Statistical Natural Language Processing

Exercise Sheet 1

Due date - 05.05.17

1 Zipf's Law

1.1 (2 points)

On the website www.gutenberg.org you will find texts in many languages. Choose a text which has a version in English, German and French. Generate word frequency tables for each version. You can apply a tokenizer of your choice for the purpose.

Plot the frequency of the words in the three texts against their rank (both axes on a log scale).

1.2 (1 points)

What do you observe from the different plots. How do you explain the differences between them?

2 Implementation of the model on slides 18 and 19 of SNLP chapter 2

2.1 (2 points)

Generate a random text with words of varying length obtained by observing the following rules and draw the corresponding Zipf's plot.

a. Hit spacebar with probability $p = 0.2$ and all other keys with a probability $(1 - p)$
b. Never hit spacebar twice in a row.

2.2 (1 points)

Plot the probability of the individual characters in the generated text.

2.3 (2 points)

Now extend the model to be more realistic. Have a separate state for each of the 26 characters a-z in addition to a state for generating a space. The transition probability into a new state should not depend on the present state but on the

probability of a characters P(c) (c is a character a-z or space). To estimate P(c) use relative frequencies as counted on the English text you used in question 1.2. Ignore other characters. Use this expanded model to generate text and create a zipf plot for the generated words.

3 Probability Theory

3.1 (1 point)

Use set theory and the definitions of probability functions to show that,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

3.2 (1 point)

Are X and Y, as defined in the following table, independently distributed ?

x	0	0	1	1
y	0	1	0	1
$P(X = x, Y = y)$	0.32	0.08	0.48	0.12

Table 1: Joint probability table

4 Submission Instructions

- You can form groups of maximum 2 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.zip

- Provide in the archive:

  i. your code, accompanied with sufficient comments
  ii. a PDF report with answers, solutions, plots and brief instructions on executing your code
  iii. a README file with the group member names, matriculation numbers and emails
  iv. Data necessary to reproduce your results

- 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 MatriculationNumber1 MatriculationNumber2

- Submit to snlp_tutors_2017@lsv.uni-saarland.de .