

3.3 Part A, all words start with A and probability:

Total is 81764926

1, word is: A

unigram probability is 0.01840724469071

2, word is: AND

unigram probability is 0.01786323392502

3, word is: AT

unigram probability is 0.00431297400061

4, word is: AS

unigram probability is 0.00399179716741

5, word is: AN

unigram probability is 0.00299925667394

6, word is: ARE

unigram probability is 0.00298969267091

7, word is: ABOUT

unigram probability is 0.00192561783765

8, word is: AFTER

unigram probability is 0.00134656759795

9, word is: ALSO

unigram probability is 0.00131001158125

10, word is: ALL

unigram probability is 0.00118181480406

11, word is: A.

unigram probability is 0.00102561090803

12, word is: ANY

unigram probability is 0.00063186016948

13, word is: AMERICAN

unigram probability is 0.00061209619391

14, word is: AGAINST

unigram probability is 0.00059596458266

15, word is: ANOTHER

unigram probability is 0.00042838661653

16, word is: AMONG

unigram probability is 0.00037429251755

17, word is: AGO

unigram probability is 0.00035657098253

18, word is: ACCORDING

unigram probability is 0.00034754510754

19, word is: AIR

unigram probability is 0.00031100132103

20, word is: ADMINISTRATION

unigram probability is 0.00029151863967

21, word is: AGENCY

unigram probability is 0.00027965536225

22, word is: AROUND

unigram probability is 0.00027685465037

23, word is: AGREEMENT

unigram probability is 0.00026278994003

24, word is: AVERAGE

unigram probability is 0.00025907196443

25, word is: ASKED
unigram probability is 0.00025822808181

26, word is: ALREADY
unigram probability is 0.00024907990499

27, word is: AREA
unigram probability is 0.00023108930595

28, word is: ANALYSTS
unigram probability is 0.00022603824041

29, word is: ANNOUNCED
unigram probability is 0.00022715118705

30, word is: ADDED
unigram probability is 0.00022121954834

31, word is: ALTHOUGH
unigram probability is 0.00021426057427

32, word is: AGREED
unigram probability is 0.00021177784714

33, word is: APRIL
unigram probability is 0.00020669009105

34, word is: AWAY
unigram probability is 0.00020205485173

3.3 Part B, ten most likely words to follow "THE".
"THE" is at index 3 of the file

THE
Count of THE is : 3855375
1 : the vocabulary is <UNK>
the bigram probability 0.61501981000551

2 : the vocabulary is U.
the bigram probability 0.01337249943261

3 : the vocabulary is FIRST
the bigram probability 0.01172026067503

4 : the vocabulary is COMPANY
the bigram probability 0.01165878805564

5 : the vocabulary is NEW
the bigram probability 0.00945148007652

6 : the vocabulary is UNITED
the bigram probability 0.00867230814123

7 : the vocabulary is GOVERNMENT
the bigram probability 0.00680348863600

8 : the vocabulary is NINETEEN
the bigram probability 0.00665071491100

9 : the vocabulary is SAME
the bigram probability 0.00628706675745

10 : the vocabulary is TWO
the bigram probability 0.00616074960283

3.3 Part c, comparing Log-likelihood of unigram and bigram
The unigram Log-likelihood is -64.50944034364879
The bigram Log-likelihood is -40.91813213378977

Therefore, the bigram model yields the highest log-likelihood

3.3 Part d, pair that are not observed
The unigram Log-likelihood is -44.29193447313260

not observed in training corpus: curr: OFFICIALS, previous SIXTEEN
not observed in training corpus: curr: FIRE, previous SOLD
Some Pairs not find in traning corpus, Log(0) undefined, can't calculate bigram prob
ability

3.3 Part e, mixture model

Lambda: 0.00 total is -44.29193447313261
Lambda: 0.01 total is -44.16477944425115
Lambda: 0.02 total is -44.05295254780151
Lambda: 0.03 total is -43.95346021097132
Lambda: 0.04 total is -43.86413813195733
Lambda: 0.05 total is -43.78337001938518
Lambda: 0.06 total is -43.70991643503021
Lambda: 0.07 total is -43.64280574287029
Lambda: 0.08 total is -43.58126193439200
Lambda: 0.09 total is -43.52465530426293
Lambda: 0.10 total is -43.47246780585678
Lambda: 0.11 total is -43.42426813430323
Lambda: 0.12 total is -43.37969343104693
Lambda: 0.13 total is -43.33843560311345
Lambda: 0.14 total is -43.30023092618666
Lambda: 0.15 total is -43.26485202821000
Lambda: 0.16 total is -43.23210162766372
Lambda: 0.17 total is -43.20180758478801
Lambda: 0.18 total is -43.17381894871572
Lambda: 0.19 total is -43.14800276949136
Lambda: 0.20 total is -43.12424150428679
Lambda: 0.21 total is -43.10243089009294
Lambda: 0.22 total is -43.08247818620268
Lambda: 0.23 total is -43.06430071250665
Lambda: 0.24 total is -43.04782462643607
Lambda: 0.25 total is -43.03298389397349
Lambda: 0.26 total is -43.01971941967341
Lambda: 0.27 total is -43.00797830790595
Lambda: 0.28 total is -42.99771323314158
Lambda: 0.29 total is -42.98888190145040
Lambda: 0.30 total is -42.98144658880390
Lambda: 0.31 total is -42.97537374446272
Lambda: 0.32 total is -42.97063364987885
Lambda: 0.33 total is -42.96720012526013
Lambda: 0.34 total is -42.96505027733244
Lambda: 0.35 total is -42.96416428296298
Lambda: 0.36 total is -42.96452520423230
Lambda: 0.37 total is -42.96611883130494
Lambda: 0.38 total is -42.96893355008312
Lambda: 0.39 total is -42.97296023216013
Lambda: 0.40 total is -42.97819214504175
Lambda: 0.41 total is -42.98462488099085
Lambda: 0.42 total is -42.99225630318629
Lambda: 0.43 total is -43.00108650818365
Lambda: 0.44 total is -43.01111780393088
Lambda: 0.45 total is -43.02235470283430
Lambda: 0.46 total is -43.03480392959781
Lambda: 0.47 total is -43.04847444377423
Lambda: 0.48 total is -43.06337747718079
Lambda: 0.49 total is -43.07952658654352
Lambda: 0.50 total is -43.09693772195611
Lambda: 0.51 total is -43.11562931196909
Lambda: 0.52 total is -43.13562236637604
Lambda: 0.53 total is -43.15694059803586
Lambda: 0.54 total is -43.17961056537597
Lambda: 0.55 total is -43.20366183756634
Lambda: 0.56 total is -43.22912718474976
Lambda: 0.57 total is -43.25604279617114
Lambda: 0.58 total is -43.28444852958410
Lambda: 0.59 total is -43.31438819594176
Lambda: 0.60 total is -43.34590988412617
Lambda: 0.61 total is -43.37906633136214
Lambda: 0.62 total is -43.41391534603244

```
Lambda: 0.63 total is -43.45052029090613
Lambda: 0.64 total is -43.48895063636574
Lambda: 0.65 total is -43.52928259514337
Lambda: 0.66 total is -43.57159985244380
Lambda: 0.67 total is -43.61599440826144
Lambda: 0.68 total is -43.66256755234718
Lambda: 0.69 total is -43.71143099685251
Lambda: 0.70 total is -43.76270819744925
Lambda: 0.71 total is -43.81653590105604
Lambda: 0.72 total is -43.87306596769493
Lambda: 0.73 total is -43.93246752612482
Lambda: 0.74 total is -43.99492953868475
Lambda: 0.75 total is -44.06066387152073
Lambda: 0.76 total is -44.12990899388962
Lambda: 0.77 total is -44.20293446713083
Lambda: 0.78 total is -44.28004643392828
Lambda: 0.79 total is -44.36159438715697
Lambda: 0.80 total is -44.44797959311978
Lambda: 0.81 total is -44.53966567873692
Lambda: 0.82 total is -44.63719208538813
Lambda: 0.83 total is -44.74119137371222
Lambda: 0.84 total is -44.85241178211638
Lambda: 0.85 total is -44.97174707671224
Lambda: 0.86 total is -45.10027671668379
Lambda: 0.87 total is -45.23932093174328
Lambda: 0.88 total is -45.39051789123121
Lambda: 0.89 total is -45.55593453189626
Lambda: 0.90 total is -45.73823035873303
Lambda: 0.91 total is -45.94090784433090
Lambda: 0.92 total is -46.16871093203297
Lambda: 0.93 total is -46.42829103975528
Lambda: 0.94 total is -46.72938994160976
Lambda: 0.95 total is -47.08711090153921
Lambda: 0.96 total is -47.52675641131557
Lambda: 0.97 total is -48.09575995494593
Lambda: 0.98 total is -48.90061116859637
Lambda: 0.99 total is -50.28110891199336
The best Lambda value is: 0.35
```