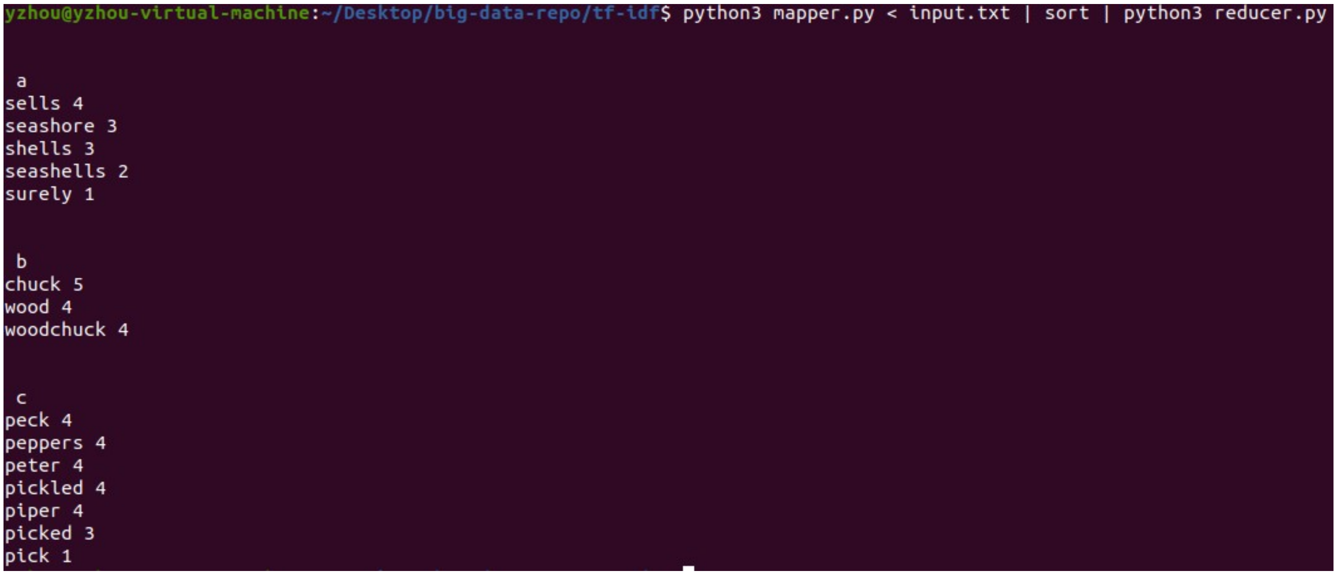
Yi Zhou
Quiz 5
Mar. 17th. 2022



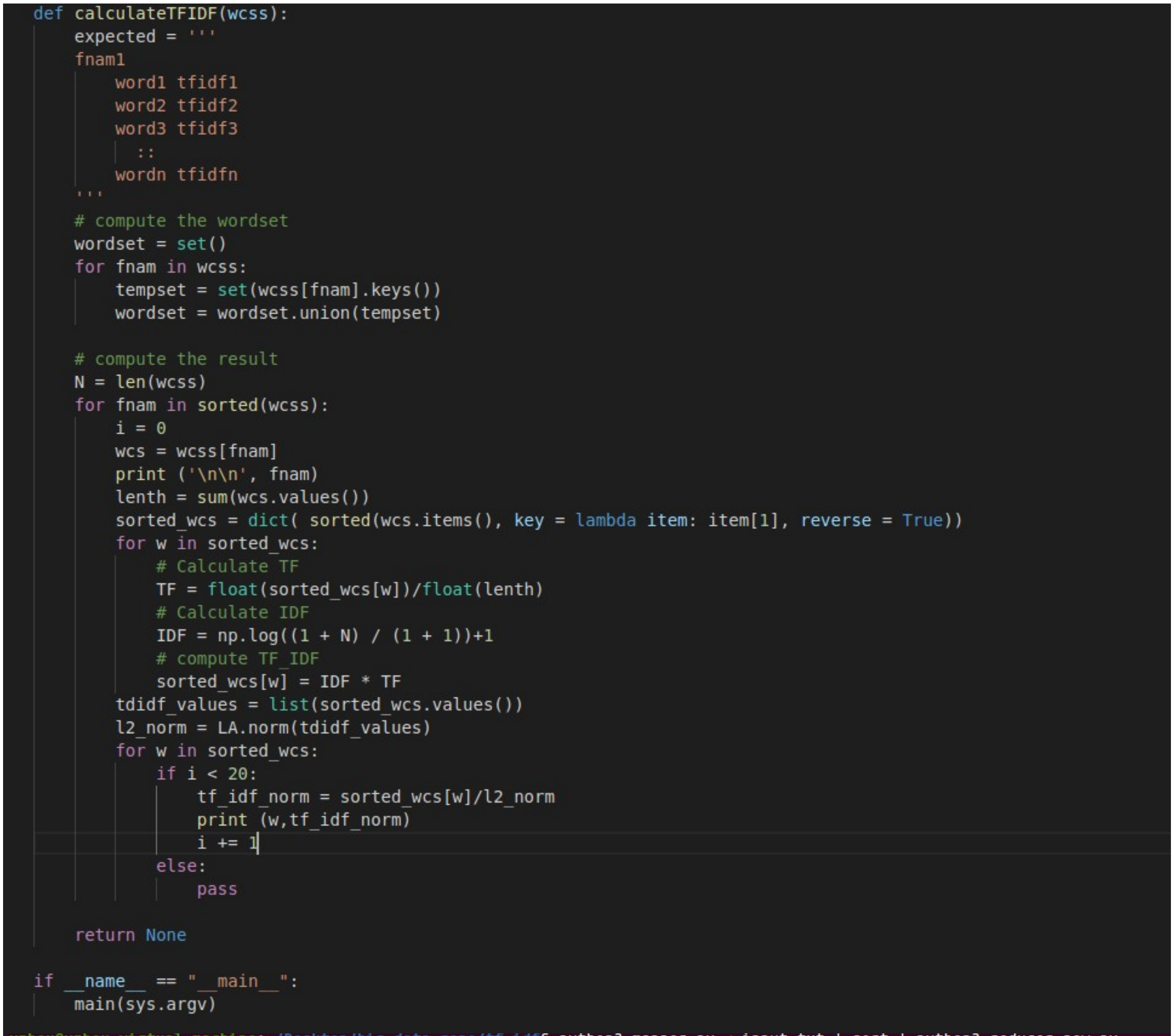
```
yzhou@yzhou-virtual-machine:~/Desktop/big-data-repo/tf-idf$ python3 mapper.py < input.txt | sort | python3 reducer.py


 a
sells 4
seashore 3
shells 3
seashells 2
surely 1


 b
chuck 5
wood 4
woodchuck 4


 c
peck 4
peppers 4
peter 4
pickled 4
piper 4
picked 3
pick 1
```

Answer for question1

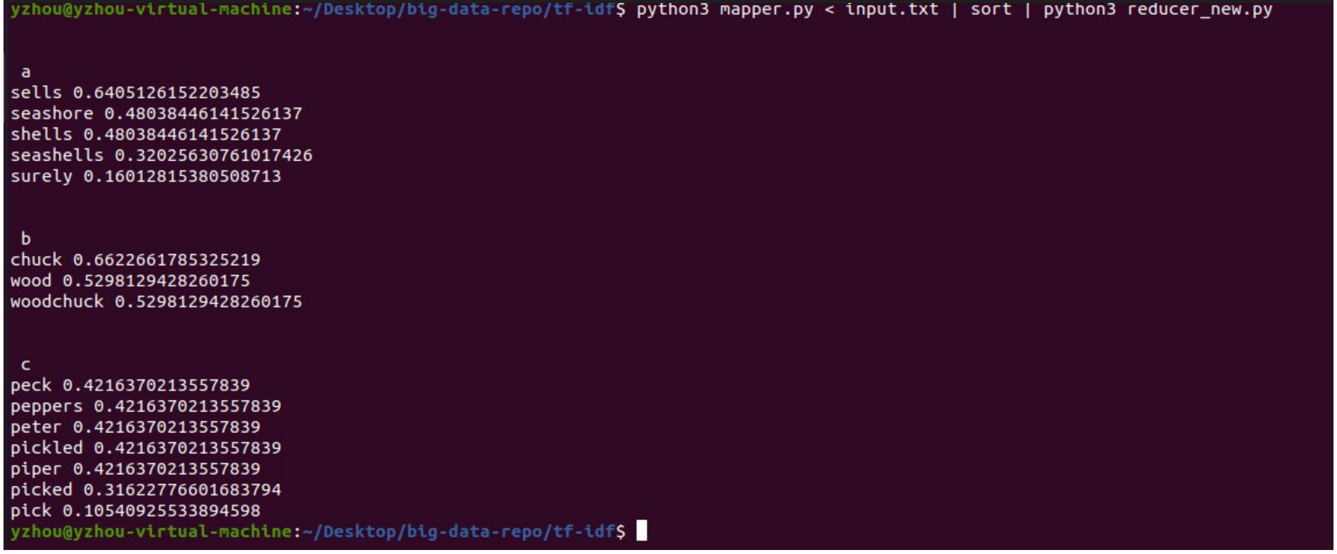


```
def calculateTFIDF(wcss):
    expected = '''
    fnam1
        word1 tfidf1
        word2 tfidf2
        word3 tfidf3
          ::
        wordn tfidfn
    '''
    # compute the wordset
    wordset = set()
    for fnam in wcss:
        tempset = set(wcss[fnam].keys())
        wordset = wordset.union(tempset)

    # compute the result
    N = len(wcss)
    for fnam in sorted(wcss):
        i = 0
        wcs = wcss[fnam]
        print ('\n\n', fnam)
        lenth = sum(wcs.values())
        sorted_wcs = dict( sorted(wcs.items(), key = lambda item: item[1], reverse = True))
        for w in sorted_wcs:
            # Calculate TF
            TF = float(sorted_wcs[w])/float(lenth)
            # Calculate IDF
            IDF = np.log((1 + N) / (1 + 1))+1
            # compute TF_IDF
            sorted_wcs[w] = IDF * TF
        tdidf_values = list(sorted_wcs.values())
        l2_norm = LA.norm(tdidf_values)
        for w in sorted_wcs:
            if i < 20:
                tf_idf_norm = sorted_wcs[w]/l2_norm
                print (w,tf_idf_norm)
                i += 1
            else:
                pass

    return None

if __name__ == "__main__":
    main(sys.argv)
```



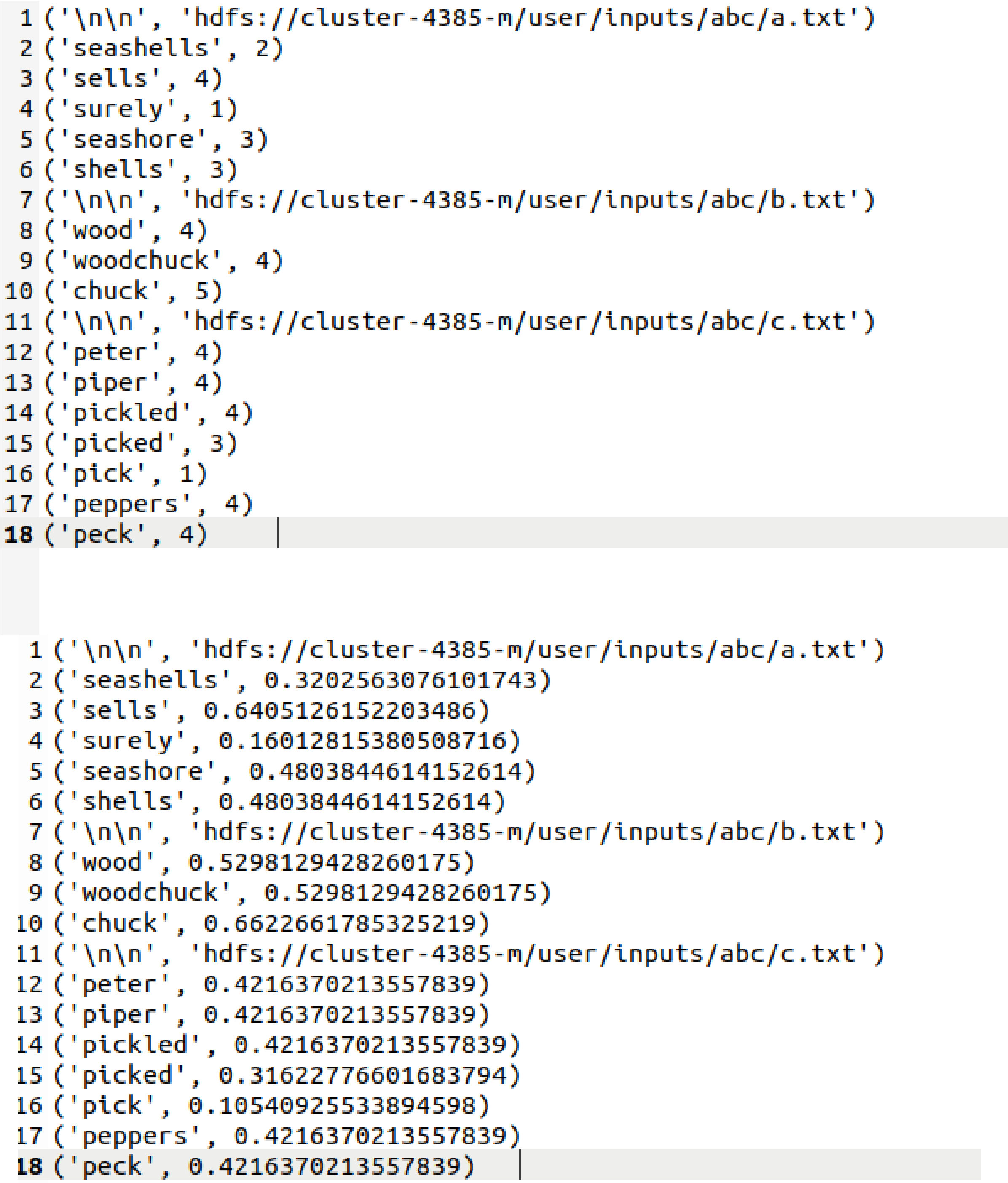
```
yzhou@yzhou-virtual-machine:~/Desktop/big-data-repo/tf-idf$ python3 mapper.py < input.txt | sort | python3 reducer_new.py


 a
sells 0.6405126152203485
seashore 0.48038446141526137
shells 0.48038446141526137
seashells 0.32025630761017426
surely 0.16012815380508713


 b
chuck 0.6622661785325219
wood 0.5298129428260175
woodchuck 0.5298129428260175


 c
peck 0.4216370213557839
peppers 0.4216370213557839
peter 0.4216370213557839
pickled 0.4216370213557839
piper 0.4216370213557839
picked 0.31622776601683794
pick 0.10540925533894598
yzhou@yzhou-virtual-machine:~/Desktop/big-data-repo/tf-idf$
```

Answer for question 2



```
('\n\n', 'hdfs://cluster-4385-m/user/inputs/abc/a.txt')
('seashells', 2)
('sells', 4)
('surely', 1)
('seashore', 3)
('shells', 3)
('\n\n', 'hdfs://cluster-4385-m/user/inputs/abc/b.txt')
('wood', 4)
('woodchuck', 4)
('chuck', 5)
('\n\n', 'hdfs://cluster-4385-m/user/inputs/abc/c.txt')
('peter', 4)
('piper', 4)
('pickled', 4)
('picked', 3)
('pick', 1)
('peppers', 4)
('peck', 4)
```

```
('\n\n', 'hdfs://cluster-4385-m/user/inputs/abc/a.txt')
('seashells', 0.3202563076101743)
('sells', 0.6405126152203486)
('surely', 0.16012815380508716)
('seashore', 0.4803844614152614)
('shells', 0.4803844614152614)
('\n\n', 'hdfs://cluster-4385-m/user/inputs/abc/b.txt')
('wood', 0.5298129428260175)
('woodchuck', 0.5298129428260175)
('chuck', 0.6622661785325219)
('\n\n', 'hdfs://cluster-4385-m/user/inputs/abc/c.txt')
('peter', 0.4216370213557839)
('piper', 0.4216370213557839)
('pickled', 0.4216370213557839)
('picked', 0.31622776601683794)
('pick', 0.10540925533894598)
('peppers', 0.4216370213557839)
('peck', 0.4216370213557839)
```

Answer for question 3

Question 5:

```
 hdfs://cluster-4385-m/user/inputs/speeches/1981.txt
government 0.35183959770683404
americans 0.19790977371009413
people 0.19790977371009413
freedom 0.17591979885341702
time 0.17591979885341702
america 0.13193984914006276
man 0.13193984914006276
nation 0.13193984914006276
called 0.10994987428338564
today 0.10994987428338564
work 0.10994987428338564
national 0.10994987428338564
strength 0.10994987428338564
president 0.10994987428338564
heroes 0.10994987428338564
god 0.08795989942670851
great 0.08795989942670851
price 0.08795989942670851
special 0.08795989942670851
history 0.08795989942670851


 hdfs://cluster-4385-m/user/inputs/speeches/1985.txt
people 0.29418525254116873
freedom 0.2390255176896996
government 0.2390255176896996
time 0.18386578283823043
human 0.16547920455440743
history 0.16547920455440743
god 0.14709262627058436
peace 0.14709262627058436
america 0.12870604798676133
american 0.12870604798676133
years 0.12870604798676133
americans 0.1103194697029383
progress 0.1103194697029383
earth 0.1103194697029383
weapons 0.1103194697029383
nuclear 0.1103194697029383
national 0.1103194697029383
citizens 0.1103194697029383
heart 0.09193289141911522
union 0.09193289141911522
```

```
 hdfs://cluster-4385-m/user/inputs/speeches/1989.txt
great 0.2131522781597438
nation 0.2131522781597438
free 0.19183705034376938
good 0.17052182252779505
hand 0.17052182252779505
things 0.17052182252779505
friends 0.17052182252779505
america 0.14920659471182066
today 0.14920659471182066
day 0.14920659471182066
work 0.14920659471182066
people 0.14920659471182066
time 0.14920659471182066
dont 0.12789136689584626
president 0.12789136689584626
freedom 0.12789136689584626
strong 0.12789136689584626
breeze 0.1065761390798719
love 0.1065761390798719
word 0.1065761390798719


 hdfs://cluster-4385-m/user/inputs/speeches/1993.txt
america 0.3774854303921253
people 0.30198834431370025
americans 0.2516569535947502
today 0.2516569535947502
change 0.20132556287580017
time 0.17615986751632515
work 0.15099417215685013
fellow 0.1258284767973751
nation 0.1258284767973751
idea 0.1258284767973751
service 0.10066278143790008
season 0.10066278143790008
americas 0.10066278143790008
millions 0.10066278143790008
generation 0.10066278143790008
renewal 0.10066278143790008
government 0.10066278143790008
american 0.10066278143790008
democracy 0.10066278143790008
celebrate 0.07549708607842506


 hdfs://cluster-4385-m/user/inputs/speeches/1997.txt
century 0.37562411321267386
nation 0.244155673588238
```

```
time 0.22537446792760432
america 0.20659326226697064
land 0.20659326226697064
people 0.18781205660633693
promise 0.18781205660633693
government 0.18781205660633693
american 0.16903085094570325
work 0.15024964528506954
citizens 0.15024964528506954
fellow 0.13146843962443586
children 0.11268723396380216
great 0.11268723396380216
americans 0.11268723396380216
lives 0.11268723396380216
strong 0.09390602830316847
power 0.09390602830316847
today 0.09390602830316847
worlds 0.09390602830316847


 hdfs://cluster-4385-m/user/inputs/speeches/2001.txt
story 0.23292312361320136
country 0.23292312361320136
america 0.20704277654506792
citizens 0.20704277654506792
nation 0.20704277654506792
courage 0.12940173534066746
nations 0.12940173534066746
americans 0.12940173534066746
common 0.12940173534066746
freedom 0.12940173534066746
promise 0.12940173534066746
president 0.12940173534066746
work 0.10352138827253396
duty 0.10352138827253396
purpose 0.10352138827253396
time 0.10352138827253396
government 0.10352138827253396
character 0.10352138827253396
justice 0.10352138827253396
public 0.10352138827253396


 hdfs://cluster-4385-m/user/inputs/speeches/2005.txt
freedom 0.4701108840278759
liberty 0.2820665304167255
america 0.22565322433338036
nation 0.15043548288892025
americans 0.15043548288892025
```

```
country 0.15043548288892025
americas 0.15043548288892025
history 0.13163104752780522
time 0.13163104752780522
free 0.13163104752780522
justice 0.11282661216669018
citizens 0.11282661216669018
hope 0.11282661216669018
human 0.11282661216669018
day 0.11282661216669018
people 0.11282661216669018
work 0.11282661216669018
states 0.09402217680557515
united 0.09402217680557515
rights 0.09402217680557515


 hdfs://cluster-4385-m/user/inputs/speeches/2009.txt
nation 0.2825289656088066
america 0.1883526437392044
people 0.16480856327180388
today 0.16480856327180388
common 0.1412644828044033
work 0.1412644828044033
time 0.1412644828044033
day 0.11772040233700276
spirit 0.11772040233700276
generation 0.11772040233700276
meet 0.0941763218696022
women 0.0941763218696022
men 0.0941763218696022
greater 0.0941763218696022
power 0.0941763218696022
seek 0.0941763218696022
peace 0.0941763218696022
long 0.0941763218696022
crisis 0.0941763218696022
god 0.07063224140220165


 hdfs://cluster-4385-m/user/inputs/speeches/2013.txt
people 0.2634022284358166
time 0.21551091417475907
america 0.1436739427831727
freedom 0.1436739427831727
citizens 0.1436739427831727
country 0.1436739427831727
journey 0.1436739427831727
nation 0.1436739427831727
```

```
god 0.11972828565264391
years 0.11972828565264391
american 0.11972828565264391
liberty 0.11972828565264391
equal 0.11972828565264391
complete 0.11972828565264391
creed 0.11972828565264391
requires 0.11972828565264391
future 0.09578262852211514
today 0.09578262852211514
oath 0.09578262852211514
americans 0.09578262852211514


 hdfs://cluster-4385-m/user/inputs/speeches/2017.txt
america 0.47075295928987204
american 0.27254118695729435
people 0.24776471541572215
country 0.22298824387414992
great 0.14865882924943327
nation 0.14865882924943327
dreams 0.12388235770786107
protected 0.12388235770786107
nations 0.12388235770786107
president 0.12388235770786107
god 0.09910588616628886
americans 0.09910588616628886
heart 0.09910588616628886
wealth 0.09910588616628886
jobs 0.09910588616628886
bring 0.09910588616628886
day 0.09910588616628886
power 0.09910588616628886
today 0.09910588616628886
citizens 0.09910588616628886
```