Name: Nooreldean Koteb

GMUID: Nkoteb – G01085380

HW 2 Write up

Part 1:

Below are the values of the unsmoothed and add-1 smoothed table for Part 1 a & b.

```
Unigram:
p(free): 0.000152748726579
p(market): 0.001808262381697
p(language): 0.000008829406161

Bigram:
p(market | free): 0.225433526011561
p(shortage | market): 0.008789062500000
p(reagan | president): 0.125806451612903
p(administration | carter): 0.066666666666667
p(profit | net): 0.117956588467441
p(loss | net): 0.165753124314843
p(language | programming): 0.333333333333333
p(language | endangered): 0.000000000000000

Trigram:
p(of | the language): 0.500000000000000
p(accord | the paris): 0.328671328671329
```

```
add-1 smoothing applied:

Unigram:
p(free): 0.000148084602265
p(market): 0.001743823850812
p(language): 0.000009361670258

Bigram:
p(market | free): 0.000939011221184
p(shortage | market): 0.000427225507611
p(reagan | president): 0.001835288651411
p(administration | carter): 0.000047125353440
p(profit | net): 0.011471502149576
p(loss | net): 0.016111182054229
p(language | programming): 0.000047138682002
p(language | endangered): 0.000023569341001

Trigram:
p(of | the language): 0.000070706356501
p(accord | the paris): 0.001127607592558
```

Below are the values of the perplexities I got for Part 1 c.

```
n-gram 1: 348505.5433189364
n-gram 2: 449010.22406680416
n-gram 3: 303293.5415853823
n-gram 4: 209624.86943659984
n-gram 5: 147960.07113139087
n-gram 6: 106466.20641090958
```

I was unable to get to Part 1 d.

Part 2:

Below are the values of the cosine similarity between the two words in the table for Part 2-1.

```
Cosine Similarity:
horseradish | spinach --> 0.43284
lingonberry | strawberries --> 0.46530
pikachu | charizard --> 0.53346
charizard | charmander --> 0.23467
math | algorithm --> 0.61035
```

analogy-predictions.txt is in this file with the predictions for Part 2-2.

Part2-2 (a):

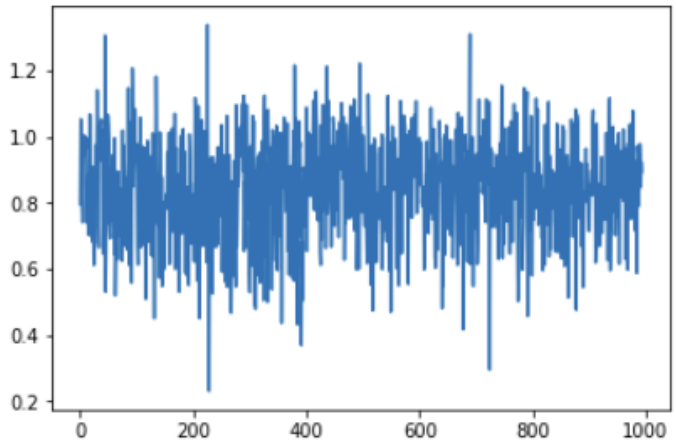
For this part I did element-wise operations on 3 of the 4-word vectors than did a cosine similarity between that new vector and the 4th word.

Operation: ((Word2 – Word1) + Word3) = New vector

Part2-2 (b):

I then decided to make some statistics on all the cosine similarity values as shown below. Given that Eve changed more than 20% but not more than 50% of the data I felt that 0.845 to 0.855 was a safe threshold. I then decided to add about 25% of the STD to the median and call that my final threshold for deciding if an analogy was right or wrong.

```
Max: 1.339512646654186
Min: 0.23058255178455855
Mean: 0.8453156937548844
Median: 0.8569925520935333
STD: 0.1570998718024798
Range of possible Wrong: [0.9798133949308251, 0.8634270821335785]
Range of possible Wrong: 0.11638631279724654
```



References:

For reading the code on part 2 I took code from piazza by Samuel Blouir because it was much faster at loading and reloading data. https://piazza.com/class/kkaenv2ty7x4tr?cid=88