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

Lab Exercise 4

Lab Date: 5th of March 2021

1. **Guidelines**

   Deadline for submitting your solution :**23:59 19th March 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: Develop your own NP chunker that converts POS tagged text into a list of tuples in the *Brown Corpus*, where each tuple consists of a verb followed by a sequence of noun phrases and prepositions.

   Example: "*the little cat sat on the mat*" becomes ('sat', 'on', 'NP') . . .

   Write out the first 20 results in the *Brown Corpus.*

   Exercise 2: Define an NP-chunk grammar consisting of the rules as below:

   $\{ < \text{NNP} > * \}$
   $\{< \text{DT} > ? < \text{JJ} > ? < \text{NNS} >\}$
   $\{< \text{NN} > < \text{NN} >\}$

   (a) Write a RegexpParser chunker with the rules above, and run the chunker on the first 5 sentences in the "*SpaceX.txt*" file. Print out the NP-chunk.

   (b) Print out the matching texts when running the NP-chunker.

   Exercise 3: Develop a context free grammar to parse sentences like:

   1. This dog runs.
   2. These dogs run.

   Write a function to do the following parsing:

   (a) Take a sentence (string) as input of the function.

   (b) Use global variable grammar by the above CFG.

   (c) Print all trees (or a notification message if the sentence is invalid)

   (d) Return the first one.

   Exercise 4: This exercise requires you to make your own language model[1] to generate new tweets based on Donald Trump's tweets using n-grams. For example, given an input "make America", your model needs to calculate the most likely complete sentence, and print out that sentence, i.e. "make America great again!".
   Trump's twitter account is @realDonaldTrump - Hint[2]

[1] https://www.nltk.org/api/nltk.lm.html

[2] You can find a repository with tweets from 2011 to 2021 here.