文本分析

文本数据

	category	theme	URL	content
0	汽车	新辉腾　４．２　Ｖ８　４座加长Ｉｎｄｉｖｉｄｕａｌ版２０１１款　最新报价	http://auto.data.people.com.cn/model_15782/	经销商　电话　试驾／订车Ｕ憬杭州滨江区江陵路１７８０号４００８－１１２２３３转５８６４＃保常...
1	汽车	９１８　Ｓｐｙｄｅｒ概念车	http://auto.data.people.com.cn/prdview_165423....	呼叫热线　４００８－１００－３００　服务邮箱　ｋｆ@ｐｅｏｐｌｅｄａｉｌｙ．ｃｏｍ．ｃｎ
2	汽车	日内瓦亮相　ＭＩＮＩ性能版／概念车－１．６Ｔ引擎	http://auto.data.people.com.cn/news/story_5249...	ＭＩＮＩ品牌在二月曾经公布了最新的ＭＩＮＩ新概念车Ｃｌｕｂｖａｎ效果图，不过现在在日内瓦车展...
3	汽车	清仓大甩卖一汽夏利Ｎ５威志Ｖ２低至３．３９万	http://auto.data.people.com.cn/news/story_6144...	清仓大甩卖！一汽夏利Ｎ５、威志Ｖ２低至３．３９万＝日，启新中国一汽强势推出一汽夏利Ｎ５、威志...
4	汽车	大众敞篷家族新成员　高尔夫敞篷版实拍	http://auto.data.people.com.cn/news/story_5686...	在今年３月的日内瓦车展上，我们见到了高尔夫家族的新成员，高尔夫敞篷版，这款全新敞篷车受到了众...

文本分析

- 停用词

  - 1.语料中大量出现
  - 2.没啥大用
  - 3.留着过年嘛？

1.!	1.一下
2."	2.一个
3.#	3.一些
4.$	4.一何
5.%	5.一切
6.&	6.一则
7.'	7.一则通过
8.(	8.一天
9.)	9.一定
10.*	10.一方面
11.+	11.一旦
12.,	12.一时
13.-	13.一来
14.--	14.一样
15..	15.一次
16...	16.一片
17....	17.一番
18.....	18.一直
19.....	19.一致
20../	20.一般
21..一	21.一起

文本分析

- Tf-idf：关键词提取
  - 《中国的蜜蜂养殖》：进行词频（Term Frequency，缩写为TF）统计
  - 出现次数最多的词是----“的”、“是”、“在”----这一类最常用的词（停用词）
  - “中国”、“蜜蜂”、“养殖”这三个词的出现次数一样多，重要性是一样的？
  - "中国"是很常见的词，相对而言，"蜜蜂"和"养殖"不那么常见

文本分析

- "逆文档频率"（Inverse Document Frequency，缩写为IDF）

  - 如果某个词比较少见，但是它在这篇文章中多次出现，那么它很可能就反映了这篇文章的特性正是我们所需要的关键词

$$\text{词频(TF)} = \frac{\text{某个词在文章中的出现次数}}{\text{该文出现次数最多的词的出现次数}}$$

$$\text{逆文档频率(IDF)} = \log\left(\frac{\text{语料库的文档总数}}{\text{包含该词的文档数} + 1}\right)$$

文本分析

- Tf-idf：关键词提取

$$TF-IDF = 词频(TF) \times 逆文档频率（IDF）$$

- 《中国的蜜蜂养殖》：假定该文长度为1000个词，"中国"、"蜜蜂"、"养殖"各出现20次，则这三个词的"词频"（TF）都为0.02

- 搜索Google发现，包含"的"字的网页共有250亿张，假定这就是中文网页总数。
包含"中国"的网页共有62.3亿张，包含"蜜蜂"的网页为0.484亿张，
包含"养殖"的网页为0.973亿张

文本分析

- Tf-idf：关键词提取

$$TF-IDF = 词频(TF) \times 逆文档频率（IDF）$$

	包含该词的文档数（亿）	IDF	TF-IDF
中国	62.3	0.603	0.0121
蜜蜂	0.484	2.713	0.0543
养殖	0.973	2.410	0.0482

文本分析

- 相似度


北京气象专家解释“泥雪”：长期无降水空气脏
金羊网 - 4小时前
两人合撑一把伞在雨中打车。昨天，京城迎来一场雨夹雪。记者陶冉摄 今天是春分节气，昨
中到大雪，而平原地区由于气温原因以雨夹雪为主。截至昨晚8点，城区 ...
凤凰网
搜狐
每日甘肃
搜狐
腾讯网
北国网
北京暴雪清污染京城三月飘雪好预兆【组图】
www.591hx.com - 3小时前
飞雪迎春袭北京京城今晨或现“堵城”
大洋网 - 3小时前
北京普降瑞雪银装素裹树挂景观成春日美景
艾拉家居网 - 7小时前
延庆迎春雪城区下泥雪专家称系内蒙古沙尘被卷来
凤凰网 - 9小时前
昨夜北京普降大雪道路结冰早高峰注意出行安全
张家界在线 - 11小时前
北京春分降雪空气净化专家称三月下雪很正常
腾讯网 - 11小时前

文本分析

相似度

句子A：我喜欢看电视，不喜欢看电影。
句子B：我不喜欢看电视，也不喜欢看电影。

分词：
句子A：我/喜欢/看/电视，不/喜欢/看/电影。
句子B：我/不/喜欢/看/电视，也/不/喜欢/看/电影。

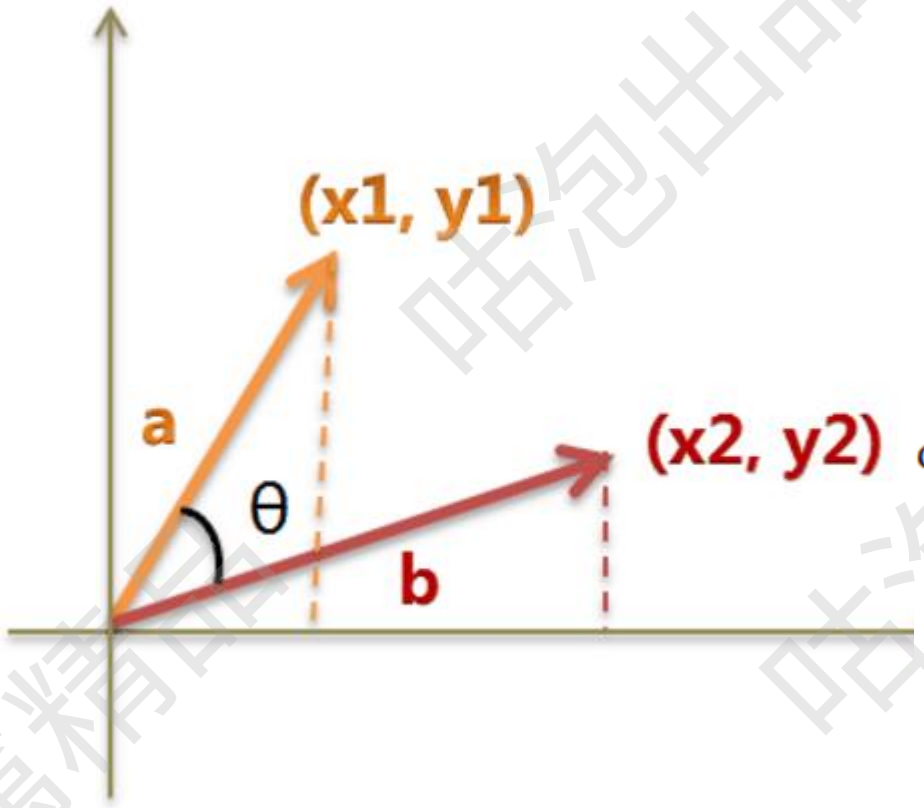
语料库：我，喜欢，看，电视，电影，不，也。

词频：
句子A：我 1，喜欢 2，看 2，电视 1，电影 1，不 1，也 0。
句子B：我 1，喜欢 2，看 2，电视 1，电影 1，不 2，也 1。

词频向量：
句子A：[1, 2, 2, 1, 1, 1, 0]
句子B：[1, 2, 2, 1, 1, 2, 1]

文本分析

- 相似度



$$\cos\theta = \frac{\sum_{i=1}^{n}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} = \frac{A \cdot B}{|A| \times |B|}$$

$$\cos\theta = \frac{1\times1+2\times2+2\times2+1\times1+1\times1+1\times2+0\times1}{\sqrt{1^2+2^2+2^2+1^2+1^2+1^2+0^2}\times\sqrt{1^2+2^2+2^2+1^2+1^2+2^2+1^2}}$$

$$= \frac{13}{\sqrt{12}\times\sqrt{16}}$$

$$= 0.938$$