Sheet 1

SNLP 2016

Due date: 06.05.2016

In this assignment you will learn about Text Preprocessing (normalization and tokenization) for text mining and information retrieval purposes. You will also examine Zipf's Law on different natural languages.

1 Text Preprocessing

1.1 (2 points)

Research briefly on common Text Preprocessing techniques (e.g used for Search Engines). Name at least four, briefly explain them and the reason they are used.

1.2 (3 points)

Provide 3 documents in 3 different languages (e.g English, German, French). You can use *Gutenberg Project* [1] if you want. (1 point)

Use the result of your research in part 1.1 to implement a reasonable tokenizer for Zipf's Law examination. Apply your tokenizer on the documents provided in the previous section. (2 points) [2]

2 Zipf's Law

2.1 (1 point)

Compute frequency and rank of each token obtained in the previous part. Report the first rank token.

2.2 (4 points)

Plot frequency vs. rank. Apply log function in your plot. (3 points)[3]
How do you explain the effect of "log" function here? Why is it applied? (1 point)

[1] `https://www.gutenberg.org/`
[2] "Stemming" would have 1 bonus points
[3] Plots for cases with "stemming" would have 1 bonus point

3 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 [4]
- 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*
  - Naszádi Kata: *b.naszadi@gmail.com*
  - Stephanie Lund: *stflund@gmail.com*

4 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.

[4] 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

- **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 information to support your claims. Failure to do so, might results in reduction of points in the relevant questions.**
- 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 for the person who attends the turoial, not for the group as such. These points will count as bonus points towards the qualification for the exam.
- Exercise points (including any bonuses) guarantee only the admittance to the exam, however have no further effect on the final exam grade.