TDT4310/2021S: Intelligent Text Analytics and Language Understanding Prof. Björn Gambäck

Lab Exercise 5

Lab Date: 19th of March 2021

1. **Guidelines**

   Deadline for submitting your solution:**time 9th April 2021**.
   Submission is a zipped folder with *your name* containing:

   (a) Source files (python): format name as Lab{LabNumber}_{Exercise Number}.py.

   (b) A summary/report file formatted as a pdf which explains and present the results with respect to the input value.

2. **Exercises**

   Exercise 1: Using the screenplay from Star Wars: The Empire Strikes Back or any other fictional work of your choice complete the following tasks. [1]

   (a) Find all entities in the text.

   (b) Find the frequency of each entity

   (c) Plot the frequency of the character (the entity class containing PERSON) to a graph of choice.

   (d) Does this graph adequately represent the importance of each character in the work. If it does, explain how, if does not, explain how you would find the most important characters in the work.

   Exercise 2: Named Entity Linking (NEL) is a an advancement of NER. NEL can be used to find every reference to the same entity in a text. Using the Knowledge Base(KB) known as Dbpedia, you're going to find all unique entities as well as extract some information from the KB about them.[2]

   (a) Using your favorite Wikipedia page (preferably for a person or character) find all unique people mentioned.

   (b) Extract some information about the people you found, that was not present in the original Wikipedia page.

   Exercise 3: Write a program to analyze sentiment of your tweet corpus using "*keras*", "*long short-term memory networks*" (LSTM). Hint: In the repository of this chapter, it contains sentiment analyzer code for reviews[3]

   (a) Analyze sentiment of a tweet to classify it in one of the classes: {positive, negative, neutral}.[4]

   (b) Using Keras and LSTM to build your model with the accuracy metric.

   (c) Evaluate the model with the {train data, test data} of {90%, 10%}

   (d) Using the trained model, plot the sentiment of Donald Trumps tweets to a graph using tweets from one month before the 2020 election until one month after the election.[5]

[1] Dump of the ESP script

[2] Resources for Dpbedia API, and Dbpedia spotlight

[3] `https://github.com/foxbook/atap/blob/master/snippets/ch12/deep_snark.py`

[4] You're allowed to use more advanced sentiment analysis like EmoLex

[5] `https://github.com/micaelaustad/TDT4310_Hint/blob/main/Lab%205/tweets/realDonaldTrump.json`