SNLP 2017
Exercise 2

Due date - 12.05.17 23:59

1) (3 points) In this exercise you will compare the probability distributions of $P(W_i|W_{i-1} = "in")$ and $P(W_i|W_{i-1} = "the")$. The distribution of the words given the previous word is "in" or "the" respectively.

   - Download the Brown corpus from the web `http://www.nltk.org/nltk_data/` or through the python NLTK toolkit.
   - Tokenize 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.
   - Plot the frequency distribution (unnormalized frequency counts) or the probability distribution for the 20 most frequent tokens for both distributions.
   - 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?

2) (2 points) Prove the following expectation properties:

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

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

Perplexity

3) (5 points) In the task, you will implement perplexity (Slide 3, page 20) for unigram and bigram language models (page 13).

   (a) First, pre-process (tokenize lowercase, remove puncutation) and tokenize the "English1.txt". You may use any functions from NLTK for this task.

   (b) Estimate the unigram and bigram conditional probabilities from the preprocessed corpus.

   (c) Implement the perplexity function for both language models and test their performance on "English2.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 given below:

$$P(w|h) = \frac{N(w,h) + \alpha}{N(h) + \alpha V} \tag{1}$$

   where $N(h)$ is the absolute frequency of h, and V the size of the vocabulary. We'll use $\alpha = 0.3$ for this example. Why is smoothing important?

   (d) Estimate bigram probabilities for 20% of the corpus ("English1.txt") and compute the perplexity on the test set. Repeat the experiment for 40% , 60%, 80% and 100% of the corpus. Plot the perplexity values as the size of corpus increases. Explain your observation.

1 Submission Instructions: Read carefully

- 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_02_MatriculationNumber1_MatriculationNumber2.zip

- Provide in the archive:
  - your code, accompanied with sufficient comments,
  - a PDF report with answers, solutions, plots and brief instructions on executing your code,
  - a README file with the group member names, matriculation numbers and emails,
  - Data necessary to reproduce your results [1]
- 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 02

- Depending on your tutorial group, send your assignment to the corresponding tutor:
  - David Adelani: *s8daadel@stud.uni-saarland.de*
  - Rajarshi Biswas: *rbisw17@gmail.com*
  - Xiaoyu Shen: *s8xishen@stud.uni-saarland.de*

2 General Information

- In your mails to us regarding the tutorial please add the tag [SNLP] in the subject accompanied by an appropriate subject briefly describing the contents.
- Feel free to use any programming language of your liking. However we strongly advise in favor of Python, due to the abundance of available tools (also note that Python3 comes with an excellent native support of UTF8 strings).
- Avoid using libraries that solve what we ask you to do (unless otherwise noted).
- Avoid building complex systems. The exercises are simple enough.
- Do not include any executable files in your submission, as this may cause the e-mail server to reject it.
- **In case of copying, all the participants (including the original solution) will get 0 points for the whole assignment. Note: it is rather easy to identify a copied solution. Plagiarism is also not tolerated.**
- **Please submit in your solutions necessary to support your claims. Failure to do so, might results in reduction of points in the relevant questions.**
- Each assignment has 10 points and perhaps some bonus points (usually 2 or 3). In order to qualify for the exams, you need to have 2/3 of the total points. For example, in case there are 12 assignments, you need to collect at least 80 out of the 120 points to be eligible for the exams. A person that gets 10 plus 2 bonus points in every exercise, needs to deliver only 7 assignments in order to be eligible for the exams, since 7*12=84.

[1] If you feel that these files are beyond reasonable size for an email submission and also reasonably convenient, please provide a means for us to access them online