Language Modelling and Basic Spell Checking

ICS2203 Natural Language Processing: Methods and Tools
Project Task 1

Claudia Borg

updated October 25, 2017

1 Problem Definition

Language Models are widely used in NLP — these models provide a probabilistic description of a language based on a given corpus. In this assignment, you will first build a Language Model based on the corpus provided from the MLRS corpus[1]. The initial tasks for this project are as follows:

1. Understand the formatting of the MLRS corpus.
2. Decide how you are going to process / store your data, and justify your reasons. Remember that although you are working with a 'small' corpus, some corpora could contain gigabytes of data.
3. Implement Unigram, Bigram and Trigram models — Vanilla Version with no smoothing.
4. Next, implement the models using Laplace smoothing (add-one).
5. Finally, implement a simple backoff model.

In order to test the use of these models, write the two following functions:

1. `GenerateText` which takes a string (sentence) as input, and will output the next most likely word.
2. `WrongWord` which takes a string (sentence) as input, and outputs which word is most likely to be misspelt.

[1] `http://mlrs.research.um.edu.mt/index.php?page=downloads`

2 Notes, Deliverables and Dates

Plagiarism You are free to discuss the problem with your colleagues, but do not provide each other with code or solutions. Remember to cite all your resources/websites used in your documentation.

Programming Language Choice of programming language is up to you. `Python`, `Java` and `C#` are accepted. If you would like to use any other programming language, please check with me.

Deliverables You should upload a `zip` file in the VLE which should include a `pdf` or `doc` file of your documentation, and your code. You do not need to include the corpus. In an effort to avoid unnecessary usage of paper, only an electronic submission is required.

Due date The deadline for this project is 23:55hr, Friday 15th December 2017.

Late submissions A 10% deduction from the total grade of this task per every late day will apply.

Grade This project constitutes 25% of your final grade for this study unit. The evaluation criteria are spilt as follows:

- Justification of processing and storage strategies: 10%
- Vanilla NGram models: 10%
- Laplace NGram models: 15%
- Backoff: 15%
- `GenerateText` function: 10%
- `WrongWord` function: 10%
- Demonstration: 10%
- Documentation: 20%

Demos Demonstrations will be held on Monday 18th December 2017, from 10a.m. till noon. You will be given the exact schedule closer to the date.

Questions Discussions can be done on the VLE through the Class Forum or you can send me an email.