Statistical Natural Language Processing

Exercise Sheet 2

Due date - 04.05.18 (23:59)

1 Probability Theory

1.1 (1 point)

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

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

1.2 (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

2 Zipf's Law

2.1 (2 points)

While Hermione was studying her book about Magical Theorie, she recalled Zipf's Law that states a relationship between frequency f and rank r of a word in a text:

$f \propto \frac{1}{r}$ or, in other words: There is a constant c such that $f \cdot r = c$

With this, she would be able to numerically calculate the training perplexity of a text with a Zipfan distribution. Assuming $c = 0.1$ for English texts and a vocabulary size of 10,000 words, you should do the same and calculate the training perplexity for a text with the given parameters.

2.2 (2 point)

How does perplexity change when you vary c in [0.01 .. 0.5] and the vocabulary size in [100 .. 1,000,000]?

3 Perplexity

3.1 (2 points)

Later that day, Professor Flitwick introduced Hermione into the basics of magic spells. The professor told the students that every spell is part of a large language system and every spell is made up of multiple components. For example, *Bombarda Maxima* consists of two components. The network given in Figure 1 displays a subset of these spells. As a newly Zipf expert, Hermione wants to check whether the language of magic spells follows the Zipfian distribution. For this, you should help her by explaining if this statement is plausible on a component level, i.e. every component is considered a "word" in the magic language, assuming a uniform distribution of all possible magic spells that can be constructed with Figure 1. [1]

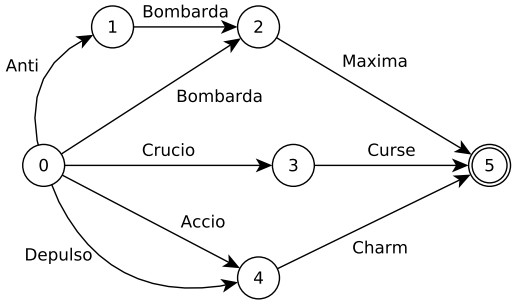

Figure 1: Finite-State network for Harry Potter spells.

3.2 (2 points)

Using the language of magic spells, calculate the perplexity of this language. Again, assume uniform distribution of all magic spells and operate on component level.

4 Submission Instructions

- You can form groups of 2 to 3 people
- Submit only 1 archive file in the ZIP format with name containing the MN of all the team members, e.g.:

  Exercise_02_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

[1] Check this paper, if you want to learn more about FST `https://web.stanford.edu/~laurik/publications/pmatch.pdf`

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` if you want to submit before Tuesday. Submission instructions will be updated post doodle results.