SNLP 2016
Exercise 6

Submission date: 10.06.2016, 23:59

Pruning Language Models

1) (6 points) For this exercise, we will train a language model using absolute discount smoothing and pruning. You can complete this task by extending the code from your solution to Part 2 of Exercise 5.

- (3 points) Implement a method *prune* in the *discounting_model* class, which takes *epsilon* as an arugment, and creates and returns a new absolute discounting model with only the bigrams and unigrams with a probability of at least *epsilon* in the original model.
- (3 points) Using the text $/twain/pg119.txt$ for training, and $/twain/pg3176.txt$ for testing, compute the perplexity for the original absolute discounting model, and the pruned models for $epsilon = 10^-n$, for $n = 3, 4, 5, 6$. Use $d = 0.9$. Report both the perplexity and the number of parameters used for each model.

Classification

2) (3 points) In this task, you will train a simple Naive Bayes classifier to perform author identification, using texts by Mark Twain and Jane Austen.

- (2 points) Using the provided materials in */twain/* and */austen/*, create a Naive Bayes classifier using the word frequencies (ie. the unigram distribution, with floor discounting (also known as add-epsilon smoothing, Lidstone smoothing ... see `https://en.wikipedia.org/wiki/Additive_smoothing`)) as features. For the class probabilities, you may assume the document counts are representative. Classify the excerpts in the */test/* folder by author, and report the results.
- (1 point) What do you think of using the unigram probabilities as features for classification? Give an example (not necessarily from the given texts) of a case where unigram probabilities would not produce optimal results for classification.

3) (1 points) Name a drawback of the Naive Bayes classifier in general, and explain at least one reason why, despite this, it is a useful model in practice.

1 Submission Instructions: Read carefully

- You can form groups of maximum 3 people.
- Submit only 1 archive file in the ZIP format with name containing the MN of all the team members, e.g.:

Exercise_01_MatriculationNumber1_MatriculationNumber2_MatriculationNumber3.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 01

- Depending on your tutorial group, send your assignment to the corresponding tutor:

  - Sedigheh Eslami: *eslami@mpi-inf.mpg.de*
  - Naszdi Kata: *b.naszadi@gmail.com*
  - Stephanie Lund: *stflund@gmail.com*

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.**

- **Missing the deadline even for a few minutes, will result in 50% point reduction. Submission past the next tutorial, is not corrected, as the solutions will already be discussed.**

- **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.

- Attending the tutorial gives 2 points increase for the corresponding assignment.

- Exercise points (including any bonuses) guarantee only the admittance to the exam, however have no further effect on the final exam grade.

[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