資管財金工作坊

情感文本分析

資管二 陳約廷

Index

- CSV 檔案 I/O (Input/Output)
  - read_csv.py
- Jieba 中文切割
  - test_jieba_v0.py
  - test_jieba_v1.py
  - test_jieba_v2.py
- 統計詞語出現頻率
  - count_freq.py
- 篩選正負面詞
  - select_pos_neg.py
- 使用新聞資料庫
  - get_data.py

編碼問題 Encode v.s. Decode

->UTF-8 Input

->UTF-8->Unicode

->Process…….

->Unicode->UTF-8

->UTF-8 Output

read_csv.py

```
15  infile = "samplenews_500.csv"
```

第15行指定出要讀入什麼檔案

注意！要把欲讀入檔案與此.py檔放在同一層資料夾才可以

中文字切割

Origin 文本 ：

我來到北京清華大學

切割後文本 ：

精確模式 (cut_all=False)

我/ 來到/ 北京/ 清華大學

全模式 (cut_all=True)

我/ 來到/ 北京/ 清華/ 清華大學/ 華大/ 大學

特別詞問題

地名，人名，公司名，字典沒有的東東

->擴充「辭典」

斷詞問題

標點符號，語助詞，冗詞贅字...

->建構「斷詞辭典」

test_jieba_v1.py

```
3 jieba.load_userdict('newword.txt')
```

擴充辭典

->新增文字檔 ”newword.txt”

詞典格式：

一個詞一行

-詞語

-詞頻（可省略）

-詞性（可省略）

test_jieba_v2.py

```
 7  re_float = re.compile('([+-]?\d+(\.\d*)?|\.\d+)([eE][+-]?\d+)?')
 8  stoplist = "cstop.dic"
 9  fstop1 = codecs.open(stoplist,'r',encoding='utf8')
10  cstop=[]
11  for aline in fstop1:
12      aline = aline.strip()
13      atoken = aline.split(' ')[0]
14      cstop.append(atoken)
15  fstop1.close()
```

新增斷詞辭典 ”cstop.dic”

test_jieba_v2.py

```
20  res=[]
21  for at in words:
22      at = at.strip(' ;=.-,/()%:"[]*\n')
23      if len(at) == 0:
24          continue
25      if at in cstop:
26          continue
27      rem1 = re_float.findall(at)
28      if len(rem1) > 0:
29          continue
30      res.append(at)
```

拔除 ;=.-,/()%:“[]*\n 這種符號字元
拔除 ctop中的所有符號
拔除 數字

test_jieba_v0.py

民進/黨/立委/林岱/樺/日前/在/主持/動保法/修法/協商時/，/一席/「/放生/論/」/引起/爭議/，/就/連黨/內/同志/都/嚴厲/批評/，/林岱/樺/因此/被/網友/戲稱/「/放生/立委/」/。

test_jieba_v1.py

民進黨/立委/林岱樺/日前/在/主持/動保法/修法/協商時/，/一席/「/放生論/」/引起/爭議/，/就/連黨/內/同志/都/嚴厲/批評/，/林岱樺/因此/被/網友/戲稱/「/放生/立委/」/。

test_jieba_v2.py

民進黨/立委/林岱樺/日前/主持/動保法/修法/協商時/一席/放生論/引起/爭議/連黨/內/同志/嚴厲/批評/林岱樺/因此/網友/戲稱/放生/立委

count_freq.py

```
67  infile = "samplenews_500.csv"
```

讀入 “samplenews_500.csv” 作為分析的文本

```
83  topn=30
```

顯示出現頻率前30高的

select_pos_neg.py

```
55  #Fetch postive list=================================
56  poslist = []
57  posfile='ntusd/NTUSD_positive_utf8.txt'
58  fhp1 = codecs.open(posfile,'r',encoding='utf8')
59  for aline in fhp1:
60      aline = aline.strip()
61      poslist.append(aline)
62  fhp1.close()
63  #End Build===========================================
64
65  #Fetch negative list=================================
66  neglist = []
67  negfile='ntusd/NTUSD_negative_utf8.txt'
68  fhp1 = codecs.open(negfile,'r',encoding='utf8')
69  for aline in fhp1:
70      aline = aline.strip()
71      neglist.append(aline)
72  fhp1.close()
73  #End Build===========================================
74
75  pos=[x for x in res if x in poslist]
76  neg=[x for x in res if x in neglist]
```

像先前一樣，完成切勾後放入res...

載入 NTU Sentiment Dictionary

-NTUSD_positive_utf8.txt

-NTUSD_negative_utf8.txt

將結果存於pos與neg之中

get_data.py

```
4  try:
5      conn = psycopg2.connect("host=%s dbname=%s user=%s password=%s" %
6          ('ana.lu.im.ntu.edu.tw', 'im_fin', 'imfin_read1', 'read998811'))
7      conn.set_client_encoding('UTF8')
8      cur = conn.cursor(cursor_factory = psycopg2.extras.DictCursor)
9  except psycopg2.DatabaseError, dbnamee:
10     print 'DatabaseError in open connection: %s' % dbe
11
12 cur.execute("select title,content from twnews where stock_id=2311 limit 10")
13 res=cur.fetchall()
```

Data base credit to 冠宗學長
選擇股票代號(stock_id)之後，放入res之中...

於text資料夾中的name_to_id.txt可以找到股票與他對應的ID

問卷～～～～～