# hebin

## 博客园 首页 新随笔 联系 订阅 管理

### 利用中文数据跑Google开源项目word2vec

一直听说word2vec在处理词与词的相似度的问题上效果十分好,最近自己也上手跑了跑Google开源的代码(https://code.google.com/p/word2vec/)。

### 1、语料

首先准备数据:采用网上博客上推荐的全网新闻数据(SogouCA),大小为2.1G。

从ftp上下载数据包SogouCA.tar.gz:

```
1 wget ftp://ftp.labs.sogou.com/Data/SogouCA/SogouCA.tar.gz --ftp-user=hebin_hit@foxmail.com
--ftp-password=4FqLSYdNcrDXvNDi -r
```

#### 解压数据包:

```
1 gzip -d SogouCA.tar.gz
2 tar -xvf SogouCA.tar
```

再将生成的txt文件归并到SogouCA.txt中,取出其中包含content的行并转码,得到语料corpus.txt,大小为2.7G。

```
1 cat *.txt > SogouCA.txt
2 cat SogouCA.txt | iconv -f gbk -t utf-8 -c | grep "<content>" > corpus.txt
```

随笔 - 4 文章 - 0 评论 - 16

## 公告

昵称:hebin

园龄:4年11个月

粉丝:11 关注:10 +加关注

<		201	.7年1	2月		>
日		_=_	=	四	五	六
26	27	28	29	30	1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31	1	2	3	4	5	6

搜索		

### 2、分词

分词结果:

用ANSJ对corpus.txt进行分词,得到分词结果resultbig.txt,大小为3.1G。

分词工具ANSJ参见 http://blog.csdn.net/zhaoxinfan/article/details/10403917 在分词工具seg\_tool目录下先编译再执行得到分词结果resultbig.txt,内含426221个词,次数总计572308385个。

## 常用链接

我的随笔 我的评论

我的参与

最新评论 我的标签

## 随笔分类

博士记事(1)

机器学习(2)

日常学习(1)

## 随笔档案

2015年10月 (1)

2014年1月 (1)

2013年4月 (2)

## 最新评论

1. Re:利用中文数据跑Google...

能给解释一下最后聚类那块,具体是怎么做的么?因为比如说"北京体育大学"在分词的那个步骤的时候就已经被分成 北京 体育 大学 三个词了,后来聚类的时候如果用向量来聚类是怎么组合这三个词的向

量.....

--香港胖仔

2. Re:利用中文数据跑Google... 不会编码,不会分词,能不能分享一下训练好的模型 谢谢

--急需帮助

3. Re:利用中文数据跑Google... @我不会编代码可以去github找 一下...

恒指 现 报 25296.17点 , 跌 446.06点 , 成交 金额 167.51亿元 ; 新浪 提示: 本文 属于 研究 报告 栏目, 仅为分析 人士对一只 华龙 网 一 重庆 商报 本报 讯 ( 记者 廖宇翔 ) 在 昨天 凌晨 过 官方 网站 讯 我们 来到 拉萨 已经 是 第四天 , 再 这 几天 里 , 5月 3日 (本周六) 22 时 重心 推介: 2.5 球 (2.098) 近 5月 7日 , 奥运 火炬手 徐真华 手持 火炬 传递 。 当日 , 北京 奥克 腾讯 音乐 讯 : [2008年 5月 15日 , 北京 , 昨天 (5月 14日 ) , 此时此刻。,所有。的。语言。都是多余。,不如。化、悲痛、为、力量、,、尽 中国 红十字 基金会 崔永元 公益 基金 原 定于 6月 上旬 启动 的 第二 著名、演员、陈、佩斯、。。(、左、)。 中新社、发。 廖攀。 摄、相关、阅 中新社 宁波 五月 十九 日电 题: 中国 再 出发 之 五: 奥运 力 中新社、北京、五月、三、十一、日、电、(、记者)、刘长忠、)、六、・、一、儿童 信报 讯 ( 记者 王菲 ) 昨天 下午 , 即 将 于 21日 远 赴 意大利 ... (图6)2008 亚洲 超级 模特 大赛 中国区 总决赛 隆重 举行 , 亻 东北网 5月 28日 讯 (记者 印蕾 ) 记者 从省 劳动部 繁华\_落\_尽\_亦\_枉然\_姓名\_: \_如\_妃\_、\_尔淳\_、\_玉莹\_、\_安茜\_来自\_: 中新网. 5月 15 日电。 14日 11 时 49分 , 黑龙江省 安达市 119 指 东吴 基金 今日 发布 公告 , 旗 下 近期 募集 成立 的 东吴 行业 轮 明天 上午 8点 至 11点 15分 , 奥运会 火炬 在 南京 境内 进行 主题 首席」记者。孙。卫东。 。昨天。在。日本。乒乓球。公开赛。首轮。比赛。中。,。 腾讯 体育 讯 北京 时间 5月 6日 凌晨 , 2008 斯诺克 世锦赛 决剥 周报、要点、深圳、、广州、成交量、大幅、上涨、、其它、城市、成交量、依然、

--kaierlong

4. Re:利用中文数据跑Google... 您好,您给的最上边那个源代码链接打不开了,可否发邮箱么》?谢

--我不会编代码

5. Re:利用中文数据跑Google... 请问一下,为什么跑出来的 corpus.txt是空的?

--cv1

# 阅读排行榜

- 1. 利用中文数据跑Google开源..
- 2. Latent Dirichlet Allocation(...
- 3. 杨锦锋师兄博士毕业答辩(2...
- 4. 博客里编公式(70)

## 评论排行榜

1. 利用中文数据跑Google开源..

## 推荐排行榜

- 1. 利用中文数据跑Google开源..
- 2. Latent Dirichlet Allocation(...

### 3、用word2vec工具训练词向量

1 nohup ./word2vec -train resultbig.txt -output vectors.bin -cbow 0 -size 200 -window 5 - negative 0 -hs 1 -sample 1e-3 -threads 12 -binary 1 &

vectors.bin是word2vec处理resultbig.txt后生成的词的向量文件,在实验室的服务器上训练了1个半小时。

### 4、分析

#### 4.1 计算相似的词:

1 ./distance vectors.bin

./distance可以看成计算词与词之间的距离,把词看成向量空间上的一个点,distance 看成向量空间上点与点的距离。

下面是一些例子:

Enter	word or	sentence	(EXI	T to	break)	: 诺基	NV.		
Word:	诺基亚	Position	in v	ocabu	lary:	10582			
							Word	 Cosine	distan
							 ]子	0.73678	5
							惠普		0.7301
						摩托罗		0.724859	9
							宇尔	0.702094	4
							思科		0.6922
							rim		0.6856
							东芝		0.6838
							索尼		0.6805
						S	ymbian		0.6793
							ibm		0.6791
							рc		0.6657
							sap		0.6536
						苹果公		0.648959	9
						爱立	广信	0.647333	3
						富士	∸通	0.640040	)
					V	indows	mobile		0.6397
							amd		0.6374
							华硕		0.6372
						cle	arwire		0.6332
							手机		0.6310

Enter word or sentence (EXIT to break): 苹果	
Word: 苹果 Position in vocabulary: 6007	
Word	Cosine distanc
	0.678530
iphone	0.65165
rim	0.61305
<b>基</b>	0.59689
土豆	0.59527
mobileme	0.59412
540万部	0.59341
itunes	0.59268
诺基亚	0.586824
芒果	0.58482
花生酱	0.583349
ipod	0.57747
google	0.57322
macbookair	0.57086
惠普	0.56500
htc	0.56429
盒装	0.56292
菠萝	0.56242
黄瓜	0.56208
imac	0.56057

Enter	word	or sentence	(EXIT to bre	ak): 法国	
Word:	法国	Position in	n vocabulary:	765	
				Word	Cosine distar
				 意大利	0.768896
				西班牙	0.766856
				德国	0.6969
				比利时	0.691224
				荷兰	0.6783
				捷克	0.650
				奥地利	0.647384
				波兰	0.6470
				爱尔兰	0.646253
				瑞典	0.6400
				英国	0.6371
				加拿大	0.634669
				瑞士	0.6260
				俄罗斯	0.622869
				匈牙利	0.606673
				葡萄牙	0.606487
				乌克兰	0.600848
				31904	0.6000
				摩洛哥	0.599123
				加纳	0.5985
				新西兰	0.596644

Enter word or sentence (EXIT to break): 上海 Word: 上海 Position in vocabulary: 202 Cosine dista Word 杭州 0.675 南京 0.635 深圳 0.622 6020.00 0.613 70班 0.598 沈阳 0.593 西安 0.589 永康路 0.583992 0.576 0.572 襄阳南路 0.570747 0.569 水荫路 0.564981 0.564 0.561 1154816.200.00 0.561 0.555 6863090.007126953.13 保叔路 0.553127 天津 0.553 梅川 0.552 0.550

```
Enter word or sentence (EXIT to break): 共产党
Word: 共产党 Position in vocabulary: 8334
                                                     Cosine dista
                                           Word
                                            党
                                                            0.697
                                 中国共产党
                                                    0.685726
                                                    0.680111
                                                    0.636323
                                                            0.632
                                   共产国际
                                                    0.618701
                                      青年团
                                                    0.616045
                                                    0.613440
                                    十月革命
                                                    0.612961
                                                    0.612714
                           新民主主义革命
                                                    0.607131
                                                    0.605031
                                                            0.602
                                                            0.598
                                   无产阶级
                                                    0.595561
                                                            0.594
                                                            0.592
                                                    0.591317
                                                            0.586
                                                    0.585549
                                                    0.585427
```

Enter	word or	sentence	(EXIT to b	reak): 宝	马			
1	<b>≐</b>			21.12				
word:	玉与 Po:	sition in	vocabular	y: 3143				
					Word		Cosine	dista
					本田			0.750
					丰田			0.7089
					雷诺			0.707
					奥迪			0.703
					奔驰			0.6883
					modus			0.667
					2008款			0.661
				<b>*</b>	每甘娜	C	.654732	2
					bmw			0.6543
					捷达			0.644
					gt			0.634
				1	呆时捷	C	.632410	)
					koleos			0.632
				225	90英镑			0.632
					奇瑞			0.630
				劳斯	莱斯	C	.62875	5
					车型			0.626
					铃木	C	.626353	3
				Ī	<b>敂篷车</b>	C	.624669	9
					车系			0.624

13/13	1 XXXXIII B C C C C C T III X X X X X X X X X X X X X X X X X	1311
Enter word or sentence (EXIT to break): ${}^{\complement}$	呼伦贝尔	
Word: 呼伦贝尔 Position in vocabulary: 5	53957	
	Word	Cosine distance
	内蒙古	0.605938
	鄂伦春	0.604082
克什克	[腾旗	0.574921
新巴尔虎		0.574637
	乌海	0.574201
	通辽	0.570246
锡林郭	<b>勒盟</b>	0.569448
	通辽市	0.568210
	响沙湾	0.567445
	正蓝旗	0.566607
	大庆	0.560020
宝	日希勒	0.551677
	海林市	0.551651
	赤峰	0.551327
	五彩湾	0.545135
	准格尔	0.543981
	穆棱市	0.543549
	高平原	0.543501
	羌塘	0.543280
	怀来县	0.540802



### 4.2 潜在的语言学规律

在对demo-analogy.sh修改后得到下面几个例子: 法国的首都是巴黎,英国的首都是伦敦, vector("法国") - vector("巴黎) + vector("英国") --> vector("伦敦")"

Enter three	words (EXIT	to break):	法国 巴黎	英国	
Word: 法国	Position in	vocabulary:	765		
Word: 巴黎	Position in	vocabulary:	3559		
Word: 英国	Position in	vocabulary:	829		
				Word	Distand
			,	 伦敦	0.5968
				纽约	0.51148
					0.481875
					0.460148
				žЩ	
				 东京	0.44092
					0.437690
				温莎	0.43604
					0.433944
				も顿	
			bankofe	ngland	0.42209
					0.420974
			墨西哥:	城	0.420085
					0.417653
			肯特	評都	0.414981
			加利福尼亚		0.413795

```
Word: 女王 Position in vocabulary: 10457
Word: 女 Position in vocabulary: 403
Word: 国王 Position in vocabulary: 3199
                                                              Distance
                                            Word
                                                              0.451989
                                       开拓者
                                                      0.407146
                                                              0.39761
                                                              0.37047
                                                              0.37029
                                       河正佑
                                                      0.369062
                                       花朱俊
                                                      0.347492
                                                              0.342962
                                    三十左右
                                                      0.342526
                                                              0.335768
                                                              0.33175
                                                              0.32570
                                        中国籍
                                                      0.322323
                                                              0.32205
                                     carapiquiba
公牛
                                                              0.32133
                                                              0.31889
```

```
Enter three words (EXIT to break): 东西南
         Position in vocabulary: 2090
Word: 东
Word: 西
         Position in vocabulary: 1237
        Position in vocabulary: 2102
Word: 南
                                                              Distanc
                                            Word
                                              北
                                                              0.54492
                                                      0.529837
                                                      0.520238
                                                      0.516953
                                                      0.507359
                                                      0.506711
                                                              0.49963
                                           以西
                                                              0.49596
                                        交汇处
                                                      0.489818
                                                      0.489263
                                                      0.488120
                                                      0.487450
                                        蓟运河
                                                      0.486349
                                        郑常庄
                                                      0.485370
                                                              0.48267
                                       4.14公里
```

```
Enter three words (EXIT to break): 学生 老师 病人
Word: 学生
           Position in vocabulary: 451
Word: 老师
           Position in vocabulary: 1011
Word: 病人 Position in vocabulary: 4974
                                            Word
                                                             Dista
                                                             0.732
                                                             0.676
                                                             0.652
                                                             0.650
                                    主治医生
                                                      0.639239
                                          打针
                                                             0.63
                                       曾利红
                                                      0.636444
                                                             0.635
                                                      0.631654
                                    北京医院
                                                      0.620679
                                                      0.620189
                                                      0.619292
                                                             0.618
                                    主治医师
                                                      0.615278
                                                             0.612
                                                      0.606430
```

#### 4.3 聚类

#### 将经过分词后的语料resultbig.txt中的词聚类并按照类别排序:

```
1 nohup ./word2vec -train resultbig.txt -output classes.txt -cbow 0 -size 200 -window 5 -
negative 0 -hs 1 -sample 1e-3 -threads 12 -classes 500 &
2 sort classes.txt -k 2 -n > classes_sorted_sogouca.txt
```

#### 例如:

119329 摩托罗拉\_122↓ 119330 耐克 122↓ 119331 尼葛洛庞蒂 122↓ 119332 诺基亚\_122↓ 119333 拍卖网\_122↓ 119334 苹果 122↓ 119335 苹果公司 122 119336 奇摩 122↓ 119337 碁 122↓ 119338 前程无忧 122↓ 119339 仁宝\_122↓ 119340 任天堂\_122↓ 119341 软银\_122↓ 119342 赛迪\_122↓ 119343 三星电子 122↓ 119344 三星公司 122↓ 119345 商业部门 122↓

http://www.cnblogs.com/hebin/p/3507609.html

```
263444 北京体育大学 291↓
263445 北京体育学院 291↓
263446 北京外国语大学 291↓
     北京舞蹈学院 291↓
263447
263448 北京物资学院 291↓
263449 北京信息工程学院_291↓
263450 北京信息职业技术学院 291↓
263451 北京印刷学院 291↓
263452 北京邮电大学 291↓
263453 北京中加学校 291↓
263454 北京中医药大学 291↓
263455 北外 2914
263456 北医 291↓
263457 北邮 291↓
263458 北语 291↓
263459 本科班 291↓
263460 本硕 291↓
263461 毕业 2914
     表演系_291↓
263462
```

#### 4.4 短语分析

先利用经过分词的语料resultbig.txt中得出包含词和短语的文件sogouca\_phrase.txt,再训练该文件中词与短语的向量表示。

```
1 ./word2phrase -train resultbig.txt -output sogouca_phrase.txt -threshold 500 -debug 2
2 ./word2vec -train sogouca_phrase.txt -output vectors_sogouca_phrase.bin -cbow 0 -size 300 -
window 10 -negative 0 -hs 1 -sample 1e-3 -threads 12 -binary 1
```

#### 下面是几个计算相似度的例子:

```
Enter word or sentence (EXIT to break): 东北 地区
Word: 东北
          Position in vocabulary: 3887
Word: 地区 Position in vocabulary: 385
                                                  Cosine dista
                                         Word
                                    小雨雪
                                                  0.765074
                                       华北
                                                         0.760
                              华北 中南部
                                                  0.757553
                              内蒙古 河套
                                                  0.751643
                              江南 中南部
                                                  0.749622
                                                         0.742
                              黄淮 西南部
                                                  0.742324
                                 西北 地区
                                                  0.740335
                              江淮 西南部
                                                  0.734904
                           川西高原 大部
                                                  0.729589
                                                  0.723757
                              黄淮 东北部
                                                  0.720520
                                 黄淮 东部
                                                  0.720274
                                                         0.716
                              贵州 中南部
                                                  0.714828
                                                         0.714
                              东北 中南部
                                                  0.711341
                              华北 东北部
                                                  0.710917
                                                  0.710592
                                                  0.709603
                                                  0.707226
```

```
Enter word or sentence (EXIT to break): 奥运 火炬
Word: 奥运 Position in vocabulary: 120
Word: 火炬 Position in vocabulary: 731
                                                      Cosine dista
                                           Word
                                   火炬 接力
                                                     0.778738
                                                            0.758
                                   奥运 圣火
                                                     0.753928
                                                            0.744
                                                     0.743820
                                                     0.734146
                                                     0.732106
                                                     0.728518
                             231名 少数医
                                                     0.723351
                                                     0.722706
                                                     0.716137
                                                            0.713
                                 境内 第30个
                                                     0.713619
                                      第138 棒
                                                            0.713
                             美景 融为
                                                     0.712487
                                    温州 第41
                                                     0.710148
                                                     0.708979
                                                     0.708147
                                桂林站
                                                     0.706491
                                                     0.703019
                                                     0.697910
```

```
Enter word or sentence (EXIT to break): 国庆 假期
Word: 国庆
          Position in vocabulary: 18035
Word: 假期
           Position in vocabulary: 5712
                                                        Cosine dista
                                              Word
                                           长假
                                                               0.739
                                         黄金周
                                                       0.738333
                                                               0.705
                                                       0.701930
                                                       0.675978
                                                       0.660498
                                                               0.645
                                  国庆 黄金周
                                                       0.640860
                                            放假
                                                               0.640
                                                       0.617687
                                                               0.61
                                                        0.616491
                                                               0.611
                                                       0.611377
                                                       0.608285
                                                       0.605670
                                                       0.601350
                                                       0.590729
                                                       0.586201
```

### 5、参考链接:

- 1. word2vec: Tool for computing continuous distributed representations of words, https://code.google.com/p/word2vec/
- 2. 用中文把玩Google开源的Deep-Learning项目 word2vec, http://www.cnblogs.com/wowarsenal/p/3293586.html

- 3. 利用word2vec对关键词进行聚
- 类, http://blog.csdn.net/zhaoxinfan/article/details/11069485

### 6、后续准备仔细阅读的文献:

- [1] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.
- [2] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of NIPS, 2013.
- [3] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. In Proceedings of NAACL HLT, 2013.
- [4] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. The Journal of Machine Learning Research, 2011, 12: 2493-2537.



http://www.cnblogs.com/hebin/p/3507609.html

评论列表

#1楼	2014-01-07 15:45 `````		
	这个好。	支持(1)	反对(0)
#2楼	[楼主] 2014-01-07 15:48 hebin		
	<u>@</u> rsync 跑了一下结果确实不错!	支持(0)	反对(0)
#3楼	2014-01-07 16:03 yhdino		
	非常感谢,最近正在研究这个。	支持(0)	反对(0)
#4楼	2014-01-08 12:51 Francis Fu		
	mark~~	支持(0)	反对(0)
# <b>5楼</b>			
	您好 请问转码是转成utf-8么? 我用word2phrase训练的,输入的是utf-8,输出是latin1的格式了,就变成乱码了= = 不知道怎么办了博主是怎么做的呢,谢谢哈~		
		支持(0)	反对(0)
#6楼	[ <u>楼主</u> ] 2014-03-17 20:21 hebin		

@ω提拉米兔℃

你为啥要输出latin1格式啊,输入输出都是utf8格式

支持(0) 反对(0)

#7楼 2015-07-29 15:25 smile\_tina

博主您好,这个SogouCA的语料下不下来,有没有别的下载链接地址,麻烦了

支持(0) 反对(0)

#8楼 [楼主] 2015-07-29 15:30 hebin

@ smile\_tina

你好,

你可以到http://www.sogou.com/labs/dl/ca.html 链接下注册下载,当时我也注册了才下载的

支持(0) 反对(0)

#9楼 2015-07-29 15:47 smile\_tina

@ hebin

恩,我试了,还是下不下来。连接不上。

支持(0) 反对(0)

#10楼 [楼主] 2015-07-29 15:49 hebin

@ smile\_tina

你试试另外几个小点的数据集行不行,我的完整版的数据删了

支持(0) 反对(0)

#11楼 2016-04-03 22:13 Janvn

嘻嘻

支持(0) 反对(0)

#12楼 2016-05-11 21:36 cy1

请问一下,为什么跑出来的corpus.txt是空的?

支持(0) 反对(0)

#13楼 2016-08-13 18:32 我不会编代码

您好,您给的最上边那个源代码链接打不开了,可否发邮箱么》?谢谢

支持(0) 反对(0)

#14楼 2016-09-23 13:31 kaierlong

@ 我不会编代码可以去github找一下

支持(0) 反对(0)

#15楼 2017-03-30 17:01 急需帮助

不会编码,不会分词,能不能分享一下训练好的模型 谢谢

支持(0) 反对(0)

#16楼 2017-04-13 22:50 香港胖仔

能给解释一下最后聚类那块,具体是怎么做的么?因为比如说 "北京体育大学" 在分词的那个步骤的时候就已经被分成 北京 体育 大学 三个词了,后来聚类的时候如果用向量来聚类 是怎么组合这三个词的向量的呢?

支持(0) 反对(0)

刷新评论 刷新页面 返回顶部

注册用户登录后才能发表评论,请 <u>登录</u> 或 <u>注册</u>, <u>访问</u>网站首页。

- 【推荐】50万行VC++源码: 大型组态工控、电力仿真CAD与GIS源码库
- 【促销】腾讯云技术升级10大核心产品年终让利
- 【推荐】高性能云服务器2折起,0.73元/日节省80%运维成本
- 【新闻】H3 BPM体验平台全面上线



#### 最新IT新闻:

- ·郭台铭详解鸿海工业互联网战略 拟分拆在上海上市
- · 扎克伯格休假照片曝光 配娃娃吃喝玩乐
- · 金刚狼死侍回归漫威, 迪士尼收购福克斯让好莱坞「变天」
- · 乐视网发布公告: 聘任刘淑青为公司总经理
- ·面试软件工程师,这些准备工作你做了吗?
- » 更多新闻...



#### 最新知识库文章:

- · 以操作系统的角度述说线程与进程
- · 软件测试转型之路
- ·门内门外看招聘
- ·大道至简,职场上做人做事做管理
- · 关于编程, 你的练习是不是有效的?
- » 更多知识库文章...

Copyright ©2017 hebin