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

Lab Exercise 1

Lab Date: 22nd of January 2021

1. **Guidelines**

   Deadline for submitting your solution: **23:59 4th of February 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 presents the results with respect to the input values.

2. **Exercises**

   Exercise 1: With the following list of words
   `['she', 'sells', 'sea', 'shells', 'by', 'the', 'sea', 'shore']` perform the following tasks:

   (a) Print all words beginning with *sh*

   (b) Print all words longer than four characters

   Exercise 2: Read in the texts of the State of the Union addresses, using the *state_union* corpus reader.

   (a) Count occurrences of *men, women*, and *people* in each document.

   (b) Explain what has happened to the usage of these words over time using graphs.

   Exercise 3: *Pig Latin* is a simple transformation of English text. Each word of the text is converted as follows: move any consonant (or consonant cluster) that appears at the start of the word to the end, then append *ay*, e.g. *string* → *ingstray*, *idle* → *idleay*
   `http://en.wikipedia.org/wiki/Pig_Latin`

   (a) Write a function to convert a word to Pig Latin.

   (b) Write code that converts text, instead of individual words.

   (c) Explain how you would decode Pig Latin, either with words or code.

   Exercise 4: Reddit is a social media where people can share posts. Using the old web-page `old.reddit.com` as the source, build a web scrapper and present the following information about the posts. [1]

   - Amount of upvotes
   - Which subreddit the post was made in
   - The title of the post
   - The time when the post was made

   Exercise 5: Write a program to build your own custom *Corpus* (to a folder on your local computer) from tweets (e.g., using the results of 10 search queries or tweets from 10 accounts) on Twitter.[2]

   (a) Write code to tokenize the tweets and remove stopwords.

   (b) Print out the 10 most common words in the corpus.

   (c) Print out the 10 most common words from each user/search.

   (d) Print out the 10 most used hashtags in the corpus.

[1] If the page is down or you have issues with the scrapping of the webpage, you can find a dump here.

[2] If you have issues getting tweets from Twitter, you can pick terms from this set.