Intermediate report part a - step 1

Julian Main, Yorick de Boer, Amor Frans

February 4, 2016

Introduction

For assignment step 1 of part A we wrote a program which can extract ngrams from a large text file. An ngram is a word sequence of the length *n*. The program takes the length of the ngram *n*, the name of the corpus *corpus* and the top most frequent results *m* as input. The output of the program was the word sequences of length *n* , the *m* most frequent sequences, and the sum of the frequencies of all sequences.

Method

To solve this assignment we chose to program in python. The python script could be executed with the command:

```
python3.4 a1-step1.py [-h] [-n N] [-m M] corpus
```

Our program is made in the following structure:
- *Read input*
Using the python argparse module the input is read from the command line.

- *Find the* m *most frequent sequences of length* n
First we split the text in all possible sequences of length *n*. Then using the standard python module *collections* the program counted the *m* most frequent sequences. The *m* tuples of the structure(word sequence, frequency) were printed.

- *Get the sum of the frequencies*
Finally the sum of all the sequences are printed.

Results

The results for n=1, n=2 and n=3 for the 10 most frequent ngrams are shown in table q. Also the sum of the frequencies is given in the table.

n	1		2		3	
	word sequence	**frequency**	**word sequence**	**frequency**	**word sequence**	**frequency**
1.	'the'	20829	'of the'	2507	'I do not'	378
2.	'to'	20042	'to be'	2235	'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'	1268	'would have been'	189
6.	'her'	11020	'to the'	1142	'I dare say'	174
7.	'I'	10396	'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'	7575	'to her'	965	'could not be'	155
sum of frequencies	135868		14383		2195	

Table 1: Results of the ngram extraction with m=10

Discussion and conclusion

The program runs in a few seconds which is impressive for such a large text. Further, results seem good if we look at the context of the text. The results show that 'her 'she had', 'of her', 'to her' and 'she could not' are appearing in table of the 10 most frequent word sequences. In contrary to that the word 'him' or 'her' etc. are not appearing in the table. Therefore the results are showing that is very likely that the text is about a girl or a woman. The corpus given here is Emma by Jane Austen and indeed the main character is a girl/woman. Therefore we can conclude that the program works well.