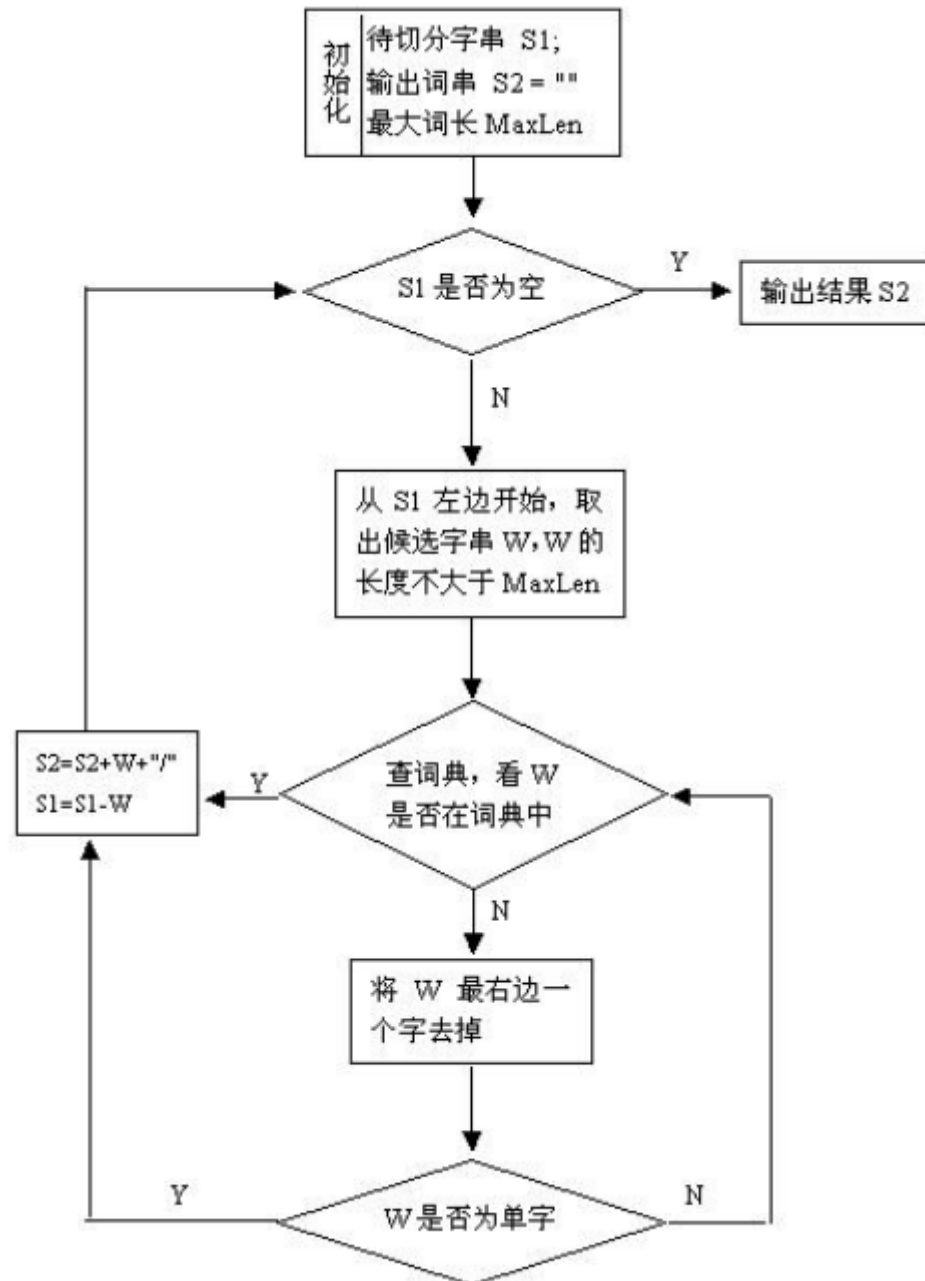


一. 正向最大匹配算法

1. 去除标点

资本结构的优化使青岛一批企业迅速崛起青啤公司海尔集团和双星集团的产品品牌已先后被认定为中国驰名商标

算法实现逻辑



```

1  # 最大正向匹配
2  def max_left_match(line, dict):
3      input_str = line
4      output_str = ""
5      # 最大词长
6      max_length = dict['max_length']
7      word_dict = dict['word_dict']
8      while input_str.strip() != '':
9          num = max_length
10         w = input_str[0:num]
11         while w not in word_dict:
12             num -= 1
13             w = w[0:num]
14             if len(w) == 1:
15                 break

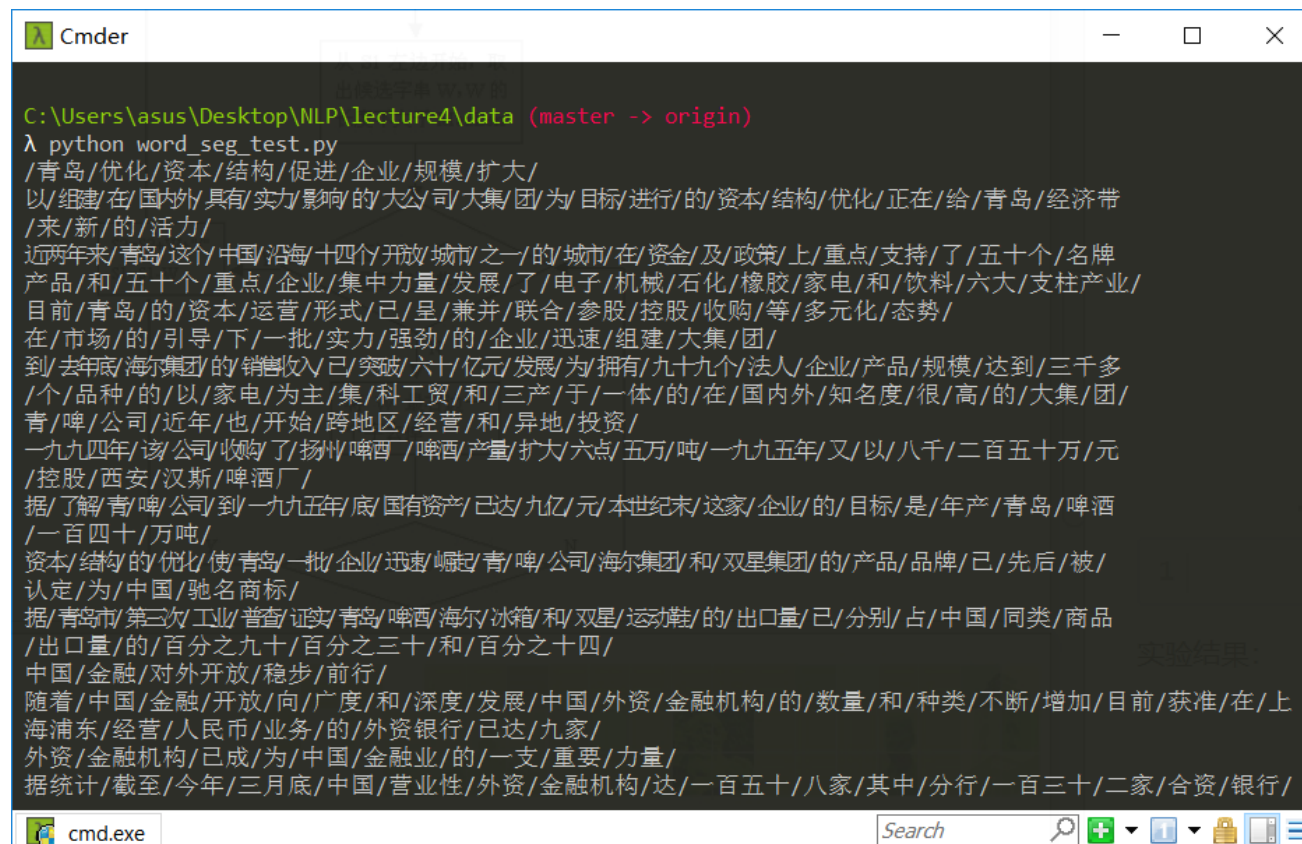
```

```

16         output_str += w + "/"
17         input_str = input_str[len(w):]
18     return output_str

```

实验结果：



```

C:\Users\asus\Desktop\NLP\lecture4\data (master -> origin)
λ python word_seg_test.py
/青岛/优化/资本/结构/促进/企业/规模/扩大/
以/组建/在/国内外/具有/实力/影响/的/大/公司/大/集团/为/目标/进行/的/资本/结构/优化/正在/给/青岛/经济带
/来/新/的/活力/
近两年来/青岛/这个/中国/沿海/十四个/开放/城市/之一/的/城市/在/资金/及/政策/上/重点/支持/了/五十个/名牌
产品/和/五十个/重点/企业/集中/力量/发展/了/电子/机械/石化/橡胶/家电/和/饮料/六大/支柱产业/
目前/青岛/的/资本/运营/形式/已/呈/兼并/联合/参股/控股/收购/等/多元化/态势/
在/市场/的/引导/下/一批/实力/强劲/的/企业/迅速/组建/大/集团/
到/去年底/海尔集团/的/销售收入/已/突破/六十/亿元/发展/为/拥有/九十九个/法人/企业/产品/规模/达到/三千多
/个/品种/的/以/家电/为主/集/科工贸/和/三产/于/一体/的/在/国内外/知名度/很/高/的/大/集团/
青/啤/公司/近年/也/开始/跨地区/经营/和/异地/投资/
一九九四年/该/公司/收购/了/扬州/啤酒/啤/酒/产量/扩大/六点/五万/吨/一九九五年/又/以/八千/二百五十万/元
/控股/西安/汉斯/啤酒厂/
据/了解/青/啤/公司/到/一九九五年/底/国有资产/已达/九亿/元/本世纪末/这家/企业/的/目标/是/年产/青岛/啤酒
/一百四十/万吨/
资本/结构/的/优化/使/青岛/一批/企业/迅速/崛起/青/啤/公司/海尔集团/和/双星集团/的/产品/品牌/已/先后/被/
认定/为/中国/驰名商标/
据/青岛市/第三次/工业/普查/证实/青岛/啤酒/海尔/冰箱/和/双星/运动鞋/的/出口量/已/分别/占/中国/同类/商品
/出口量/的/百分之九十/百分之三十/和/百分之十四/
中国/金融/对外开放/稳步/前行/
随着/中国/金融/开放/向/广度/和/深度/发展/中国/外资/金融机构/的/数量/和/种类/不断/增加/目前/获准/在/上
海浦东/经营/人民币/业务/的/外资银行/已达/九家/
外资/金融机构/已成/为/中国/金融业/的/一支/重要/力量/
据统计/截至/今年/三月底/中国/营业性/外资/金融机构/达/一百五十/八家/其中/分行/一百三十/二家/合资/银行/

```

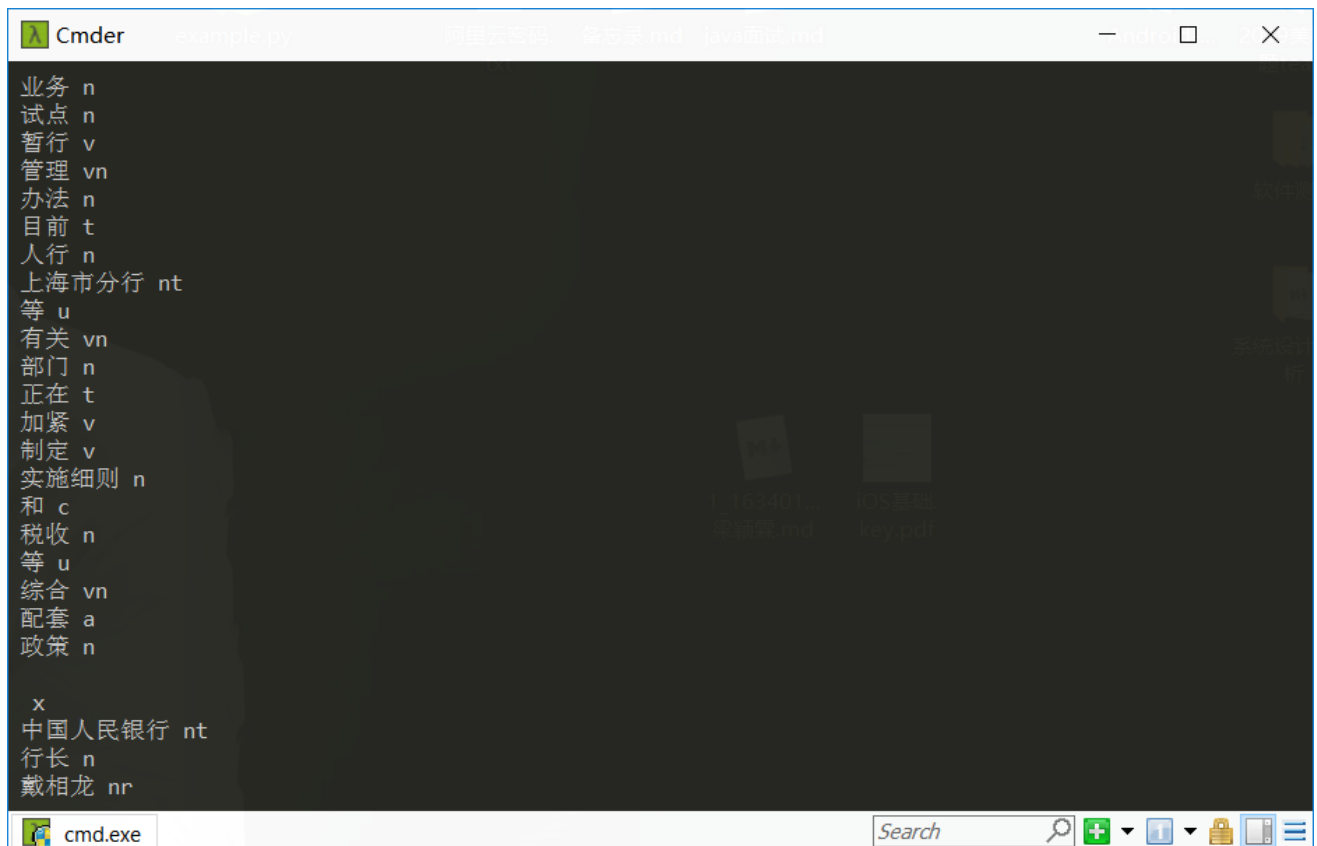
3.利用jieba库的分词功能并标记词性

```

1  # 利用jieba库的分词功能并标记词性
2  def jieba_cut(line):
3      line_seg = pseg.cut(line)
4      return line_seg

```

实验结果：



4.完整代码

```
1 # encoding=utf-8
2 import nltk
3 import string
4 import re
5 import jieba
6 import jieba.posseg as pseg
7
8 # 加载字典
9 def load_word_list():
10     max_length = 0
11     word_dict = set()
12     for line in open('./data/corpus.dict.txt', encoding='utf-8', errors='ignore').readlines():
13         tmp = len(line)
14         if(max_length < tmp):
15             max_length = tmp
16         word_dict.add(line.strip())
17     return {
18         'max_length': max_length,
19         'word_dict': word_dict
20     }
21
22 # 去标点
23 def filter_punctuation(line):
24     # 去除标点符号
```

```

25 punc = "[! ? , . \" ' # $ % & ' ( ) * + , - / : ; < = > @ [ \ ] ^ _ ` { | } ~ « » 「 」 『 』 【 】 ( ) [ ] { } ~ “ ” ‘ ’ ‘ ’ “ ” „ … … . ! \" # $ % & \ ' ( ) * + , - . / : ; < = > ? @ [ \ ] ^ _ ` { | } ~ ] + "
26 line = re.sub(punc, "",line)
27 return line
28
29 # 最大正向匹配
30 def max_left_match(line, dict):
31     input_str = line
32     output_str = ""
33     # 最大词长
34     max_length = dict['max_length']
35     word_dict = dict['word_dict']
36     while input_str.strip() != '':
37         num = max_length
38         w = input_str[0:num]
39         while w not in word_dict:
40             num -= 1
41             w = w[0:num]
42             if len(w) == 1:
43                 break
44         output_str += w + "/"
45         input_str = input_str[len(w):]
46     return output_str
47
48 # 利用jieba库的分词功能
49 def jieba_cut(line):
50     line_seg = pseg.cut(line)
51     return line_seg
52
53 # 测试
54 def main():
55     dict = load_word_list()
56     for line in open('./data/corpus.sentence.txt',encoding='utf-8',errors='ignore').readlines():
57         # 去标点
58         new_line = filter_punctuation(line)
59         # 自己写的最大匹配
60         # result = max_left_match(new_line, dict)
61         # 结果
62         # print(result)
63         # jieba库的分词并且标记词性
64         result = jieba_cut(new_line)
65         for word in result:
66             print (word.word,word.flag)
67
68 if __name__ == '__main__':
69     main()
70     # print(__name__)

```