Answers Assignment A (Step 1)

Name: Danny Dijkzeul, 10554386
Kaj Meijer, 10509534
Philip Bouman, 10668667

Corpus: AUSTEN TRAIN corpus

1. Ten most frequent sequences:

	$n = 1$	$n = 2$	$n = 3$
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)	as soon as, (173)
8	was, (9409)	had been, (995)	a great deal, (173)
9	in, (9182)	she had, (978)	it would be, (171)
10	it, (7575)	to her, (965)	could not be, (155)

2. Sum of all frequencies of all sequences:

$n = 1$	$n = 2$	$n = 3$
620918	620917	620916

3. The program can be run by entering the following command:

   ```
   ./assignmentA1.py -corpus [path] -n [value] -m [value]
   ```

   Replace path with the location of the corpus (saved as .txt-file).
   Replace the first value with the order of n-gram (n = 1 is unigram, n = 2 is bigram, etc.).
   Replace the second value with the number of most frequent sequences to print.

   See code for the working of the program.