Statistical Natural Language Processing (SS-2018)
Exercise Sheet 6

Submission Deadline: 01.06.2018, 23:59 (CEST)

n-gram LM & Kneser–Ney Smoothing

1) (2 points) Assume we have four words in our vocabulary:

$$a, b, c, d$$

Explain with a simple diagram of branching tree, how an n-gram model is represented(*Assume we consider up to trigrams*). Label each node with probabilities and mention at which level the n-grams can be read off from the branching tree. Now, for a generic case, assume that our vocabulary size is V and n in our n-gram is arbitrary. In such a case, how many parameters would we need to fully represent a trigram tree? And what problems could arise for larger vocabulary sizes (where $V >> n$)?

Hint: The probabilities in the branching tree can be represented in an abstract manner (i.e. you need not compute anything)

2) (1 points) Describe the idea behind Kneser–Ney smoothing technique. That is, explain where does it come from and how does it work?

Text Categorization

In this task, you will train a simple Naive Bayes classifier to perform *author identification*, using texts written by Henry James and Jack London.

Do some automated corpus analysis for the given files in *corpus/james/∗* and *corpus/london/∗* and

3) (1 points) List at least five features that could be useful for the task of *author identification* and justify the same (i.e. why the listed features are good *discriminators* for the task at hand).

4) (0.5 poitns) Now, produce a *histogram plot* of the frequencies of *10 most common words*, for each of the authors in the training set. (Note: don't forget to label the axes and indicate the scales)

5) (2.5 points) Now, using the files in *corpus/james/∗* and *corpus/london/∗* create a Naive Bayes classifier using the word frequencies (i.e. the unigram distribution, with floor discounting (a.k.a add-epsilon smoothing, Lidstone smoothing ... see Additive Smoothing)) as features. For the class probabilities, you may assume the document counts are representative. Classify the excerpts in the *corpus/test/∗* by author, and report the results.

6) (1 points) What do you think of using the unigram probabilities as features for classification task? Give an example of a case (not necessarily from the given texts) where unigram probabilities would *not* produce optimal results for classification.

7) (1 points) How can one show that a set of data points from two classes is **not** linearly separable? Also, explain how can one solve (i.e. put a linear decision boundary for) such not linearly separable data points?

8) (0.5 points) While *training* language models, what are we estimating? And what is/are the parameter(s)?

9) (0.5 points) What does *minimizing* the perplexity correspond to? Explain very briefly.

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_06_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*