Project Text Analysis – Assignment 1

Malvina Nissim
m.nissim@rug.nl

15 April 2019

The aim of this assignment is to get you acquainted with text manipulation at the word level using `nltk`. You will have to use several `nltk` functions, whose theory we have seen in class. Some of the functions we have seen in the slides, some others you will have to figure out yourself (please, refer to the `nltk` online material). Please, embed your functions in proper python code.

- use Nestor for submission
- **Deadline**: 29 April 2019, 23:59
- hand in three files, two `.py` and one `.txt/.pdf`:
  - one python script that generates all the information required in Exercise 1
  - one python script that generates all the information required in Exercise 2
  - a `.txt` or `.pdf` file (please no `.docx`) with all comments/answers to the questions in Exercise 2. If you want, you can also included comments about Exercise 1. If your scripts generate (almost) all the information you need in a nice informative way you can just copy and paste it from your code's output, of course. Add comments where appropriate.

For the first exercise you will have to use data which is already part of the NLTK distribution, while for the second one you will have to use the text file which is provided within the assignment, so as to start with completely raw data.

Exercise 1 – Trying out NLTK

In this simple part of the assignment you just have to try out all of the code that is included in the slides from Lecture 1. There isn't much reporting about this, but you have to hand in the python code you put together commenting appropriately in the code what your lines are doing. You can choose which file you will work with out of those already included in the NLTK distribution (possibly pick one which is **not** the one used in the slides).

Exercise 2 – N-grams and Sentences

For this part of the assignment you will work with n-grams. The text file you will have to work with as input is loaded on Nestor. Differently than for the previous exercise, you will not work with a pre-included text in NLTK, but you will have to start from scratch with the text that it's provided for you. After doing some manipulation, you will have to answer some questions. Please, remember the difference between tokens and types.

Note: all of the tasks that you have to perform for this assignment must be run within a **single** python script. It's nice if your script prints a few headings/comments referring to what it's doing.

1. work with sentences, and print:
   (a) the longest sentence (if there is more than one of the same length, print both or pick one at random, but specify this)
   (b) the shortest sentence (if there is more than one of the same length, print both or pick one at random, but specify this)
   (c) the distribution of sentences in terms of length (how many sentences of length 1, how many sentences of length 2, and so on, the same as you would do for types and tokens)
   (d) the average sentence length in the whole document
2. work with n-grams, and list:
   (a) how many **character types** are found, followed by the character type list, alphabetically ordered.
   (b) how many **word types** are found, followed by the word type list, alphabetically ordered
   (c) the top 20 **character-level unigrams, bigrams, and trigrams**, ordered by descending frequency (from the most frequent to the least). Please, add a comment specifying what you think this information could be useful for.
   (d) the top 20 **word-level unigrams, bigrams and trigrams**, ordered by descending frequency. Please, add a comment specifying what you think this information could be useful for.