February 6, 2015

Natural Language Models and Interfaces: Assignment Part A, Step 1

Cornelis Boon - 10561145, Markus Pfundstein - 10452397,
Thomas Meijers - 10647023

Abstract

[TODO]

1. Introduction

In this assignment we have built n-grams out of the Austen corpus which can be found `http://www-nlp.stanford.edu/fsnlp/statest/austen.txt`. We have done this in python(see Appendix for details on how to run and the results). From these n-grams, we will extract statistics such as the frequency of words and word sequences.

2. Problem

A probabilistic approach to language models can make use of n-grams. For this assignment we will create unigrams, bigrams and trigrams.

3. Approach

3.1. Step 1

The main approach to building n-grams out of the corpus is to split the corpus into separate words and then build sequences of length n. To count the frequencies of these n-grams, we use a Counter. Finally we order the results using an ordered dictionary and print the results as well as the sum of the frequencies, which is always equal to the total of n-grams in the corpus. (Not the total of unique n-grams)

3.2. Step 2

[TODO]

3.3. Step 3

[TODO]

4. Results

Please refer to the appendices B.1, B.2 and B.3 for the results of respectively step 1, 2 and 3.

5. Conclusion

Step 1 required a simple implementation of a n-gram counter combined with a few print statements. For $n = 3$ the Python script takes about three seconds to run and gives the correct output.

Appendices

A. Run instructions

A.1. Step 1

```
usage: a1-step1 [-h] [-corpus INPUT_FILE] [-n N] [-m M]

Assignment A, Step 1

optional arguments:
  -h, --help          show this help message and exit
  -corpus INPUT_FILE  Path to corpus file
  -n N                Length of word-sequences to process (n-grams)
  -m M                Number of n-grams to show in output
To exit: use 'exit', 'quit', or Ctrl-D.
An exception has occurred, use \%tb to see the full traceback.
```

A.2. Step 2

[TODO]

A.3. Step 3

[TODO]

B. Results

B.1. Step 1

B.1.1. 10 most frequent n-gram sequences

m'th most frequent n-gram	n=1	n=2	n=3
1	the 20829	of the 2507	I do not 378
2	to 20042	to be 2233	I am sure 366
3	and 18331	in the 1917	in the world 214
4	of 17949	I am 1366	she could not 202
5	a 11135	of her 1264	would have been 189
6	her 11007	to the 1142	I dare say 174
7	I 10381	it was 1010	a great deal 173
8	was 9409	had been 995	as soon as 173
9	in 9182	she had 978	it would be 171
10	it 7573	to her 964	could not be 155

B.1.2. Sum of all frequencies

- **For n = 1**
  Sum of frequencies = 617091
- **For n = 2**
  Sum of frequencies = 617090
- **For n = 3**
  Sum of frequencies = 617089

B.2. Results Step 2

[TODO]

B.3. Results Step 3

[TODO]