Homework 1

Problem 1

Corpus	Doc #	Words	Class
Training	1	fun, couple, love, love	comedy
	2	fast, furious, shoot	action
	3	couple, fly, fast, fun, fun	comedy
	4	furious, shoot, shoot, fun	action
	5	fly, fast, shoot, love	action
Test	6	fast, couple, shoot, fly	action

Priors

$$P(comedy) = \frac{2}{5}$$

$$P(action) = \frac{3}{5}$$

Conditional Probabilities

$$P(fast|comedy) = \frac{1+1}{9+8} = \frac{2}{17}$$

$$P(couple|comedy) = \frac{2+1}{9+8} = \frac{3}{17}$$

$$P(shoot|comedy) = \frac{0+1}{9+8} = \frac{1}{17}$$

$$P(fly|comedy) = \frac{1+1}{9+8} = \frac{2}{17}$$

$$P(fast|action) = \frac{2+1}{11+8} = \frac{3}{19}$$

$$P(couple|action) = \frac{0+1}{11+8} = \frac{1}{19}$$

$$P(shoot|action) = \frac{4+1}{11\ +\ 8} = \frac{5}{19}$$

$$P(fly|action) = \frac{1+1}{11\ +\ 8} = \frac{2}{19}$$

Choosing a class:

$$P(comedy|D) = \frac{2}{5} * \frac{2}{17} * \frac{3}{17} * \frac{1}{17} * \frac{2}{17} = \frac{24}{417605} = 0.000057$$

$$P(action|D) = \frac{3}{5} * \frac{3}{19} * \frac{1}{19} * \frac{5}{19} * \frac{2}{19} = \frac{18}{130321} = 0.00013$$

The most likely class for the document D is "action".

Problem 2

2.1

refuse - 6
believe – 5
bank – 10
justice – 4
is – 13
bankrupt – 1

refuse - 6
believe – 5
there – 1
are – 13
insufficient - 1
funds - 4
great – 6
vaults – 4
opportunity – 1
nation – 4

so – 10
have – 19
come – 21
cash - 1

check – 13
check - 13
will – 3
give - 44
demand - 5
riches - 1
freedom - 2
security - 9
justice – 4

2.2

First sentence: 15 600

Second sentence: 149 760

Third sentence: 32 043 211 200

Problem 3

The ZIP file attached to the homework contains a folder called Language Model. Inside this folder, is the solution of Problem 3. Please find a README.txt to read a description of how to run the code.

1. See attached code.

2. See attached code.

3. See attached code

4. Comparing the two smoothing methods

I developed my code by creating many small train sets and test sets and testing the code as I progressed through the work. Largely based on this strategy, the fact that my tests were many and exhaustive, and on my custom parser, I managed to get good results from each of the smoothing methods.

My perplexity for add-1 smoothing was 128.51. My perplexity if interpolation was used was 88.41. While add-1 performed well compared to the expected values for

perplexity we were provided, interpolation certainly resulted in better distribution of the data.

I think the reason behind this is the fact that interpolation allowed me to use various sources of information. Because it includes considering probabilities of a word as a unigram, bigram, and trigram, I was able to use various sources of authoritative data. In this way if in a training corpus, I have not seen the combination X Y Z, I can still get reliable information if I have seen Y Z in case that X was a very rare or unknown word. And if I don't have any good data for Y Z, I can at least use the unigram Z. If I don't have information about this even, I could rely on the UNK class.

In add-1 smoothing, I didn't not have this variety of options. Yes, add-1 smoothing allowed me to eliminate 0 probabilities and I could still use the UNK class, but it did not give me the insight that a bigram in interpolation would. Alongside with this, add-1 smoothing can unfairly change the distribution of the data. So if some combination X Y Z was likely to happen, after applying add-1 smoothing, X Y Z is now less likely because parts of the probability body have been moved to previous places that earlier had 0 probability. While this solves the problem with 0 probabilities that prevents Language Models from generating correct probabilities for sentences, it also leads to less accurate predictions for more likely sentences. The sentence X Y Z, unfairly, has a smaller probability after add-1 smoothing compared to before the smoothing was applied.

All of this is logical for add-1 smoothing and as it is the simplest option, it performs as well as we could expect it. But because interpolation uses a more flexible method that includes various sources of insight and because interpolation does not lead to that big and unfair redistribution of the probability body, interpolation leads to better perplexity and is this a better smoothing method.