SNLP 2016
Exercise 4

Submission date: 27.05.2016, 23:59

You can use NLTK for text normalization (incl. punctuation removal, stemming/lemmatizatio and tokenization) you need in this exercise sheet.

Short and Long Range Dependencies

1) (5 points) In this exercise you will experience analyzing the Short/Long Range Dependencies by Correlation Function.

- (1 points) Use the document ”poem.txt” provided in the materials. Start with text normalization and change all different versions of the word ”you” like ”your, you'll, you've”into ”you”.
- (2 points) Use the correlation function provided in page 47 of Slide ”Chapter 4” and compute it for the word ”you” with different distances of 1 to 50 ($\forall$d$\in [1, 50]$, d$\in \mathbb{N}$)
- (2 points) Plot the correlation vs. distance with the values obtained in the previous section. How do you explain the plot?

OOV

2) (5 points) In this exercise you will experience the relationship between Out Of Vocabulary rate and Size of Vocabulary.

- (1 point) Use each of the documents provided in *train* folder to construct 5 different vocabularies.
- (1 point) Use the ”test.txt” document provided in the materials which is about *Physics* as a test to compute OOV using each of the vocabularies.
- (1 point) Plot OOV vs. Size of the Vocabulary. How do you explain the plot?
- (1 point) How do you explain the importance of OOV rate? Do you think out of vocabulary words will cause issues in SNLP like computing probability of a sequence of words? If yes, how do you offer to solve this issue?
- (1 point) How many different bigrams does the training data contain? How many bigrams in the test file do not occur in the training data. Give the same numbers for trigrams.

1 Submission Instructions: Read carefully

- You can form groups of maximum 3 people.
- Submit only 1 archive file in the ZIP format with name containing the MN of all the team members, e.g.:

Exercise_01_MatriculationNumber1_MatriculationNumber2_MatriculationNumber3.zip

- Provide in the archive:
  - your code, accompanied with sufficient comments,
  - a PDF report with answers, solutions, plots and brief instructions on executing your code,
  - a README file with the group member names, matriculation numbers and emails,
  - Data necessary to reproduce your results [1]
- The subject of your submission mail must contain the string [SNLP] (including the braces) and explicitly denoting that it is an exercise submission, e.g:

[SNLP] Exercise Submission 01

- Depending on your tutorial group, send your assignment to the corresponding tutor:
  - Sedigheh Eslami: *eslami@mpi-inf.mpg.de*
  - Naszdi Kata: *b.naszadi@gmail.com*
  - Stephanie Lund: *stflund@gmail.com*

2 General Information

- In your mails to us regarding the tutorial please add the tag [SNLP] in the subject accompanied by an appropriate subject briefly describing the contents.
- Feel free to use any programming language of your liking. However we strongly advise in favor of Python, due to the abundance of available tools (also note that Python3 comes with an excellent native support of UTF8 strings).
- Avoid using libraries that solve what we ask you to do (unless otherwise noted).
- Avoid building complex systems. The exercises are simple enough.
- Do not include any executable files in your submission, as this may cause the e-mail server to reject it.
- **In case of copying, all the participants (including the original solution) will get 0 points for the whole assignment. Note: it is rather easy to identify a copied solution. Plagiarism is also not tolerated.**
- **Missing the deadline even for a few minutes, will result in 50% point reduction. Submission past the next tutorial, is not corrected, as the solutions will already be discussed.**
- **Please submit in your solutions necessary to support your claims. Failure to do so, might results in reduction of points in the relevant questions.**

[1] If you feel that these files are beyond reasonable size for an email submission and also reasonably convenient, please provide a means for us to access them online

- Each assignment has 10 points and perhaps some bonus points (usually 2 or 3). In order to qualify for the exams, you need to have 2/3 of the total points. For example, in case there are 12 assignments, you need to collect at least 80 out of the 120 points to be eligible for the exams. A person that gets 10 plus 2 bonus points in every exercise, needs to deliver only 7 assignments in order to be eligible for the exams, since 7*12=84.
- Attending the tutorial gives 2 points increase for the corresponding assignment.
- Exercise points (including any bonuses) guarantee only the admittance to the exam, however have no further effect on the final exam grade.