the section 'Part-of-Speach (POS) tagging', choose your own text example and derive POS tags for this example. In the Jupyter notebook, briefly explain the meaning of at least 5 POS-tags from the output of the POS tagging. Use the following webpage for help: https://pythonexamples.org/nltk-pos-tagging.

To be submitted on Moodle:

- The Jupyter notebook as html-file 'NLP_text_preprocessing.html' with the changes and short explanations according to b), c), d), e) and f).

Task 2

In this exercise, you will learn to use NLP methods and a classification model to predict sentiments from hotel reviews. In addition, you will perform your own sentiment classification based on movie reviews. The tasks are:

a) Run the Jupyter notebook 'NLP_hotel_reviews_sentiment_prediction.ipynb' step by step and try to understand what the Python code does.

b) Use the Jupyter notebook from a) as template to perform your own sentiment classification analysis. For this, use the movie reviews data from Moodle. Note that the movie reviews data are delimited by a tab (.tsv) instead of the comma used in .csv files. Below you can see how the data can be imported. So, to read a .tsv file, you must use `sep= '\t'` instead of `sep= ','` in `pd.read_csv()`. The movie sentiments are stored in the variable 'Rating' (0=low rating, 1=high rating).

Import movie reviews

```
# Import original data
data_orig = pd.read_csv('movie_reviews.tsv', sep="\t")[['sentiment', 'review']]
data_orig = data_orig.rename(columns={"sentiment": "Rating", "review": "Review"})

# Random sample from the data (to reduce computing time)
sub_00 = data_orig.loc[data_orig['Rating'].isin([0])].sample(n=1000, random_state=42)
sub_01 = data_orig.loc[data_orig['Rating'].isin([1])].sample(n=1000, random_state=42)

# Concatenating the subsets to one data frame
data = pd.concat([sub_00, sub_01]).reset_index(drop=True)
print('Number of movie reviews in data: {}'.format(len(data)))

data
```

```
Number of movie reviews in data: 2000
```

	Rating	Review
0	0	I am a back Batman movie and TV fan. I loved t...
1	0	Pay no attention to the comments behind the cu...
2	0	I happen to be the director's nephew. It's tak...
3	0	SLIGHT SPOILERS (but it doesn't matter anyway)...
4	0	I actually had quite high hopes going into thi...

To be submitted on Moodle:

The Jupyter notebook with your own sentiment classification according to b) as html-file 'NLP_movie_reviews_sentiment_prediction.html'.