**TDT4310/2020S: Intelligent Text Analytics and Language Understanding** Prof. Björn Gambäck

# Lab Exercise 1

*Lab Date: 14th Jan 2020*

I. **Guidelines**:

Deadline for submitting your solution: **12pm 27th Jan 2020**. Submission is in zip format, containing:

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

(b) A summary/report file (pdf or word) to explain and present the results with respect to the input value.

II. **Exercises**: The *Brown* corpus is in *nttk.corpus*.

**Exercise 1 :** Understanding Task. Here we have a program from the code repository[1], which is used for Exercise 1 (renamed to Lab1_1.py). Thoroughly reading the code, understand what the code is, and write your comments to the code in the file to concisely describe/guide the purpose of the program, functions, arguments, and code.

**Exercise 2 :** Use the *book* corpus. A sentence is considered as a set of words between 2 '.'

(a) Write a program to find and print the slice for the complete sentence that contains the word at position text9.index("*sunset*") of the *book* corpus.

(b) Write a program to get a word as input, then write all the sentences in text9 that contain the input word.

**Exercise 3 :** 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 with 10 search words, or tweets from 10 accounts) on Twitter[2]. Write code to tokenize tweets (e.g. hashtag) and print out the ten most common words in the Corpus (remove English stop words), or the most common words in each users tweets, print out the ten most used hashtags. Hint: use *tweepy*.

**Exercise 4 :** Write a program to find all words that occur at least $k$ times in the first $n$ words in *Brown* Corpus. $k$ and $n$ are input parameters of the function.

**Exercise 5 :** "https://www.ntnu.edu/vacancies" is an address where NTNU post their "Vacancies and Job Openings". Write a program to do with the list of jobs on that page:

(a) Print how many jobs/vacations are currently posted on the address.

(b) Print all titles of jobs/vacations.

(c) Extract deadline of each post.

[1] https://github.com/foxbook/atap/blob/master/snippets/ch01/gender.py
[2] http://twitter.com/