Natuurlijke Taalmodellen & Interfaces
Tejaswini Deoskar
Part A (Step 2)

Lucas van Berkel, 10747958
Ruben Gerritse, 10760326
Gideon Mooijen, 10686290

Report

Our second program has a few steps. They are described generally in the following paragraphs. A more technical description can be found in the code as comments. The required results can be found as tables on the bottom of this report.

Read the text-file and create the so-called *'sentence list'*, a list that contains lists, which represents all paragraphs of the corpus. The second step counts the amount of times that a sequence of n words is contained in the corpus. If a sequence is encountered that is not yet present in the dictionary, it is clearly the first occurrence of that sequence and it is added to the dictionary. If the sequence was already present it is merely a new occurrence and the counter of said sequence is incremented by one. What follows is a little procedure that simply orders the *'sequence dictionary'*. After all occurrences are counted, they are rearranged from high to low and the desired amount of results are printed (along with the corresponding frequencies).

Then, the program reads the conditional probability file. It calculates the posterior probability and is printed in the terminal. To achieve this, it needs the dictionary containing the sequences of size n (and their corresponding frequencies), and the dictionary containing the sequences of size n-1 (and their corresponding frequencies).

Ultimately, we were able to prevent the program from rounding off very small numbers to 0, and did manage for the program to accept a list of sentences as input.

Answers

1. 10 most frequent bigrams:

	sequence	**frequency**
1	of the	2507
2	to be	2232
3	in the	1917
4	I am	1365
5	of her	1264
6	to the	1142
7	it was	1010
8	had been	995
9	she had	978
10	to her	964

4. Set A:

#	permutation	probability
1	I do not know what your opinion may be	1.0154707761629874e-12
2	I do not know what may be your opinion	1.2486963401507535e-13

Set B:

#	permutation	probability
1	I do not know	4.744737109484349e-05
2	I know not do	2.1618102147881128e-07