2019年11月3日 23:44

CS practice project report template
Project Name: The main priorities of the 19th national congress
Your Name: Cathy YangTianrui

1. Problem definition

1.1 What is the project about? (What is it? Why you want to do it? Possible applications)

It is a text mining about the report on the 19th National Congress of the Communist Party of China. I want to test which noun that Xi Jinping mentioned with the highest frequency to find which field's state will become most concerned in the coming years. I will use python to realize it.

1.2 What are the success criteria? (What are your goals that will guide you through design and implementation and can be evaluated)

This program could make a list with the rank of those noun's frequency, and make a word cloud to show the frequency of those most importance nouns.

2. Design overview

How to solve the problem in part 1? You need to design your software system, design about the input, process, and output parts.
Please list the Identifier tables if needed.

Identifier	type	Explanation
text	file	the needed file
content	string	the information in the file
stop_words	array	list those words that we do not need or appear in our result
stat	dictionary	to store needed words
segs	generator	to store words those are already sliced form the whole paragraph
seg	generator	single items from the 'segs'

INPUT:
The txt file of the content of the 19th national congress

PROCESS

Describe your key algorithms in flowchart or pseudocode.

Key algorithm 1:

```
OPENFILE ’19.txt’ FOR READ
WHILE NOT EOF("Test.txt")
READFILE "Test. txt", temp
content←content + temp
ENDWHILE
CLOSEFILE
```

```
stop_words ←['我们', '以', '把', '了', '到', '上', '有','中国','国家','全面','党',
            '问题','世界','时代','领导','战略','力量','特色','关系','全党']
```

Key algorithm2:

```
FOR seg IN segs:
   IF seg IS A NOUN:
      THEN
      IF seg NOT IN stop_words:
         THEN
         stat ← stat+seg
      ENDIF
   ENDIF
EDNFOR
```

Key algorithm 3:

```
cloud.fit_words(words)
plt.imshow(cloud)
plt.axis('off')
```

OUTPUT:
Print word cloud and sorted list

3. Implementation

List the key techniques used in your software, like arrays, searching, sorting, dictionaries, python modules (third-party, like matplotlib, numpy, scipy, sklearn etc.)

Tech #1:jieba

Why you use tech 1 in your project?

To cut a whole paragraph into different words and mark each word with a part of speech.

How do you use it? Put your code screenshots here.

```
segs = pseg.cut(content)
…
zh_pattern.search(seg.word)
…
if str(seg.flag)[0] == 'n':
```

Tech #2:wordcloud

Why you use tech 2 in your project?

making a word cloud to visualize the result

How do you use it? Put your code screenshots here.

```
cloud = WordCloud(font_path='C:\\simhei.ttf', background_color='white')
    words = pt_stat['19'].to_dict()
    cloud.fit_words(words)
    plt.imshow(cloud)
    plt.axis('off')
    plt.show()
```

Tech #3:dictionary

Why you use tech 3 in your project?

Label each items that easily to use the label to sort

How do you use it? Put your code screenshots here.

```
stat.append({'from':'19','word': seg.word})
```

4. Results

Show the outcomes/results of your projects

```
from    19
word
一体        2
本质特征      2
本质属性      2
本位主义      2
服务网络      2
服务水平      2
服务型       2
机会        2
朋友        2
有效率       2
有效性       2
有度        2
最大公约数     2
暴力        2
智库        2
```

有益 2
机场 2
机械化 2
机遇 2
案件 2
格调 2
根本保证 2
根本任务 2
根基 2
校企 2
标志 2
架构 2
林田湖 2
构成威胁 2
极端 2
...
依法治国 38
强国 38
基础 40
体制 40
协商 42
机制 42
事业 42
政策 42
理论 44
人类 46
国际 48
群众 50
能力 54
历史 54
民族 60
思想 60
基本 62
法治 62
生态 64
文明 66
中华民族 86
民主 88
经济 118
文化 132
体系 136
社会 160
政治 160
制度 166
社会主义 292
人民 314

5. Evaluation and conclusion

Evaluate the project based on your success criteria in Part 1.
It is successful- this program successfully makes a clearly word cloud to shows needed noun's frequency.

What is the conclusion of the project?
the 19th CPC National Congress was mainly focuses on 'people' and social system. As for those particular parts, he mentioned most about legal system, civilization and enthronement.

References
List all the materials you referred to.
Jianshu(2019) Python3 文本挖掘https://blog.csdn.net/likunkun__/article/details/81707883

CSDN(2018) python之re模块详解https://www.jianshu.com/p/ba3798562dea

CSDN(2015)python的jieba分词词性标注
https://blog.csdn.net/li_31415/article/details/48660073

the Xinhua News Agency(2017) 习近平在中国共产党第十九次全国代表大会上的报告（全文）
https://www.guancha.cn/politics/2017_10_27_432557_s.shtml