Natural Language Processing 2020-2021
Homework 1:
Words: Morphology, Spelling and Normalization

Deadline: 10 September (23:59).
Questions? Post them in the HW1 discussion on Canvas.

This assignment consists of a number of small exercises. Not much programming is required. The main thing to hand in is a report with the answers to the questions below. Please make your report not more than 10 pages long, excluding figures! (And font not smaller than 10pt)

Exercise 1: Morphology

Have a look at the first page of Chapter 2 of the book by J&M. On this page, try to find 3 examples (types, not tokens) of each of the following word formation categories (see lecture slides):

- Inflection
- Derivation
- Compounding
- Cliticisation

Don't include the ones that were already discussed in class! Some categories may occur less frequently, so if you can't find three examples of each category on the page, you can list less. Explain for at least one example per category why it is an example of this type of word formation.

Exercise 2: Stemming

In this exercise you will stem the word types (unique words) found on the page from Exercise 1. As input you use the file `sorted_types.txt` provided with this assignment on Canvas. It lists all the word types in alphabetical order. (We skip the step of *tokenization*, which normally comes before stemming. Tokenization will be addressed in homework 2.)

Stem the word list using the Porter Stemmer within NLTK (Natural Language Toolkit), a very useful NLP library. Here you can find a tutorial on how to do that:

`https://www.datacamp.com/community/tutorials/stemming-lemmatization-python`

Based on the stemmed word list, answer the following questions:

2a The original word list contained N word types. How many word types (= unique word stems) are left after stemming?

2b Did you find anything interesting about the results? Give examples of where the stemming worked well and where there are errors of omission and/or commission.

Exercise 3: Lemmatization

In this exercise you are going to lemmatize the words from the same page of the book, again using NLTK. (Optionally, you can also try to do it using the NLP library SpaCy, which also includes a lemmatizer: `https://spacy.io/`.)

Lemmatize the words. Based on the lemmatized word list, answer the following questions:

3a How many word types (= unique lemmas) are left after lemmatization?

3b Give 3 examples of lemmatization errors, if you can find them. Explain why they are errors.

Exercise 4: Creative spelling analysis

In this exercise, 'creative spelling' means the intentional misspelling of existing words. For this you use the blog dataset that is provided on Canvas with this assignment. In the data set, locate file M-train-136.txt. Find as many examples as you can of *intentional* (as far as you can tell) misspellings in this document. You can do this through manual inspection, or by using a spell checker.

4a Have a look at the transformation categories of Mosquera & Moreda (2014) (see the slides of 3 September or the paper itself). Provide examples of 3 misspellings per category, if you can find them.

4b Give a rough estimate of the ratio of misspellings per category. Which trends (if any) can you see in the types of word transformations used by this blogger? Which types of misspelling are used more or less often?

4c Have a look at Section 2.3 of the J&M book, which discusses language variation and how this reflects demographic characteristics of the speaker. Which observations can you make about the language use in this blog, and what do you think their language use says about the blogger? There is no need for an extensive analysis; just briefly (in one or two paragraphs) mention anything that you found noticeable or remarkable.

Exercise 5: Number transliterations: analysis and normalization

A common spelling variation in 'Netspeak' is to replace sequences of characters with numbers that sound the same (for example, *gr8!* instead of *great!*). Regular expressions can be used to find such number transliterations in a text corpus.

Various tools exist that allow you to search a collection of text documents using regular expressions. An example tool for Windows that allows both search and replacement is PowerGrep (15 days free evaluation trial): `https://www.powergrep.com/download.html`. If you like working with Python, this also has good functions for regular expressions search and replacement.

Use PowerGrep or another tool of your own choice to find all cases of number transliteration in the blog corpus (or at least as many as you can). Use the entire blog corpus for this exercise, ignoring the distinction between training and test files. Do it for male and female bloggers separately. Blogs of female bloggers start with F; blogs of male bloggers start with M.

5a Provide the regular expression (or expressions) you used to get your results and explain it. Also mention which tool you used.

5b Per number (0...9) provide a list of the top 5 number transliterations, once for the male bloggers and once for the female bloggers. In other words, for each number find the 5 most frequent transliterations involving that number, where the frequency is the number of the same tokens in the document with a particular number transliteration. Use *case folding* so that tokens with and without capitalisation are grouped together. For each number, also give the total number of transliterations (tokens) involving that number (for the male and female bloggers separately).

5c Briefly discuss your findings. Which numbers are most often used in number transliterations? In which words? Are there any differences between male and female bloggers?

5d Try to normalize the number transliterations with 8 in the corpus (other numbers not necessary) using *string substitution*; see Section 2.1.6 from the book. Explain how you did this and which (if any) problems you came across. You are not required to come up with a perfect solution; just give it a try, and describe your experiences / any problems you encountered in about half a page.

Handing in

Hand in the following things on Canvas (submission as a group):

- Your answers in a Word or pdf document. Please include the name of both group members in the document!
- For exercises 2 and 3, also submit the stemmed and lemmatized word lists.