nl_ex

April 19, 2016

1 自然语言处理实验报告

专业： 智能科学
班级： 1301
学号： 0909111024
姓名： 张日月明

1.1 实验 1

怎样载入自己的英文语料库（obama.txt），在自己的语料库中找出频率小于8，排名前5的词和其频率。（使用nltk的英文分

```
In [7]: import nltk

        text = open('obama.txt', 'rb').read()
        tokens = nltk.word_tokenize(text)
        freqdist = nltk.FreqDist(tokens).items()
        answer = [k for k in freqdist if k[1]<8]
        answer.sort(lambda x,y:cmp(y[1],x[1]))
        print answer[0:5]
```

```
[('go', 7), ('work', 7), ('learn', 7), ('give', 7), ('something', 7)]
```

1.2 实验 2

写程序处理布朗语料库，找到一下的答案：

(1) 哪些名词常以他们复数形式而不是它们的单数形式出现？（只考虑常规的复数形式，-s后缀形式的）。

```
In [8]: print('(1):')
        wsj = nltk.corpus.treebank.tagged_words()
        word_tag_fd = nltk.FreqDist(wsj)
        #print([word + '/' + tag for (word, tag) in word_tag_fd if tag.startswith('NNS')])
        print([word for (word, tag) in word_tag_fd if tag.startswith('NNS')])
```

```
(1):
'shipments', u[illegible]
```

(2) 选择布朗语料库的不同部分（其他目录），计数包含wh的词，如：what，when，where，who和why。

```
In [9]: print('(2):')
        news_text = nltk.corpus.brown.words(categories='news')
        freqdist = nltk.FreqDist([w.lower() for w in news_text])
        demands = [w.lower() for w in freqdist if w.startswith('wh')]
        for k in demands:
                print(k + ':',freqdist[k])
```

```
(2):
(u'whose:', 22)
(u'whirling:', 1)
(u'whiz:', 2)
(u'whip:', 2)
(u'wherever:', 1)
(u'where:', 59)
(u'whitfield:', 1)
(u"what's:", 1)
(u'wheeler:', 2)
(u'wheeled:', 2)
(u'whopping:', 1)
(u'whelan:', 1)
(u'whereby:', 3)
(u'whichever:', 1)
(u'whipped:', 2)
(u'wheels:', 1)
(u'wheat:', 1)
(u'whatever:', 2)
(u'while:', 55)
(u'whisking:', 1)
(u'white-clad:', 1)
(u'whitey:', 1)
(u'whites:', 2)
(u'whitney:', 1)
(u'whipple:', 1)
(u'wholly:', 1)
(u'wheel:', 4)
(u'when:', 169)
(u'whiplash:', 1)
(u'white:', 57)
(u'whee:', 1)
(u'wholly-owned:', 1)
(u'wholesale:', 1)
(u'what:', 95)
(u'whom:', 8)
(u'whims:', 1)
(u'which:', 245)
(u'who:', 268)
(u'why:', 14)
(u'whole:', 11)
(u'whether:', 18)
```

1.3 实验3

输出brown文本集名词后面接的词性

```
In [10]: brown_lrnd_tagged = nltk.corpus.brown.tagged_words(categories='learned')
         tags = [b[1] for (a, b) in nltk.bigrams(brown_lrnd_tagged) if a[1].startswith('N')] #a[0] == '
         freqdist = nltk.FreqDist(tags)
         freqdist.tabulate()
```

```
IN         .         ,         NN        CC        NNS       BEZ       MD        CS        VBN       VBZ
    11913      4998      4968      3725      2577      1737      1336      1080       827       732
```

1.4 实验4

句法分析演示
运行以下代码即会弹出 GUI 界面查看演示

```
nltk.app.srparser()
nltk.app.rdparser()
```

1.5 实验5

中文分词
实验数据：

- word_freq_list.txt 分词词典
- pku_test.txt 未经过分词的文档文件
- pku_test_gold.txt 经过分词的文档文件

```
In [11]: import jieba
         #import sys
         #reload(sys)
         #sys.setdefaultencoding('utf-8')

         text = open('pku_test.txt', 'rb').read()

         jieba.load_userdict('word_freq_list.txt')
         seg_list = jieba.cut(text)

         #open('pku_test_seg.txt', 'wb').write(' '.join(seg_list))
         print(' '.join(seg_list))
```

```
 - -
                1

 2001                        21


                                        20
                          20
                          20

                                                                  "


      12  31
```

2001 1 1

12 31 - - - 2001

4

2000

40

12 31 12 30 55.6 11.1 25 2000

12 31

30 3000 28

20

" "

" … … "

" "

12 31

20 " " - - -

- - - 21

WTO 20

- - -

21

2001 1 1

20

0

3 ℃ 8 ℃

2000 12 31 " " 82

2000 12 31 23 12 2001 1 1 9 23

2000 12 31 9 30 3000
" " 30 1

" "
" "

" " " "

" "

- - -

20
1900 2000

" "
20
2001 " "
" " 6300

17502

2000 12 22 2 23 2 " 2000 " 16
12 24 6
12 31 6
12 31 12 6
2000 12 31

1 1

" " " " "
7 " " " "
2000 12 31 " " " … …

2000 12 31 21 2001
2000
2000 12 31 2000 2000
2 21 25 " " " "
3 15

6 28 29
9 15 10 1 28
10 9 11

11 8 14 3 8 9 14
11 9 200 5 " " 841 1200
12 27
12 30 1

2000 12 31

21 20
10 10 1993 9
1993 " " 7 " " 15
" " 6
" "

12 31 31

12 19 89 24
19 " " .

12 30 · 30

12 31 20

28 3

2000 " " 9

" " "
12 30

12 31 2001

31 10 2020

2001 " "

" "

" "

" "

2000

12 31 31

50

61

" " " "

" "

"

" "

" "

2000 12 29

12 30 30 " "

12 8 4 26 29 26

· 1913 1943 1973

" " " " " "

" U - 2 "

8

1 " " 1955

2 " " " "

3

4 568 302

5 " "

6 300 70 35

7 1

8 7000

1988 9 5

1994 5 12 2000 12 28

1989 1 1 1978 8 18 1978 8 19
2000 12 28

2000 12 28

2000 12 28

2000 12 28

" "
" "
"
"
" "
" "
" "
" "
" " " "
" " " " " " "
"
"
"
"
"
"

1 .
2 .
3 .
4 .
5 .
6 .

2000 12 28

12 29 2000 - 6.16 1981

10

- 1999 30 1997 1998 21
10

2000 12 2001 1 2 4

]]

2000

12 30 " " " " 250
1999

" "
IT
21 " "
IT 2000 23
90 " "

1915 3 4 47

- - -

" AM21B " 37 1 49

" " " happynewyear . txt . vbs "

" "
60
" " "
1916 1920
5 1932 90 10 60
" " 6 6
300
" " 1 5

" "
" " " " " "

1
1 1
2001 " "

" "

A D

" "

.

1 1

1

" "

" " " "

- " " " " " "

A B

"

" "

" " " "

1 1

1 1 21

12 31 21 23 - - - " " 100 2.5

2000 12 31 8 30 1

9 4000 "

2001

" " 9 2100 21

12 3000 " 2001 "

1 1

1 1 15

8 50 15

6 15 " " 3

1 1

" " " " " "

600

1 1

" " 500

1 1 20
3 60
" " " "
5 " " " " 5
60 40 240 10 20
" "

1 1 - -
" " 112.7 100

2000
" " " " " " " " " " " " "

" "

1 1
1999 346 104
"

1 1 1 1 " "
" " 50
2002
2001 2002 1 1
2002 2003 1 1 2003 4 1

1 1
1 1
2001 1 1
2001 1 1

4
2001 1 1 21

1
2
3
4

1 1 2000 12 31 10
2000 1 1 8 12 31 8 1998 12 19

1 1 - - - 20

" "
" "

40
" "

12 31
31
31 2001 " "
31
30 " " 2000
21
12 31
31
.

" "

2001 " "
1 1

10

1 1

12 31 31
.

12 31 31 242 338

1 1 1

12 30 16

12 28 1947

1932 1954 1963 30 1969
1998 " "
21
29
12 30
4
1 2000 12 31
2 2000 12 31
3 2001 1 1 2001
4 2000 12 31

2001 42 20

1
12 31

12 31 " "
30
" " "

1 1

1 1 21
21 21

12 31 2000

12 30

50
2000

" "
1 1
" - - - "
" - - - "

100 2008 6000
1 1 " 250

250
" "
1 1 " "
1998 3 28 56
3 160 7.8 39 6.3 2000
3 10 1.8

1 1

1 2 20 - 1 3 20

1 2 3 4 - 6 6 - 7

9 ℃ 0 ℃
4 ℃ 10 ℃
9 ℃ 0 ℃
10 ℃ 19 ℃
5 ℃ 6 ℃
6 ℃ 14 ℃
10 ℃ 5 ℃
3 ℃ 5 ℃
17 ℃ 5 ℃
5 ℃ 1 ℃
25 ℃ 13 ℃
1 ℃ 8 ℃
12 ℃ 4 ℃
3 ℃ 10 ℃
23 ℃ 17 ℃
4 ℃ 8 ℃
30 ℃ 23 ℃

12 ℃ 23 ℃
4 ℃ 12 ℃
15 ℃ 20 ℃
0 ℃ 9 ℃
13 ℃ 24 ℃
4 ℃ 9 ℃
7 ℃ 10 ℃
7 ℃ 10 ℃
16 ℃ 21 ℃
3 ℃ 7 ℃
4 ℃ 7 ℃
7 ℃ 17 ℃
24 ℃ 34 ℃
6 ℃ 10 ℃
17 ℃ 28 ℃
1 ℃ 6 ℃
10 ℃ 20 ℃
6 ℃ 4 ℃
11 ℃ 21 ℃
14 ℃ 3 ℃
5 ℃ 2 ℃
8 ℃ 4 ℃
2 ℃ 6 ℃
11 ℃ 4 ℃
2 ℃ 5 ℃
14 ℃ 21 ℃
3 ℃ 6 ℃
16 ℃ 20 ℃
2 ℃ 5 ℃

3 ℃ 5 ℃
2 ℃ 16 ℃
5 ℃ 12 ℃
2 ℃ 6 ℃
3 ℃ 7 ℃
2 ℃ 6 ℃
13 ℃ 5 ℃
35 ℃ 25 ℃
7 ℃ 10 ℃
4 ℃ 8 ℃
- - -

2001 1 1 " "
9 30
" " " " " "
20 20
20 20 30 " " " 20
" " "
17

20 79
" " - - -

1 1 6 1000
2000 12414
" " 9500 3110 100
13 55 100 85.93 301

100
1 1 " "
46 " " " " "
100 400 " " 30 300
" " 48 180

2001 1 1 5000 ·
1999 12 20 33 "
- - - 4
12 20 12 11
" "

1 1 2000 12 30 2001 1 1

1 1 50 " " " "
" "
500 1500 2500
1 1 2005 1.4 500
93 81 66 " "
2005 " "

2001 1 1 16.4 15.3 6.6 3462
1996 4 1 4971
WTO 1996 APEC " 2000 15

4 13
WTO
70
1 1 70

1 1
9 31 5 257 1029
500
- - - 2000 12 28
" " 500 81 110 8.8 5.35

2001 " "
9 9

- - -

" "

" CytoAnalysis2000 "

- - - 2

2000 12 31 5 2
10 - - -
" " 12 "
" 50 15 " " " " "
" " 11
2 " 8
12 31 2 3 800 3 405

2000 12 30 98 750
" " 12 31 8 " 24 " 31 11
31 8 1 1 12 30 2.5 3 8000

" "
1 1 10 500 30
- - - " " 12 25 1 1 50 5 90

1 677 119 510 48 12
" " " "
41 " " 148 " "
100
1 1 100

500 5 500 375 500 " " 2700

1 1 35 " " 80

174 " " " "

" " 11 " " 76
12 1 8
12 15 12 " " 10 1.2 16

A
A A

2000 12 31 " "

90
1 1 " " 11 331 2001 1 1
9975 9878 99 " "

" " " "

" "

" " " " " " " "

" " " " " "

" " " "

1 1

5000 " "

1 1 " " "

7 36 "

9 44

" " 2001 1 2

190144 281640

1 1 138

1 1 13 45 1004 1004

160 3 97 WTO

" "

52 " " 70

" "

PPA

5000

" "

" "

" " 50

" "

2000 12 31 9

" "

1 1 " "

" "

10

" " " "

1 1 " " 5 55 - - -
175 60
. . "
5 55 " " 15 - - -

" "
8 150 - 300

" "

2000 12 28 5000 19 85
75

1 1
2001 2008

2000 12 31 7 1 25 46

" "
12 31
" " www . in paku . go . jp 100 1 50
" "
202

IT 1 IT 90
1 1

7.58 1 1.45 - - 12 31 6 2000

1

1.3 1963 153

90

" "

2000

22
18 40 " 21 " 8500 12 15 1700
20 600
2

2001 573 20
2000 2480 2030 450 2000 280
80 " "
2000 6
3 4 19 2000 2000 27 2001
14

2000 14 97

12 31 · –

6.6

1 6.6 3 66

20 – – –

10 ·

"

1

2

" "

3

" " – – –

2000 12 31 ·

" 160 50

5

"

" " "

1

12 20 2001 2

" "

1999 20

160 50 30

4 " "

1963 4 2

4500

118 7000

1996 1996 40

1999 50 3

" "

" "

1890 1910

1840 " "

1900

1868

1949

" " " " " "

" " " "

" "

1996

2001

" "

- - -

100733 010 - 65092341 caibian3 peopledaily com cn

300 3

" " - - - " "

" " " "

- - -

2000 11 22 1999 11 · 16
1999 11 16 10 25
" "
4 24

" "
" " " " " "
" " " " " " " "
" " " " " " " "

162 2465 85 93
162 39

188 3402

206

" " "

1
"
"

" "
1

- - - … …

1

1958

1

2001

2000

- - -

1

" " " " 2000 " " " "

2000 5 13 6

2000

1

1

1

2000 12 31

1

- - -

" "

" "

- - -

2000 9 7

100 1900 8 9 11

100 " "

1949 10 1 " "

A C

19

1955

50

1971 10

" " 80

1979 1
1989 5 " "
10
" "

- - - 21

" "
10
10 100
1991 1996

1100

1 2 60 " " " "
" " 2001 30 6

- - - " "
" "
" "
" "
" "

11 204
1 2
17 220 4 63.5

- - - 17
1995 QTL

" "
" " " " " " " "
" " " " " " " " " "
" " " " " " " " " "
" " " " " "
" "

1 2 " "
60 10 60

1 2 - - -
20

" "

1 2 12 26 6
5 4 3000

" " 42
1 1 42 " "
" " " "

" "

1 2

1991 200 1000 2000 3 4
800

" 3 5 " " 10 1 " … …
1 1

" "
" " – –

300 16
300.5 16
11 20 13.6
244 18.3
2000 244.8 1999 18.3 " "
3 18.8
25
2000 1141.6 24.6 449.9 12.3 497.9
" "
682
2000 682.6
" " 1996 300 1998 1999 400 500 2000
600 17.6
2000 600.3 90 17.6
22 10 224

1994
1 2 WTO 8
2 3 100
" "
24
" "
10 – – –
171
1 2 " "
" " " " 5 6250 3200

1 2 1 10
" " " "
" " 3
45

1
2
3 " 1 "

" "

1 2

1931 1997 7 70

3 4000
12 31 2002 1 1

" "
" " 90

1
1 2
45

5 30 25
1 1 1

1 1 1
. " " 120 61
1948 300

1 1 25 1

21
2001
12 31 1 1
20 21 57 21
1 2
15 " "
1 2 15 " " 2 9
12 31 60 " " " "
3 500 1 0.58 3 5000

244 2 50
33
2 2000 33 1990 11 2 365 1990

. . 360 56.9 2000 12 30
19 53
20
90
50

20

" "
" "

1 1
12 30
12 31
1

1

1 42
1 " "

12 31 8 1 200
6000 27 10
" " 1999 1 1

1 1 31 300
8
5 8 130 200

" "
1 1 " "
. . 1996 " "
" " " " 1940
4 " " 50 2
- - -

… … … …
10

1996 3 21 6
3

5 2000 100 1
3 10
10 2300 6000 1

8 5000
" "
" " 50
5 12
683

" "
" " 3000
" " 4
10 105 20 3.4 161 7
1 2

GDP
2000 " " 2000
2000 1 - 11 1812 1999 22.0 1228 2 .. 2

1 2 5.5 207 138 76
4 5200
2000 9 2001 9
9957.5 1.3 317.2 11.1 32.8

2000 12 28 75

7000 8 183 810

1 2 500

1 2

1
2000 12 20 2000 12 11 88

1928 1929 1935
1952 1956 197

" "

1 2

1 2

" "
17 " "

1 2 9 20
20 20 2001
" "
" "
" "
5 10
1 3 20 - 1 4 20

3 4 4

10 ℃ 2 ℃
4 ℃ 10 ℃
7 ℃ 1 ℃
9 ℃ 18 ℃
5 ℃ 4 ℃
8 ℃ 13 ℃
8 ℃ 6 ℃
6 ℃ 3 ℃
16 ℃ 6 ℃
5 ℃ 1 ℃
27 ℃ 16 ℃
2 ℃ 4 ℃

12 ℃ 4 ℃
3 ℃ 9 ℃
29 ℃ 20 ℃
3 ℃ 7 ℃
30 ℃ 22 ℃
14 ℃ 22 ℃
6 ℃ 11 ℃
12 ℃ 16 ℃
0 ℃ 7 ℃
17 ℃ 25 ℃
5 ℃ 9 ℃
8 ℃ 10 ℃
8 ℃ 12 ℃
16 ℃ 21 ℃
3 ℃ 6 ℃
4 ℃ 9 ℃
7 ℃ 18 ℃
24 ℃ 34 ℃
6 ℃ 9 ℃
17 ℃ 30 ℃
2 ℃ 8 ℃
11 ℃ 29 ℃
5 ℃ 3 ℃
12 ℃ 22 ℃
13 ℃ 4 ℃
4 ℃ 1 ℃
11 ℃ 2 ℃
1 ℃ 5 ℃
10 ℃ 4 ℃
4 ℃ 6 ℃
14 ℃ 23 ℃
6 ℃ 10 ℃
16 ℃ 20 ℃
5 ℃ 1 ℃

2 ℃ 7 ℃
4 ℃ 13 ℃
4 ℃ 14 ℃
2 ℃ 7 ℃
3 ℃ 9 ℃
4 ℃ 12 ℃
10 ℃ 4 ℃
30 ℃ 19 ℃
6 ℃ 11 ℃
4 ℃ 8 ℃

2000 12 30 8.9

1

1

1 1952 679 679

1

1 1 1998 87599

- - -

" "

1 2

19.8 7.1

12 3 9

28 50

3

- -

" "

" " " - - -

50 20

" "

"

" "

1 2

1 11 5 260 64 24 42 13

7 "

" "

" " " " " " " " " "

31 30 440

" " " " " " " " "

" "

" " - - " " " " " " " "

" " " "
- - -

21

3 21 37.2 7 14 5

2686 128 94.1

39 300 240

" " "

63 60 3375 1928

21 10

"

2001 1 1

50

1 2

" "

12 1 12 20

1998 9

1 2 3 17

1 2 28

15.8 56.6 4 14.4

"

" "

"

70 " "

" " " "

" "

" "

" "

" "

" "

" "

" "

2001 " " " "

2002 1 1 1.4

1999 1 1
10 " " 4 6
500 40
76 3.2 8 23 25.7 7 950

1 2

3 " " 12 " "

12 " " 1962 .
6 × 6 1841 1887 1908
24 3 70

5 1 1 1999 94 100
" " 1999 3 " "
" " 50 4
1
1000 12
1000
2640 1998 78 2000 245 "
" " IMB " " " " 69
" " " "
2004 100 " " " "
3

2000 2 .. 3 - 3 .. 1 5 3.6
3

9400 1 1999 12 31
50 1 9 33.39

3.2 - 4
1
2000 12 31 " "
" " 8000
9 10 4 10
31 5 30 5
500 " "

1 2 27
11 10 2001 400 24 150

2000 12 30 500

2001 3 2001 2001 6

10
1 2 10 1 4 17

6740 " " 1991 1 4 17 2

A

1 2 A " "

2001 1 1 " "
9 300

1 2 2000 28 16 15
2000 18 110 1 2 15 4
2000 14 2 30 22 8 21 14 3 2 6 5 3 3 3 4
2000 1408 1042 2000 1378 865

www . people . com . cn sports

3000

4000 1000

" " 17

20

1 2 " "
3

4

2000 - 2001 " " 2000 12 30

2000 12 24

2000 12 28 - - -
2003 3
7.4 217
4
1 1 2001

1 2000 12 30 2001 1 1 " "
2
3
4

" " " " " " " "

11 22 11 30 " " " "
12 5

" "

477 6 26 1

" " " " " "

60 200 60

200

11 30 180 12 5

12 5

" "

" "

" " " "

" " " " " "

" " " " " "

" "

2000

- - - 100

WTO

" " " " " "

3

1999

1
2
3

" "

" " " "

" "

- - -

" "

" "

" " " "

" "

- - -

" "

" "

- - -

" "

" "

" " " "

- - -

- - -

" " - - -

" "

10

" " " "

" "

" "

.

1

2000 11 5

" "

" "

400 380 3000

1996 1000

1997 700 1

8000 5000

1998 11 15

1999 8 2

8 21

1999 10 9

1999 11 1

11 10

1999 12 8

1983 100

5000

2000 1 21 " "

2 7000 5

"

" "

" "

" "

" "

" "

– – –

50 84 1998

– – –

… …

1998

79

15

65

20 " "

" " " "

" "

1999 8

" " " " "

1999 7

1 6 63 36 57

67 33

82 700

" " " "

40

– – –

" " "

… …

"

" "

1999 4 1997 10

1999 480 300 350

" "

" " " " 1999 30

" "

" " "

" "

– – –

" "

1998

8 13