计算语言学第五次作业: 搭配实验报告

1. 介绍

本次实验利用提供的经过人工分词和词性标注的语料 `corpus.json` ，选取搭配自动抽取方法进行搭配发现和抽取（注意需考虑词性信息，例如词性为名词的“诊断”和词性为动词的“诊断”应视作不同的词语分别考虑）。

具体来说，该实验包括如下三个部分：

- 对于每种方法，请算出10个最可能的词语搭配及其得分，按照可能性从大到小的顺序排列。
- 通过考察每一个可能的搭配在有语法信息标注的语料`ud.conllu`中的语法关系来进一步确定其是否真正的搭配。
- 分析不同方法得出的10个搭配的区别。

在这里，我们选择了t-test 假设检验方法、卡方(Pearson's Chi-square)检验方法、对数似然比率(Log Likelihood Ratio)假设检验方法和点对互信息(Mutual Information)方法抽取搭配。

2. 程序实现及运行

2.1 实现方法

该部分介绍进行搭配自动抽取的相关程序代码，该项目共包含四份代码文件，其用途如下：

文件名	用途
readCorpus.py	读取corpus.json数据集，并提取出词项出现频次列表，以及bigram共现矩阵(稀疏阵)
collocationCalculation.py	采用提取出来的信息，进行四种方法下最优搭配的计算及输出
result_organize.py	根据前两个程序的输出，整理top 10的结果进行展示
relation_analysis.py	根据前两个程序的输出，统计top 10的结果具有依存关系的数量

其中，`readCorpus.py`中包含两个函数，其具体介绍如下：

函数名	输入	功能
getTermList_TermFrequency()	--	读取corpus.json数据集，并提取出词项频次列表
getFilterTermList_CoappearMatrix()	词项频次列表	生成bigram共现矩阵(稀疏阵)

`collocationCalculation.py`包含四个函数，其具体介绍如下：

函数名	输入	功能
get_t_statistic_topK()	词项频次列表、bigram共现矩阵、选取的最可能搭配数量	t-test假设检验最优搭配的计算及输出
get_chi_square_topK()	词项频次列表、bigram共现矩阵、选取的最可能搭配数量	卡方检验最优搭配的计算及输出
get_LLR_statistic_topK()	词项频次列表、bigram共现矩阵、选取的最可能搭配数量	对数似然比率假设最优搭配的计算及输出
get_MI_statistic_topK()	词项频次列表、bigram共现矩阵、选取的最可能搭配数量	点对互信息方法最优搭配的计算及输出

值得注意的是，考虑到标点符号的搭配意义不大，因此我们在统计的时候，去掉了所有的标点符号，并且被标点符号分隔开的term我们不视为bigram。

2.2 程序运行

搭配自动抽取和验证的程序（使用卡方检验）运行命令如下：

```
python readCorpus.py
python collocationCalculation.py
python result_organize.py --function chi_square
python relation_analysis.py --function chi_square
```

而使用其他三种检验方式的参数分别为`LLR_statistic`（对数似然比率假设）、`MI_statistic`（点对互信息方法）和`t_statistic`（t-test假设检验）。

3. 实验结果

3.1 最可能的词语搭配及其得分

t-test假设检验结果如下：

Rank	Left_term	Right_term	t_statistic	Cnt_left	Cnt_right	Cnt_bigram
1	一/NUM	种/NOUN	24.65	8624	990	626
2	这/PRON	一/NUM	24.14	3353	8624	645
3	这/PRON	是/VERB	22.58	3353	10347	584
4	两/NUM	国/NOUN	22.38	2297	843	505
5	新/ADJ	的/PART	22.29	2179	62193	768
6	一/NUM	年/NOUN	21.57	8624	4277	544
7	就/ADV	是/VERB	21.56	2475	10347	520
8	本报/PRON	讯/ADP	21.49	1467	699	464
9	江/PROPN	泽民/PROPN	21.11	591	451	446
10	北京/PROPN	1月/NOUN	21.06	1438	1781	449

卡方检验结果如下：

Rank	Left_term	Right_term	chi_square	Cnt_left	Cnt_right	Cnt_bigram
1	甲方/NOUN	乙方/NOUN	899682	8	8	8
2	机不可失/NOUN	时不再来/NOUN	899682	1	1	1
3	还有/ADV	别的/DET	899682	1	1	1
4	亲密/VERB	起来/AUX	899682	1	1	1
5	有点儿/ADV	恼火/ADJ	899682	1	1	1
6	喀斯特地貌/NOUN	柞水溶洞/PROPN	899682	1	1	1
7	红色/ADJ	运动衣/NOUN	899682	1	1	1
8	奥巴/PROPN	贝勒/NOUN	899682	1	1	1
9	齐齐/ADV	声演/VERB	899682	1	1	1
10	这次/DET	失事/NOUN	899682	1	1	1

对数似然比率假设结果如下：

Rank	Left_term	Right_term	t_statistic	Cnt_left	Cnt_right	Cnt_bigram
1	江/PROPN	泽民/PROPN	6965.90	591	451	446
2	本报/PRON	讯/ADP	5229.63	1467	699	464
3	两/NUM	国/NOUN	5016.06	2297	843	505
4	一/NUM	种/NOUN	4569.05	8624	990	626
5	附/VERB	图片/NOUN	4399.52	295	628	293
6	据/ADP	新华社/NOUN	4228.31	1025	1177	410
7	年/NOUN	来/ADP	3939.18	4277	544	419
8	北京/PROPN	1月/NOUN	3931.11	1438	1781	449
9	钱/PROPN	其琛/PROPN	3612.84	250	205	205
10	不/ADV	能/VERB	2860.72	5037	1436	442

点对互信息方法结果如下：

Rank	Left_term	Right_term	t_statistic	Cnt_left	Cnt_right	Cnt_bigram
1	机不可失/NOUN	时不再来/NOUN	19.78	1	1	1
2	还有/ADV	别的/DET	19.78	1	1	1
3	亲密/VERB	起来/AUX	19.78	1	1	1
4	有点儿/ADV	恼火/ADJ	19.78	1	1	1
5	喀斯特地貌/NOUN	柞水溶洞/PROPN	19.78	1	1	1
6	红色/ADJ	运动衣/NOUN	19.78	1	1	1
7	奥巴/PROPN	贝勒/NOUN	19.78	1	1	1
8	齐齐/ADV	声演/VERB	19.78	1	1	1
9	这次/DET	失事/NOUN	19.78	1	1	1
10	东垣/PROPN	正中/NOUN	19.78	1	1	1

3.2 搭配关系验证

t-test假设检验结果如下：

Rank	Left_term	Right_term	Cnt_left_right	Cnt_right_left
1	一/NUM	种/NOUN	94	3
2	这/PRON	一/NUM	3	0
3	这/PRON	是/VERB	21	0
4	两/NUM	国/NOUN	9	0
5	新/ADJ	的/PART	0	34
6	一/NUM	年/NOUN	23	0
7	就/ADV	是/VERB	21	0
8	本报/PRON	讯/ADP	0	0
9	江/PROPN	泽民/PROPN	0	0
10	北京/PROPN	1月/NOUN	0	0

卡方检验结果如下：

Rank	Left_term	Right_term	Cnt_left_right	Cnt_right_left
1	甲方/NOUN	乙方/NOUN	0	0
2	机不可失/NOUN	时不再来/NOUN	0	1
3	还有/ADV	别的/DET	0	0
4	亲密/VERB	起来/AUX	0	1
5	有点儿/ADV	恼火/ADJ	1	0
6	喀斯特地貌/NOUN	柞水溶洞/PROPN	1	0
7	红色/ADJ	运动衣/NOUN	1	0
8	奥巴/PROPN	贝勒/NOUN	1	0
9	齐齐/ADV	声演/VERB	1	0
10	这次/DET	失事/NOUN	0	0

对数似然比率假设结果如下：

Rank	Left_term	Right_term	Cnt_left_right	Cnt_right_left
1	江/PROPN	泽民/PROPN	0	0
2	本报/PRON	讯/ADP	0	0
3	两/NUM	国/NOUN	9	0
4	一/NUM	种/NOUN	94	3
5	附/VERB	图片/NOUN	0	0
6	据/ADP	新华社/NOUN	0	0
7	年/NOUN	来/ADP	0	14
8	北京/PROPN	1月/NOUN	0	0
9	钱/PROPN	其琛/PROPN	0	0
10	不/ADV	能/VERB	0	0

点对互信息方法结果如下：

Rank	Left_term	Right_term	Cnt_left_right	Cnt_right_left
1	机不可失/NOUN	时不再来/NOUN	0	1
2	还有/ADV	别的/DET	0	0
3	亲密/VERB	起来/AUX	0	1
4	有点儿/ADV	恼火/ADJ	1	0
5	喀斯特地貌/NOUN	柞水溶洞/PROPN	1	0
6	红色/ADJ	运动衣/NOUN	1	0
7	奥巴/PROPN	贝勒/NOUN	1	0
8	齐齐/ADV	声演/VERB	1	0
9	这次/DET	失事/NOUN	0	0
10	东垣/PROPN	正中/NOUN	0	0

3.3 结果分析

通过比较我们可以发现，t-test 假设检验方法和对数似然比率假设检验方法更倾向于将词频较大的结果排在前面，而卡方检验方法和点对互信息方法则更倾向于将词频较小的结果排在前面。在这里我们对这一现象产生的原因进行简要地阐述。

假设p_1, p_2分别表示在一个搭配pair = (A, B)之中，词项A和B在bigrams中所出现的概率，p_{12}表示该pair在所有bigrams中出现的概率。假设$p_{12} = \alpha p_1 p_2$，其中，参数α表示$\frac{p_{12}}{p_1 p_2}$的比例大小。当$\alpha > 1$且越大时，A和B存在共现性，且共现状况更明显。

在t-test假设检验方法中，$t = \frac{p_{12} - p_1 p_2}{\sqrt{\frac{p_{12}(1-p_{12})}{N}}} \approx \sqrt{Np_{12}}(\alpha - 1)$，可以看出，t统计量不仅和$\alpha$有关，还跟A、B出现的频率相关，因此最终排名靠前的结果能在两者之间找到一个权衡，同样在对数似然比率假设检验方法中，我们可以推断总结出相同的结论。

而在卡方检验方法$\chi^2 \approx Np_1p_2(\alpha - 1 - \alpha p_1 - \alpha p_2)^2$，可以发现，它与A和B出现的频率有一定的正相关性，但主要还是跟α有关，这就导致了在该检验方法排名靠前的pair中， 大部分α很大但词频为1，但也有词频不为1的bigram。但在点对互信息方法中，这种不平衡性则更为严重。这是因为$MI = \log_2 \alpha$，它和词频没有关系，所以排序靠前的全是α很高但是词频为1的bigram。

进一步分析搭配关系验证的结果，我们可以发现，除了对数似然比率的方法，其他方法抽取的top10的结果，均有7条具有依存关系，而对数似然比率的方法仅有3条具有依存关系，但是具体看抽取的10条结果搭配效果还是不错的，原因可能是受到了依存关系数据集偏差的影响。