UNIVERSITÄT DES SAARLANDES

Statistical Natural Language Processing

Summer Semester 2020

Prof. Dr. Dietrich Klakow

Exercise Sheet III

Submission Deadline: May 29th, 23:59

Basics of Language Modeling

1) **N-gram probabilities (4 points)**
In this exercise, you will compare the probability distributions of $P(w_i|w_{i-1} = "in")$ and $P(w_i|w_{i-1} = "the")$. This is the probability distribution of words given the preceding word is "in" or "the", respectively.

- Download the Brown corpus through the python NLTK toolkit or from the web: http://www.nltk.org/nltk_data/
- Tokenize (split on white-space) and lowercase each token.
- Estimate the conditional probability distributions $P(w_i|w_{i-1} = "in")$ and $P(w_i|w_{i-1} = "the")$ using relative frequencies.
- For the 20 most frequent tokens for both distributions, plot either the frequency distribution (unnormalized frequency counts) or the probability distribution.
- Compute the expected value of $-log_2(P(X))$ i.e $E[-log_2 P(X)]$ for both distributions. Which distribution has a higher expected value? What could be a reason for the difference?

2) **Perplexity (6 points)**
Next, we want to have a look at how to calculate and interpret perplexity for unigram and bigram language models (See lecture 3 from slide 13).

(a) First, pre-process (tokenize, lowercase, remove punctuation) the "English_train.txt". Please do not use NLTK for this.

(b) Now, estimate both unigram and bigram conditional probabilities from the pre-processed corpus.

(c) Implement the perplexity function from lecture 3, slide 21 for both language models and test their performance on "English_test.txt". Report the perplexity values. Which model has a lower perplexity value and why?

Hint: Apply Lidstone smoothing on the test sample to account for words that are missing in the training sample. The formula is:

$$P(w|h) = \frac{N(w,h) + \alpha}{N(h) + \alpha V}$$

where $N(h)$ is the absolute frequency of h, and V the size of the vocabulary. Use $\alpha = 0.03$ for this example. Why is smoothing important?

(d) Estimate bigram probabilities for 20% of the corpus ("English_train.txt") and compute the perplexity on the test set. Repeat the experiment for 40% , 60%, 80% and 100% of the corpus. Plot the change in perplexity and explain your observation.

Bonus:

Perplexity vs. Alpha (1 point)

Add a plot showing how perplexity changes for both unigram and bigram probabilities with varying alpha. Use the alpha range [1.0, 0.3, 0.1, 0.03, 0.01, 0.003, 0.001]. What do you observe? Why do you think this happens?

Expected values (1 point)

Prove the following expectation properties:

(a) $E[-logP(X,Y)] = E[-logP(Y|X)] + E[-logP(X)]$

(b) $E[-logP(X)] - E[-logP(X|Y)] = E[-logP(Y)] - E[-logP(Y|X)]$

Submission Instructions

The following instructions are mandatory. Please read them carefully. If you do not follow these instructions, the tutors can decide not to correct your exercise solutions.

- You have to submit the solutions of this exercise sheet as a team of 2 students.
- If you submit source code along with your assignment, please use Python unless otherwise agreed upon with your tutor.
- NLTK modules are not allowed, and not necessary, for the assignments unless otherwise specified.
- Make a single `ZIP` archive file of your solution with the following structure
  - A `source_code` directory that contains your well-documented source code and a `README` file with instructions to run the code and reproduce the results.
  - A `PDF` report with your solutions, figures, and discussions on the questions that you would like to include. You may also upload scans or photos of high quality.
  - A `README` file with group member names, matriculation numbers and emails.
- Rename your `ZIP` submission file in the format

  `exercise02_id#1_id#2.zip`

  where `id#n` is the matriculation number of every member in the team.
- Your exercise solution must be uploaded by only one of your team members under *Assignments* in the *General* channel on Microsoft Teams.
- If you have any problems with the submission, contact your tutor before the deadline.