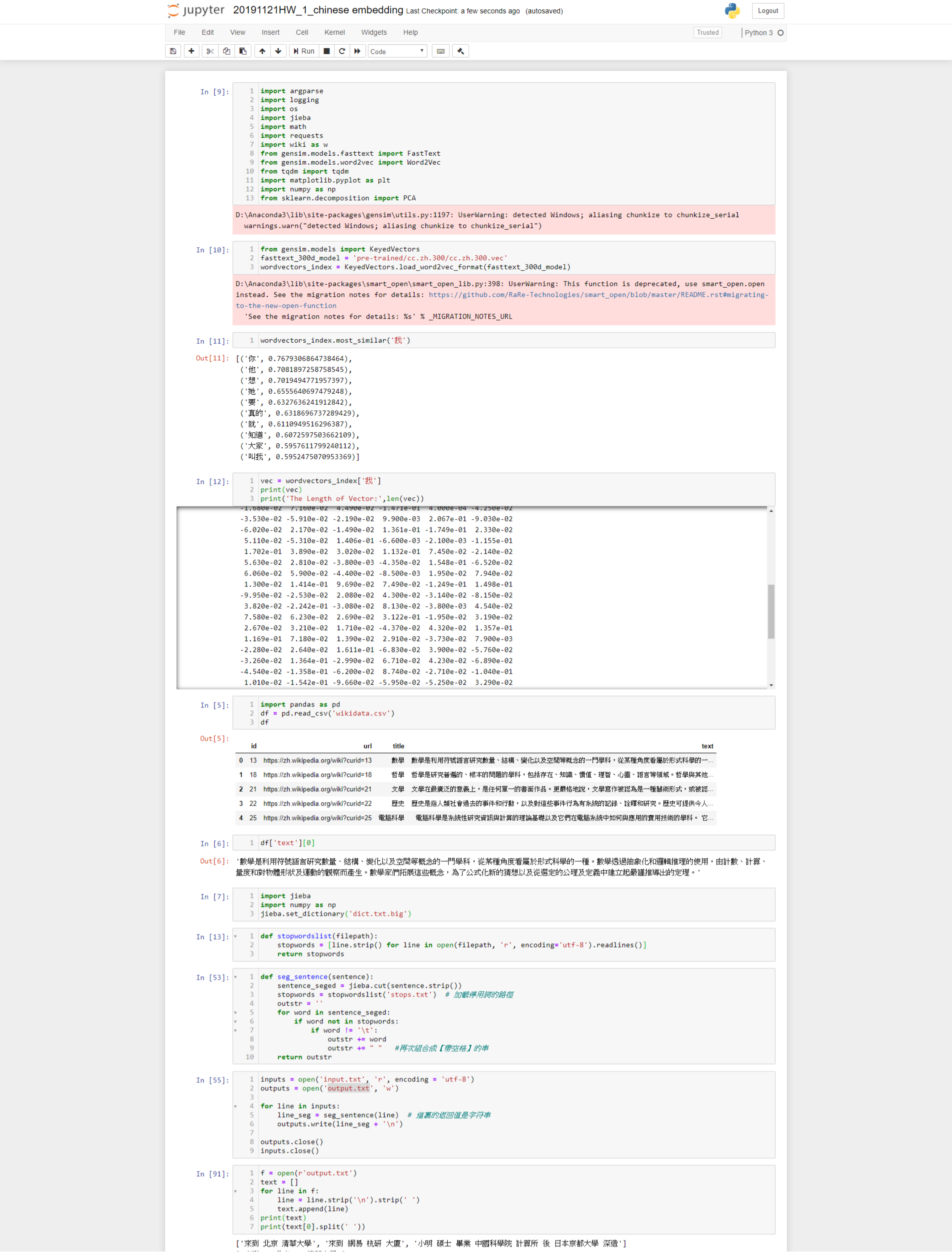

jupyter
20191121HW_1_chinese embedding Last Checkpoint: a few seconds ago (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3
Run
Code

In [9]:

```python
import argparse
import logging
import os
import jieba
import math
import requests
import wiki as w
from gensim.models.fasttext import FastText
from gensim.models.word2vec import Word2Vec
from tqdm import tqdm
import matplotlib.pyplot as plt
import numpy as np
from sklearn.decomposition import PCA
```

```
D:\Anaconda3\lib\site-packages\gensim\utils.py:1197: UserWarning: detected Windows; aliasing chunkize to chunkize_serial
  warnings.warn("detected Windows; aliasing chunkize to chunkize_serial")
```

In [10]:

```python
from gensim.models import KeyedVectors
fasttext_300d_model = 'pre-trained/cc.zh.300/cc.zh.300.vec'
wordvectors_index = KeyedVectors.load_word2vec_format(fasttext_300d_model)
```

```
D:\Anaconda3\lib\site-packages\smart_open\smart_open_lib.py:398: UserWarning: This function is deprecated, use smart_open.open
instead. See the migration notes for details: https://github.com/RaRe-Technologies/smart_open/blob/master/README.rst#migrating-
to-the-new-open-function
  'See the migration notes for details: %s' % _MIGRATION_NOTES_URL
```

In [11]:

```python
wordvectors_index.most_similar('我')
```

Out[11]:

```
[('你', 0.7679306864738464),
 ('他', 0.7081897258758545),
 ('想', 0.7019494771957397),
 ('她', 0.6555640697479248),
 ('要', 0.6327636241912842),
 ('真的', 0.6318696737289429),
 ('就', 0.6110949516296387),
 ('知道', 0.6072597503662109),
 ('大家', 0.5957611799240112),
 ('叫我', 0.5952475070953369)]
```

In [12]:

```python
vec = wordvectors_index['我']
print(vec)
print('The Length of Vector:',len(vec))
```

```
 -1.680e-02  7.160e-02  4.490e-02 -1.471e-01  4.000e-04 -4.250e-02
 -3.530e-02 -5.910e-02 -2.190e-02  9.900e-03  2.067e-01 -9.030e-02
 -6.020e-02  2.170e-02 -1.490e-02  1.361e-01 -1.749e-01  2.330e-02
  5.110e-02 -5.310e-02  1.406e-01 -6.600e-03 -2.100e-03 -1.155e-01
  1.702e-01  3.890e-02  3.020e-02  1.132e-01  7.450e-02 -2.140e-02
  5.630e-02  2.810e-02 -3.800e-03 -4.350e-02  1.548e-01 -6.520e-02
  6.060e-02  5.900e-02 -4.400e-02 -8.500e-03  1.950e-02  7.940e-02
  1.300e-02  1.414e-01  9.690e-02  7.490e-02 -1.249e-01  1.498e-01
 -9.950e-02 -2.530e-02  2.080e-02  4.300e-02 -3.140e-02 -8.150e-02
  3.820e-02 -2.242e-01 -3.080e-02  8.130e-02 -3.800e-03  4.540e-02
  7.580e-02  6.230e-02  2.690e-02  3.122e-01 -1.950e-02  3.190e-02
  2.670e-02  3.210e-02  1.710e-02 -4.370e-02  4.320e-02  1.357e-01
  1.169e-01  7.180e-02  1.390e-02  2.910e-02 -3.730e-02  7.900e-03
 -2.280e-02  2.640e-02  1.611e-01 -6.830e-02  3.900e-02 -5.760e-02
 -3.260e-02  1.364e-01 -2.990e-02  6.710e-02  4.230e-02 -6.890e-02
 -4.540e-02 -1.358e-01 -6.200e-02  8.740e-02 -2.710e-02 -1.040e-01
  1.010e-02 -1.542e-01 -9.660e-02 -5.950e-02 -5.250e-02  3.290e-02
```

In [5]:

```python
import pandas as pd
df = pd.read_csv('wikidata.csv')
df
```

Out[5]:

	id	url	title	text
0	13	https://zh.wikipedia.org/wiki?curid=13	數學	數學是利用符號語言研究數量、結構、變化以及空間等概念的一門學科，從某種角度看屬於形式科學的一...
1	18	https://zh.wikipedia.org/wiki?curid=18	哲學	哲學是研究普遍的、根本的問題的學科，包括存在、知識、價值、理智、心靈、語言等領域。哲學與其他...
2	21	https://zh.wikipedia.org/wiki?curid=21	文學	文學在最廣泛的意義上，是任何單一的書面作品。更嚴格地說，文學寫作被認為是一種藝術形式，或被認...
3	22	https://zh.wikipedia.org/wiki?curid=22	歷史	歷史是指人類社會過去的事件和行動，以及對這些事件行為有系統的記錄、詮釋和研究。歷史可提供今人...
4	25	https://zh.wikipedia.org/wiki?curid=25	電腦科學	電腦科學是系統性研究資訊與計算的理論基礎以及它們在電腦系統中如何與應用的實用技術的學科。 它...

In [6]:

```python
df['text'][0]
```

Out[6]:

```
'數學是利用符號語言研究數量、結構、變化以及空間等概念的一門學科，從某種角度看屬於形式科學的一種。數學透過抽象化和邏輯推理的使用，由計數、計算、
量度和對物體形狀及運動的觀察而產生。數學家們拓展這些概念，為了公式化新的猜想以及從選定的公理及定義中建立起嚴謹推導出的定理。'
```

In [7]:

```python
import jieba
import numpy as np
jieba.set_dictionary('dict.txt.big')
```

In [13]:

```python
def stopwordslist(filepath):
    stopwords = [line.strip() for line in open(filepath, 'r', encoding='utf-8').readlines()]
    return stopwords
```

In [53]:

```python
def seg_sentence(sentence):
    sentence_seged = jieba.cut(sentence.strip())
    stopwords = stopwordslist('stops.txt')  # 加載停用詞的路徑
    outstr = ''
    for word in sentence_seged:
        if word not in stopwords:
            if word != '\t':
                outstr += word
                outstr += " "   #再次組合成【帶空格】的串
    return outstr
```

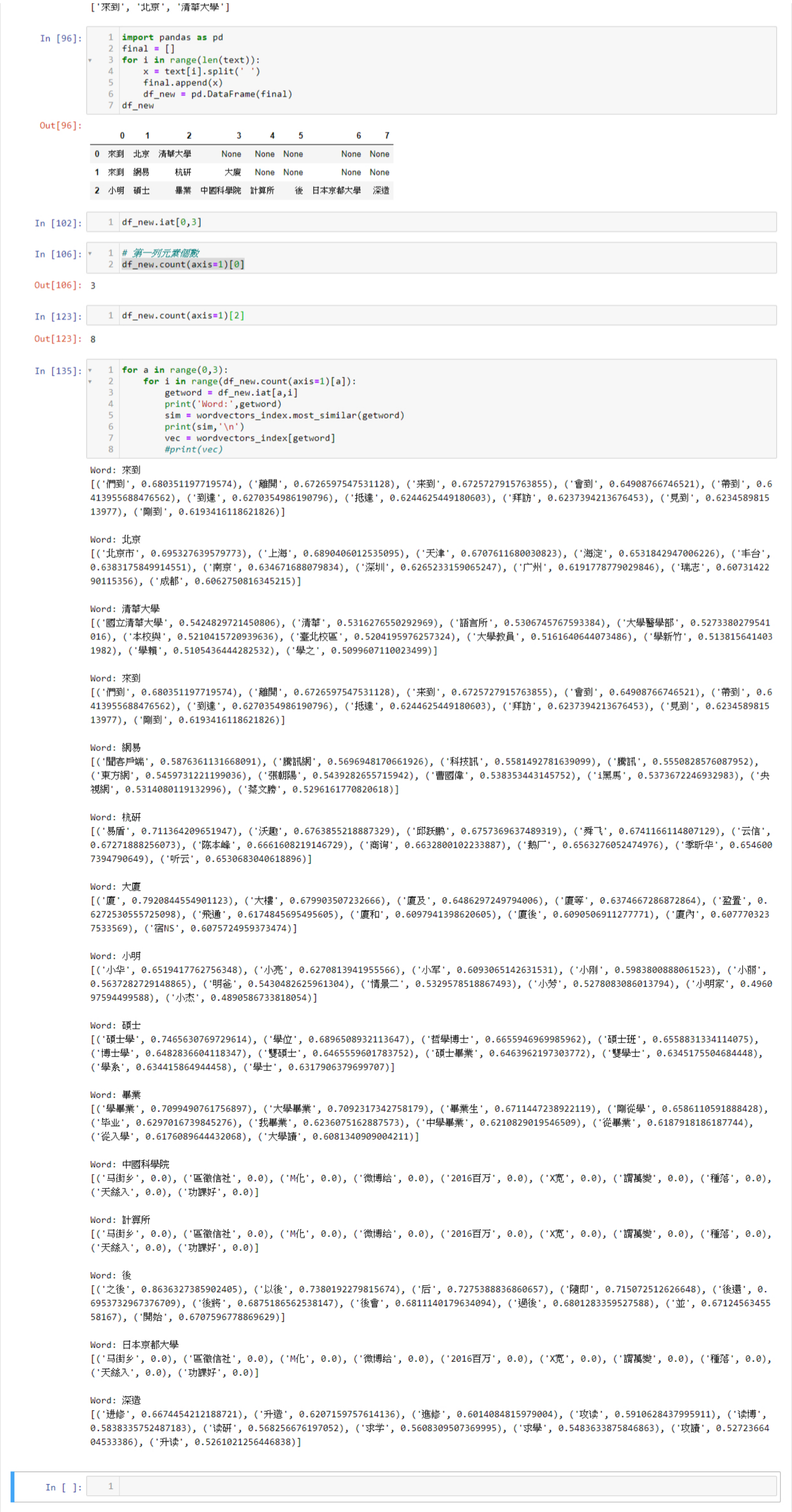
In [55]:

```python
inputs = open('input.txt', 'r', encoding = 'utf-8')
outputs = open('output.txt', 'w')

for line in inputs:
    line_seg = seg_sentence(line)  # 這裏的返回值是字符串
    outputs.write(line_seg + '\n')

outputs.close()
inputs.close()
```

In [91]:

```python
f = open(r'output.txt')
text = []
for line in f:
    line = line.strip('\n').strip(' ')
    text.append(line)
print(text)
print(text[0].split(' '))
```

```
['來到 北京 清華大學', '來到 網易 杭研 大廈', '小明 碩士 畢業 中國科學院 計算所 後 日本京都大學 深造']
```



```
['來到', '北京', '清華大學']
```

In [96]:

```python
import pandas as pd
final = []
for i in range(len(text)):
    x = text[i].split(' ')
    final.append(x)
    df_new = pd.DataFrame(final)
df_new
```

Out[96]:

	0	1	2	3	4	5	6	7
0	來到	北京	清華大學	None	None	None	None	None
1	來到	網易	杭研	大廈	None	None	None	None
2	小明	碩士	畢業	中國科學院	計算所	後	日本京都大學	深造

In [102]:

```python
df_new.iat[0,3]
```

In [106]:

```python
# 第一列元素個數
df_new.count(axis=1)[0]
```

Out[106]: 3

In [123]:

```python
df_new.count(axis=1)[2]
```

Out[123]: 8

In [135]:

```python
for a in range(0,3):
    for i in range(df_new.count(axis=1)[a]):
        getword = df_new.iat[a,i]
        print('Word:',getword)
        sim = wordvectors_index.most_similar(getword)
        print(sim,'\n')
        vec = wordvectors_index[getword]
        #print(vec)
```

```
Word: 來到
[('們到', 0.680351197719574), ('離開', 0.6726597547531128), ('来到', 0.6725727915763855), ('會到', 0.64908766746521), ('帶到', 0.6
413955688476562), ('到達', 0.6270354986190796), ('抵達', 0.6244625449180603), ('拜訪', 0.6237394213676453), ('見到', 0.6234589815
13977), ('剛到', 0.6193416118621826)]

Word: 北京
[('北京市', 0.695327639579773), ('上海', 0.6890406012535095), ('天津', 0.6707611680030823), ('海淀', 0.6531842947006226), ('丰台',
0.6383175849914551), ('南京', 0.634671688079834), ('深圳', 0.6265233159065247), ('广州', 0.6191778779029846), ('瑞志', 0.60731422
90115356), ('成都', 0.6062750816345215)]

Word: 清華大學
[('國立清華大學', 0.5424829721450806), ('清華', 0.5316276550292969), ('語言所', 0.5306745767593384), ('大學醫學部', 0.5273380279541
016), ('本校與', 0.521041572093963 6), ('臺北校區', 0.5204195976257324), ('大學教員', 0.5161640644073486), ('學新竹', 0.513815641403
1982), ('學賴', 0.510543644428253 2), ('學之', 0.5099607110023499)]

Word: 來到
[('們到', 0.680351197719574), ('離開', 0.6726597547531128), ('来到', 0.6725727915763855), ('會到', 0.64908766746521), ('帶到', 0.6
413955688476562), ('到達', 0.6270354986190796), ('抵達', 0.6244625449180603), ('拜訪', 0.6237394213676453), ('見到', 0.6234589815
13977), ('剛到', 0.6193416118621826)]

Word: 網易
[('聞客戶端', 0.5876361131668091), ('騰訊網', 0.5696948170661926), ('科技訊', 0.5581492781639099), ('騰訊', 0.5550828576087952),
('東方網', 0.545973122119903 6), ('張朝陽', 0.5439282655715942), ('曹國偉', 0.538353443145752), ('i黑馬', 0.5373672246932983), ('央
視網', 0.5314080119132996), ('蔡文勝', 0.5296161770820618)]

Word: 杭研
[('易盾', 0.711364209651947), ('沃趣', 0.6763855218887329), ('邱跃鹏', 0.6757369637489319), ('舜飞', 0.6741166114807129), ('云信',
0.67271888256073), ('陈本峰', 0.6661608219146729), ('商询', 0.6632800102233887), ('鹅厂', 0.6563276052474976), ('季昕华', 0.654600
7394790649), ('听云', 0.6530683040618896)]

Word: 大廈
[('廈', 0.792084455490112 3), ('大楼', 0.679903507232666), ('廈及', 0.6486297249794006), ('廈等', 0.6374667286872864), ('盈置', 0.
6272530555725098), ('飛通', 0.6174845695495605), ('廈和', 0.6097941398620605), ('廈後', 0.6090506911277771), ('廈內', 0.607770323
7533569), ('宿NS', 0.6075724959373474)]

Word: 小明
[('小华', 0.6519417762756348), ('小亮', 0.6270813941955566), ('小军', 0.6093065142631531), ('小刚', 0.5983800888061523), ('小丽',
0.5637282729148865), ('明爸', 0.5430482625961304), ('情景二', 0.5329578518867493), ('小芳', 0.5278083086013794), ('小明家', 0.4960
97594499588), ('小杰', 0.4890586733818054)]

Word: 碩士
[('碩士學', 0.7465630769729614), ('學位', 0.6896508932113647), ('哲學博士', 0.6655946969985962), ('碩士班', 0.6558831334114075),
('博士學', 0.6482836604118347), ('雙碩士', 0.6465559601783752), ('碩士畢業', 0.6463962197303772), ('雙學士', 0.6345175504684448),
('學系', 0.634415864944458), ('學士', 0.6317906379699707)]

Word: 畢業
[('學畢業', 0.7099490761756897), ('大學畢業', 0.7092317342758179), ('畢業生', 0.6711447238922119), ('剛從學', 0.6586110591888428),
('毕业', 0.6297016739845276), ('我畢業', 0.6236075162887573), ('中學畢業', 0.6210829019546509), ('從畢業', 0.6187918186187744),
('從入學', 0.6176089644432068), ('大學讀', 0.6081340909004211)]

Word: 中國科學院
[('马街乡', 0.0), ('區徽信社', 0.0), ('M化', 0.0), ('微博給', 0.0), ('2016百万', 0.0), ('X宽', 0.0), ('謂萬變', 0.0), ('種落', 0.0),
('天絲入', 0.0), ('功課好', 0.0)]

Word: 計算所
[('马街乡', 0.0), ('區徽信社', 0.0), ('M化', 0.0), ('微博給', 0.0), ('2016百万', 0.0), ('X宽', 0.0), ('謂萬變', 0.0), ('種落', 0.0),
('天絲入', 0.0), ('功課好', 0.0)]

Word: 後
[('之後', 0.8636327385902405), ('以後', 0.7380192279815674), ('后', 0.7275388836860657), ('隨即', 0.71507251262664 8), ('後還', 0.
6953732967376709), ('後將', 0.6875186562538147), ('後會', 0.6811140179634094), ('過後', 0.6801283359527588), ('並', 0.67124563455
58167), ('開始', 0.6707596778869629)]

Word: 日本京都大學
[('马街乡', 0.0), ('區徽信社', 0.0), ('M化', 0.0), ('微博給', 0.0), ('2016百万', 0.0), ('X宽', 0.0), ('謂萬變', 0.0), ('種落', 0.0),
('天絲入', 0.0), ('功課好', 0.0)]

Word: 深造
[('进修', 0.6674454212188721), ('升造', 0.6207159757614136), ('進修', 0.6014084815979004), ('攻读', 0.5910628437995911), ('读博',
0.5838335752487183), ('读研', 0.568256676197052), ('求学', 0.5608309507369995), ('求學', 0.5483633875846863), ('攻讀', 0.52723664
04533386), ('升读', 0.5261021256446838)]
```

In [ ]: