BBM 497: Introduction to NLP Lab.

ASSIGNMENT I

Handed out: 03.03.2017 *Handed in: 24.03.2017*

Please submit your solution (code and a PDF of your report) by 17:00pm on the due date. Please describe your code in a separate report. Your reports should not exceed a page.

Code and data

All files that are necessary to do the assignment are contained in a zip file which you can get from Piazza.

Language models and smoothing

You will train a unigram, bigram and trigram model on a training corpus and then test the three models on a testing corpus.

Training and test data: We have one training corpus that consists of Shakespeare's comedy plays and a test corpus that consists of Shakespeare's historical plays. All texts include some tags that are marked with <*> for the beginning of the text and </*> for the end of the text. You need to take out all the tags in order to have the real text for both training and testing. You may use regular expressions for this task. You also need to separate punctuation marks by whitespaces in order to take those as also tokens.

Smoothing Use add-one smoothing to smooth the models for zero probabilities.

Task 1: Estimating the probability of a given sentence Use the training texts in order to estimate the probability of the following sentences according to your smoothed bigram model.

- To work or not to work, that is the problem!
- Shall sleep more, Theodore shall sleep more.
- It does not matter how slowly you go so long as you do not stop.
- Imagination is more important than knowledge…
- Thou seest, the heavens, as troubled with man's act

Task 2: Generating sentences Please use your unigram, (unsmoothed) bigram, and trigram language models to generate 10 sentences, and compare the probabilities of the unigram, (smoothed) bigram, and (unsmoothed) trigram models assigned to these sentences. Whenever the next token is one of the end of sentence punctuation marks (i.e. ".", "!", "?", "…"), terminate your sentence. Otherwise, limit the number of words in each sentence by 20.

Task 3: Computing the perplexity of the test data You need to compute the perplexity (normalized inverse log probability) of the test corpora according to the unigram model, the smoothed bigram model, and the smoothed trigram model.

For a corpus W with N words, remember how the perplexity is calculated:

$$Perplexity(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{p(w_1 w_2 ... w_N)}}$$

In order to avoid any underflow error, you will have to use the log probabilities for the calculation. According to the base 2, the perplexity is calculated as follows:

$$Perplexity(W) = 2^{-\frac{1}{N}\sum_{i=1}^{N} \log_2 p(w_1 w_2 ... w_N)}$$

Submission

You need to implement either in Java or Python. Please submit your source codes and a one-page report in the following submission format.

Submit Format:
This file hierarchy must be zipped before
→ <student id>
→ code.zip
→ report.pdf