Statistical Natural Language Processing

Exercise Sheet 1

Due date - 27.04.18

1 Zipf's Law

1.1 (2 points)

In the lecture, you came across the concept of Zipf's Law. When Hermione and Parvati learned about it at Hogwarts, they tried to make their own conclusions about the law. We want you to join their team and help them. Parvati was familiar with her Indic languages and reads usually Indic books, while Hermione reads novels in European languages. Here is what they are curious to know:

What does the word-token frequency table for a novel of Parvathi and Hermione look like? You can apply a tokenizer of your choice for the purpose. Feel free to tell Hermione and Parvathi that you want to include the books from your native language. There should be at least three novels in different languages which have different mother language. For example: English, Hindi, Chinese, Arabic have different mother language.

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

1.2 (1 point)

Write down the conclusions made by Hermione after observing the different plots. How do you explain the differences between them to Hermione?

1.3 (1 point)

Write your explanation to Parvathi on how does the dependency between the morphology of the languages and Zipf's law looks like.

2 Implementation of Zipf's Law

2.1 (2 points)

Malfoy wants to generate some random text in English and somehow complete his essay. Help him as well by following the below rules and draw the corresponding Zipf's plot:

a. Type a vowel (a,e,i,o,u) with probability $p = 0.4$ and all other keys with a probability $(1 - p)$.
b. Never hit a vowel twice in a row.

2.2 (1 points)

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

2.3 (1 points)

Now build another model to help Ron. 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 1.1. Ignore other characters. Use this expanded model to generate text and create a Zipf plot for the generated words.

Shout 'expecto patronum' to help you solve the assignment if required!

3 Where artificial intelligence went wrong? Or didn't it?(2 points)

There has been a discussion about statistical methods. Chomsky opposes them:

"Chomsky argued that the field's heavy use of statistical techniques to pick regularities in masses of data is unlikely to yield the explanatory insight that science ought to offer. For Chomsky, the "*new AI*—focused on using statistical learning techniques to better mine and predict data— is unlikely to yield general principles about the nature of intelligent beings or about cognition."

Source https://www.theatlantic.com/technology/archive/2012/11/noam-chomsky-on-where-artificial-intelligence-went-wrong/261637/

Peter Norvig from Google wrote a reply in favor of statistical methods http://norvig.com/chomsky.html .

Please summarize both views and discuss them critically. What is your own view?

Don't write more than one page!

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_2018@lsv.uni-saarland.de` .