A Computational Linguistic Analysis of Party Congress Reports

对党的十九大报告进行计算语言学分析

李亦萌

2017-10-27

Our target: a “word cloud” graph

Steps necessary

- **Step 1:** Analyze a Chinese sentence and perform text segmentation:

“今天天气很好” → [“今天”,”天气”, “很”, “好”]

- **Step 2:** Calculate the frequencies of each word

- **Step 3:** Plot a graph of high-frequency words
  - Caution: remove function words as “的”, “了”

Step 1: Chinese Segmentation

- Python package: Jieba
- 安装：打开cmd或terminal，运行pip install jieba

```
import jieba
string = '今天天气特别好，很开心'
result = jieba.cut(string)
print(list(result))
```

结果：

```
['今天天气', '特别', '好', '，', '很', '开心']
```

Step 2: Count frequency

- Function 1: Calculate frequency

```
>>> from collections import Counter
>>> Counter('adffdsads')
Counter({'d': 3, 'f': 2, 's': 2, 'a': 2})
```

- Function 2: Find most common words

```
>>> c = Counter('adffdsads')
>>> c.most_common(3)
[('d', 3), ('a', 2), ('f', 2)]
```

Step 3: Plot “word cloud” graph

```
# 导入 wordcloud 模块和 matplotlib 模块
from wordcloud import WordCloud
import matplotlib.pyplot as plt
# 读入一个txt文件
text = open('Jane Eyre.txt','r').read()
# 生成词云
wordcloud = WordCloud().generate(text)
# 显示词云图片
plt.imshow(wordcloud)
plt.axis('off')
plt.show()
# 保存图片
wordcloud.to_file('test.jpg')
```

Code Analysis (from Jupyter Notebook)

Code

18th Party Congress

19th Party Congress

Thank You

谢谢大家

李亦萌

2017-10-27