Statistical Natural Language Processing (SS-2018)

Submission Deadline: 18.05.2018, 23:59

Entropy (5.5 points)

1) (1 point) In a certain village, the women who live beyond the age of 70 outnumber men in the same age by three to one ratio. How much information, in bits, is gained by learning that a certain person who lives beyond 70 happens to be male?

2) (1 point) Why are fixed length codes inefficient for alphabets whose letters are not equiprobable? Discuss this in relation to Morse Code.

3) (1.5 points)

$X\vert^{Y}$	0	1
0	1/4	1/4
1	1/2	0

Calculate the conditional entropies H(Y|X), H(X|Y) and mutual information I(X|Y)?

4) (1 point) A successful bet on the occurrence of an event i returns T$_i$ units of wealth (including the original stake) for each unit wagered. Otherwise the stake is lost. Let p$_i$ be the probability of event i. On each trial a gambler allocates all available funds, betting a fraction b$_i$ of total funds on event i. Write down an equation for the expected rate of increase of the gamblers wealth. How should the b$_i$ be chosen in order to maximize this expected rate of increase? Prove that your strategy gives the maximum value. Any properties of the cross entropy proved in the course may be used without proof, if carefully stated.

5) (1 point) Prove H(Y|X)$\leq$ H(Y). *Hint: Start with definition of H(Y|X) and use Jensen's inequality.*

Out of Vocabulary (OOV) (3 points)

6) Construct a different vocabulary using each of the documents provided in the training folder in Materials, e.g., for n documents in the training folder you will create n different vocabularies.

   - Now use the document provided in the test folder in Materials to compute **OOV** using each of the vocabularies that you created. (1 point)
   - Plot **OOV** vs size-of-vocabulary and explain the plot. (1 point)
   - How do Out-Of-Vocabulary words affect tasks like computing probability for a bigram i.e. P(w$_0$|w$_{-1}$)? Explain a possible remedy. (1 point)

Correlation vs Distance (1.5 points)

7)
   - From NLTK obtain the text “carroll-alice.txt”. Start with text normalization and change all different versions of the word “you” like “your, you'll, you've” into “you”. Print out the modified version in a text file. (0.5 point)
   - Compute the correlation for the word “you” with different distances of 1 to 50 ($\forall d \in [\,1, 50]\,, d \in \mathbb{N}$). Use the correlation function provided in slides of chapter 4 in SNLP. Now produce the plot correlation vs. distance. (1 point)

Submission Instructions

- You must 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_04_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
- Depending on your tutorial group, please send your assignment to the **corresponding tutor**:
  - ∗ Mo. 16-18: Lukas Lange *s9lslang@stud.uni-saarland.de*
  - ∗ Th. 14-16: Harshita Jhavar *snlp18.thursday@gmail.com*
  - ∗ Fr. 10-12: Marimuthu Kalimuthu *mkalimuthu@lsv.uni-saarland.de*