Natural Language Processing 2023-2024
Homework 1:
Words: Morphology, Spelling and Normalization

Deadline: 13 September (23:59).
Questions? Post them in the HW1 discussion on Canvas or send them to nlp-course@utwente.nl.

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 8 pages long, excluding figures. (And font not smaller than 10pt)

Also, please adhere to the Guidelines for using AI during your studies at UT.

Exercise 1: Morphology (1.5 pt)

Have a look at the *second* (!) 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

Some categories may occur less frequently, so if you can't find three examples of each category on the page, you can list fewer examples. Explain for one example per category WHY it is an example of this type of word formation.

Exercise 2: Stemming (1.5 pt)

In this exercise you will stem all the word types (unique words) from the book page mentioned in 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.)

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 (0.5 pt) The original word list contained 245 word types. How many word types (= unique stems) are left after stemming?

2b (1 pt) Do you notice anything interesting about how some words are stemmed? Briefly discuss your observations about the stemming. Give examples of where the stemming worked well, and where there are errors of omission and/or commission (if you can find any).

Exercise 3: Lemmatization (1.5 pt)

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 (0.5 pt) How many word types (= unique lemmas) are left after lemmatization?

3b (1 pt) Do you notice anything interesting about how some words are lemmatized? Briefly discuss your observations about the lemmatization. Give 3 examples of lemmatization errors, if you can find them. Explain WHY they are errors.

Exercise 4: Creative spelling analysis (1 pt)

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 F-train-146.txt. Read the blog text and look at how the words are spelled. (The file contains the text of multiple blog posts from the same blogger, so don't expect a coherent story.)

4a (0.5 pt) Discuss the blogger's spelling in terms of the transformation categories from the paper of Mosquera & Moreda (2014) (listed in Table 1 and explained on the next page of the paper). What kind of (intentional) misspellings does this blogger tend to make? Give some typical examples from the main categories used by the blogger.

4b (0.5 pt) Read Section 2.3 of the J&M book, which discusses language variation and how this reflects the demographic characteristics of the speaker. What do you think their language use says about the blogger? Briefly (in one or two paragraphs) mention anything that you found noticeable or remarkable.

Exercise 5: Vowel duplications (4 pt)

A common spelling variation used online is to duplicate characters for emphasis (for example, *reeeeaaallllyyyy?!*). In this exercise we limit ourselves to duplication of vowels (a,e,i,o,u) and use regular expressions to find words with such duplications in a text corpus.

If you like working with Python, this has good functions for regular expressions. If you prefer to use a specific tool, 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, which offers a 15 days free evaluation trial: `https://www.powergrep.com/grep.html`. If you like working with Python, this also has good functions for regular expressions search and replacement.

Use Python, PowerGrep or another tool of your own choice to find all cases of vowel duplication 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.

5a (1 pt) Provide the regular expression (or set of regular expressions) you used to get your results. Provide an explanation and motivation of the regular expression(s) you used. Also mention which tool you used.

5b (2 pt) For each vowel (a,e,i,o,u) , provide the top 3 most frequent words with duplications of that vowel. The frequency is the total number of tokens in the document for a specific word type involving duplication of that vowel. The number of times the vowel gets duplicated does not matter. As an example, for the purpose of this exercise we regard *coool, cooool* and *coooool* all as tokens of the word type *cool*). So if you find *coool* 2 times, *cooool* 6 times and *coooool* 3 times (and no other spelling variations of the word *cool* that involve duplication of the vowel), then the frequency of *cool* with vowel duplication is 2 + 6 + 3 = 9. Use *case folding* so that tokens with and without capitalisation can be counted together.

5c (1,5 pt) Use *string substitution*(see Section 2.1.6 from the book) to normalize the vowel duplications in the corpus. Explain how you did this and which (if any) problems you came across. You don't need to come up with a perfect solution: just give it a try, and describe your experiences / any problems you encountered in around half a page.

5d **BONUS QUESTION** (optional, for 1 bonus point) Do exercise 5b separately for for male and female bloggers. Blogs of female bloggers start with F; blogs of male bloggers start with M. Show the results and discuss the differences you found between male and female bloggers.

Handing in

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

- Your answers in a 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.