Collocation

计算语言学 第五章 搭配

使用经过人工分词后的北京大学《人民日报》标注语料库，实现**搭配自动发现程序**，要求至少实现以下几种方法：

1. 频率方法（请参考lesson5课件第8-10页）
2. 均值-方差方法（请参考lesson5课件第11-14页）
3. 假设检验方法（请参考lesson5课件第23-35页）
4. 点对互信息方法（请参考lesson5课件第37-42页）

鼓励根据课堂上讲的原则提出自己新的方法。

最终提交报告要求每种方法列出**前10个搭配的词对及其得分**，并对**实验结果进行较为充分的对比**，予以**必要的分析**。

✻ 先给出关于原始数据的一些统计信息：

File	Sentence	Token	Bigram
3148	137339	1120721	457709

其中，分句的分隔符为 。 ？ ！ ： ； ， —— 、( )

注：报告的全文围绕两个词作为搭配进行讨论，对超过两个词的搭配暂不考虑。

1. 频率方法

频率方法是在语料中找搭配最简单直接的方法。如果两个单词一起出现的频率高，那么我们有充分的理由相信这两个词在一起形成一个搭配。

下表是最原始的出现频率最高的10个bigram

rank	w_1w_2	$C(w_1w_2)$
1	”的	970
2	的 一	844
3	的 “	843
4	新 的	734
5	这 一	645
6	电 （	574
7	这 是	569

rank	w_1w_2	$C(w_1w_2)$
8	的 发展	537
9	（ 记者	530
10	一 种	529

表1 在语料中bigram的原始频率

虽然从上表中这些bigram出现的频率非常高，看不出有趣的搭配。

根据Justeson和Katz在95年提出的方法，将人的先验知识引入进来，使用POS的pattern筛选出感兴趣的搭配。

使用哈工大研发的pyltp进行pos标注。LTP 使用的是863词性标注集[1]。

Tag Pattern	Example
a n	红 苹果
n n	领导 干部

表2 Part of speech tag pattern for collocation filtering
n 包括ns，nd，nh，ni，nl，ns，nt，nz，n

经过上表pattern的filter后

筛选后频率排序结果如下表：

rank	w_1	w_2	$C(w_1w_2)$
1	北京	1 月	449
2	江	泽民	446
3	新华社	北京	286
4	新华社	记者	271
5	领导	干部	252
6	讯	记者	225
7	电	记者	219
8	李	鹏	216
9	钱	其琛	205
10	周	恩来	190

太多人名地名机构名和时间，这些都不是感兴趣的搭配，修改在pos pattern n的集合，去掉关于人名、地名、机构名和时间的POS。得到最后下表：

rank	w_1	w_2	$C(w_1w_2)$
1	领导	干部	252
2	讯	记者	225
3	电	记者	219
4	金融	危机	180
5	人民	群众	139
6	社会主义	市场经济	120
7	人民	检察院	120
8	两岸	关系	115
9	多	人	113
10	金融	机构	95

从表中可以看出，除了“讯 记者”、“电 记者”、“多 人”，其他都是比较有用的搭配。

频率方法的优点在于，这种方法可以很精确地给出搭配，比如“领导 干部”、“金融 危机”。缺点也是很明显的，有些搭配的出现频率差不多，此时就需要更加复杂的分析了。这个方法也十分受限于语料的大小。这种方法两个词相隔的距离是固定的。但事实上，很多搭配中间会插入一些不同的词。这就引入了下面的均值方差方法。

2. 均值–方差方法

为了适应搭配的灵活性，即搭配中的两个词不一定相邻出现，可能相隔长度不同的几个词。我们定义一个大小为3的搭配窗口，根据这个窗口中的词生成bigram，如此一来，可以灵活地找到距离范围在3内的搭配词组。

通过计算两个词距离的均值和方差（为了方便，取标准差），来发现搭配。

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} \qquad s = \sqrt{\frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n-1}}$$

其中n是两个词一起出现的频率大小，d_i是两个词第i次出现时的距离，$\overline{d}$是两个词所有距离的均值。

如果每次出现距离一样，则方差为0，如果每次出现的距离随机分布，方差会很高。均值方差方法刻画了两个词在语料中出现距离的分布情况。我们现在的目标就是要找到低标准差的两个词，这样找到的两个词很可能就是固定搭配。

去掉出现频率（count）仅为1的两个词。得到下表：

rank	s	$\overline{d}$	Count	W1	W2
1	0.0	1.0	2	允许	特委会
2	0.0	1.0	11	德国	人
3	0.0	**2.0**	2	国内外	具有
4	0.0	**2.0**	2	时间	４０分
5	0.0	1.0	6	据	巴黎
6	0.0	**2.0**	2	听众	超过
7	0.0	1.0	2	政治	口号
8	0.0	1.0	2	（	严重
9	0.0	**2.0**	2	当代	歌剧
10	0.0	1.0	2	第四十七	条

因为有很多只出现两次的词，每次出现的位置都相同，导致标准差都为0。从中有一些有用的搭配，如“德国 人”，“政治 口号”，“当代 歌剧”，但是依然有很多没有用的词组。那些$\overline{d} = 1$的搭配与用频率方法找到的相似。现在关注点在那些均值大于1且有低标准差的搭配。过滤掉那些均值$\overline{d}$为1的词语对，得到下表：

rank	s	$\overline{d}$	Count	W1	W2
1	0.0	2.0	12	口	航道
2	0.0	2.0	11	记者	林昌
3	0.0	2.0	33	气象	预报
4	0.0	3.0	11	联合国	小组
5	0.0	2.0	13	记者	西平

rank	s	$\overline{d}$	Count	W1	W2
6	0.0	2.0	30	之	》
7	0.0	3.0	20	已	家
8	0.0	2.0	25	泽民	说
9	0.0	3.0	26	新华社	8 日
10	0.0	2.0	17	２４日	（

如果需要找距离灵活的搭配，均值方差方法是一个较为合适的方法。

3. 假设检验方法

均值方差的方法中高频率和低方差比较偶然。此时我们就要知道是否两个词一起出现是否是因为偶然。所以用到假设检验方法。

首先定义一个null hypothesis H0:除非偶然发生，两个词不会出现在同一句子中，即这两个词不会组成一个搭配。计算事件发生概率p，如果p特别小，小于显著性水平0.05，即p<0.05，则拒绝假设H0。

因为两个词的随意组合，可知两个词是相互独立的。

$$H_0 : P(w_1w_2) = P(w_1)P(w_2) = \frac{C(w_1)}{N} \times \frac{C(w_2)}{N}$$

其中N是bigram的数量。

给出t的计算式：$t = \dfrac{\overline{x} - \mu}{\sqrt{\frac{s^2}{N}}}$

其中是$\overline{x}$样本均值，s^2是样本方差，N是样本数量，μ是分布均值.

如果t值足够大时，可以拒绝假设H0，

rank	t	C(W1)	C(W2)	C(W1W2)	W1	W2
1	23.060231	3196	7253	645	这	一
2	22.314605	7253	849	529	一	种
3	22.090207	1952	806	496	两	国
4	21.418616	1467	700	464	本报	讯
5	21.085800	602	451	446	江	泽民

rank	t	C(W1)	C(W2)	C(W1W2)	W1	W2
6	20.896126	1364	1781	449	北京	1 月
7	20.778731	7253	2516	521	一	年
8	20.485805	3196	9819	569	这	是
9	20.148402	1008	1175	412	据	新华社
10	19.632228	2516	1618	406	年	来

两个词中只要一个单词频率很高，那么就会使t的值变高。

这些结果看起来不是我们关注的那些搭配，高频词汇中停用词很多，所以考虑引进停用词，把停用词去掉后，得到如下结果：

rank	t	C(W1)	C(W2)	C(W1W2)	W1	W2
1	21.84635	1952	806	496	两	国
2	21.25405	1467	700	464	本报	讯
3	21.04142	602	451	446	江	泽民
4	20.50037	1364	1781	449	北京	1 月
5	17.07914	294	369	293	附	图片
6	16.34178	1175	1364	286	新华社	北京
7	15.54849	1175	2129	271	新华社	记者
8	15.51512	1467	2129	277	本报	记者
9	15.47787	1131	926	252	领导	干部
10	14.7841	1280	355	224	改革	开放

去掉停用词后的前十名的结果看起来更有意义了一些。

4. 点对互信息方法–PMI(Pointwise Mutual Information)

引入信息论中PMI（Pointwise Mutual Information）这个指标来衡量两个事物之间的相关性（比如两个词)在概率论中，我们知道，如果x跟y不相关，则p(x,y)=p(x)p(y)。二者相关性越大，则p(x,y)就相比于p(x)p(y)越大。在y出现的情况下x出现的条件概率p(x|y)除以x本身出现的概率p(x)，自然就表示x跟y的相关程度。当对p(x)取log之后就将一个概率转换为了信息量

$$I(w_1, w_2) = log_2 \frac{p(w_1 w_2)}{p(w_1)p(w_2)}$$

rank	I	C(W_1)	C(W_2)	C(W_1W_2)	W_1W_2
1	12.86593	1	1	1	开航 黄田
2	12.86593	1	1	1	声韵 悠悠扬扬
3	12.86593	1	1	1	藤椅 吧嗒
4	12.86593	1	1	1	企获 重赏
5	12.86593	1	1	1	白云山 云台
6	12.86593	1	1	1	卷发 美容器
7	12.86593	1	1	1	昆曲 研习班
8	12.86593	1	1	1	高棉 民族党
9	12.86593	1	1	1	稳稳地 蹬立
10	12.86593	1	1	1	离石市 前瓦村

从上表中看到的全是低频词汇，由低频词组成的bigram会比由高频词组成的bigram得分更高。尽管算出来的I排序最高，但是我们应该关注那些出现频率次数高的词组，因为出现频次高，我们有更充足的理由相信两个词是搭配。

一种方法，引入bigram的频度，得到如下公式

$$I'(w_1, w_2) = C(w_1w_2)I(w_1, w_2)$$

利用上式子重新计算I，结果如下表：

rank	I'	C(W_1)	C(W_2)	C(W_1W_2)	W_1W_2
1	1612.837017	602	451	446	江 泽民
2	1269.726458	294	369	293	附 图片
3	813.877633	1467	700	464	本报 讯
4	776.980742	223	175	158	邓 小平
5	754.616183	582	205	205	钱 其琛
6	680.432075	567	209	190	周 恩来
7	651.296875	369	495	210	图片 1
8	612.437408	1952	806	496	两 国
9	571.959892	271	221	139	反 腐败
10	567.202648	1008	1175	412	据 新华社

另一种方法是过滤掉频度小于3的bigram，得到如下结果：

rank	I	$C(W_1)$	$C(W_2)$	$C(W_1W_2)$	W_1W_2
1	11.280967	3	3	3	丹参 滴丸
2	11.280967	3	3	3	胡图族 叛乱者
3	11.280967	3	3	3	波分 复用
4	11.280967	3	3	3	货仓式 自选商场
5	11.280967	3	3	3	管理课 课长
6	11.280967	3	3	3	孔雀 开屏
7	11.280967	3	3	3	诸葛 仓麟
8	11.280967	3	3	3	上虞 风机厂
9	11.280967	3	3	3	宫内 节育器
10	11.280967	3	3	3	± %

在去掉停用词后，得到的结果如下，和未去掉停用词的结果变化不大：

rank	I	$C(W_1)$	$C(W_2)$	$C(W_1W_2)$	W_1W_2
1	12.512658	3	3	3	管理课 课长
2	12.512658	3	3	3	孔雀 开屏
3	12.512658	3	3	3	胡图族 叛乱者
4	12.512658	3	3	3	波分 复用
5	12.512658	3	3	3	丹参 滴丸
6	12.512658	3	3	3	诸葛 仓麟
7	12.512658	3	3	3	上虞 风机厂
8	12.512658	3	3	3	文传 电讯社
9	12.512658	3	3	3	草浆 书写纸
10	12.512658	3	3	3	货仓式 自选商场

用点对互信息的方法优点：点对互信息是判断两个词是否独立的好测量方法，如果I接近0说明，从频率的角度来讲，两个词独立。但是，缺点也很明显，这不是一个判断两个词互相依赖的好方法。