



(12)发明专利申请

(10)申请公布号 CN 111104794 A

(43)申请公布日 2020.05.05

(21)申请号 201911353732.8

G06F 40/30(2020.01)

(22)申请日 2019.12.25

G06F 16/31(2019.01)

G06K 9/62(2006.01)

(71)申请人 同方知网(北京)技术有限公司

地址 100084 北京市海淀区清华园清华大学36区华业大厦B1410、1412、1414室

申请人 同方知网数字出版技术股份有限公司

(72)发明人 杨雷 段飞虎 吕强 印东敏

冯自强 张宏伟

(74)专利代理机构 北京天奇智新知识产权代理有限公司 11340

代理人 陈新胜

(51) Int. Cl.

G06F 40/216(2020.01)

G06F 40/289(2020.01)

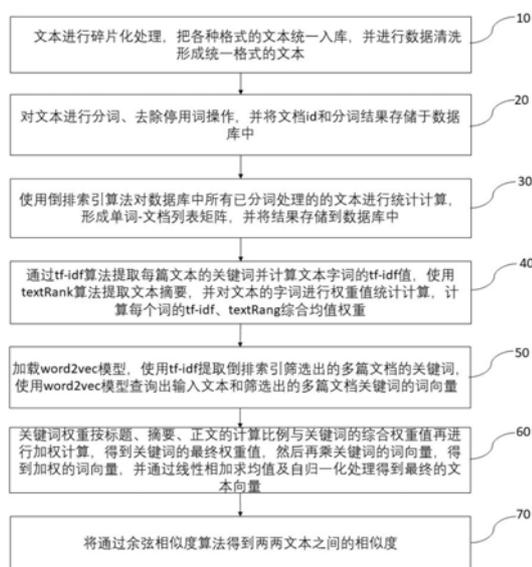
权利要求书2页 说明书6页 附图5页

(54)发明名称

一种基于主题词的文本相似度匹配方法

(57)摘要

本发明公开了一种基于主题词的文本相似度匹配方法,该方法基于倒排索引进行文档筛选,从大批量文档中精确过滤出需要的相关文档进行相似度比较,极大提升了检索效率。然后基于关键词的tf-idf、textRank综合权重值给词向量加权,用加权词向量进行文档向量计算并进行余弦相似度比较。最后对两篇相似文档,分别计算句子向量,并计算两篇文档的两两句子相似度,设置阈值以判别句子是否相似,对于句子相似度高于阈值的即判断相似并标红。本发明用于在各领域语料系统的查相似的工作,提高系统查重效率及准确率,减少人力资源浪费。



1. 一种基于主题词的文本相似度匹配方法,其特征在于,所述方法包括:

A将文本进行碎片化处理,把各种格式的文本统一入库,并进行数据清洗形成统一格式的文本;

B对文本进行分词、去除停用词操作,并将文档id和分词结果存储于数据库中;

C使用倒排索引算法对数据库中所有已分词处理的的文本进行统计计算,形成单词-文档列表矩阵,并将结果存储到数据库中;

D通过tf-idf算法提取每篇文本的关键词并计算文本字词的tf-idf值,使用textRank算法提取文本摘要,并对文本的字词进行权重值统计计算,计算每个词的tf-idf、textRang综合均值权重;

E加载word2vec模型,使用tf-idf提取倒排索引筛选出的多篇文档的关键词,使用word2vec模型查询出输入文本和筛选出的多篇文档关键词的词向量;

F关键词权重按标题、摘要、正文的计算比例与关键词的综合权重值再进行加权计算,得到关键词的最终权重值,然后再乘关键词的词向量,得到加权的词向量,并通过线性相加求均值及自归一化处理得到最终的文本向量;

G通过余弦相似度算法得到两两文本之间的相似度。

2. 如权利要求1所述的基于主题词的文本相似度匹配方法,其特征在于,所述tf-idf为用于信息检索与文本挖掘的加权技术,其中,tf为词频,idf为逆文档频率:

$$\text{词频}(TF) = \frac{\text{某个词在文章中的出现词数}}{\text{文章的总词数}}$$

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

$$TF-IDF = \text{词频}(TF) \times \text{逆文档频率}(IDF)。$$

3. 如权利要求1所述的基于主题词的文本相似度匹配方法,其特征在于,所述textRank算法公式为

$$S(v_i) = (1-d) + d \sum_{(j,i) \in \epsilon} \frac{w_{ji}}{\sum_{v_k \in \text{out}} w_{jk}} s(v_j)$$

TextRank中一个单词的权重*i*取决于与在前面的各个*j*组成的(*j*, *i*)这条边的权重,以及*j*这个点到其他边的权重之和;

将文本中的每个句子看做一个节点,如果两个句子有相似性,则认为两个句子对应的节点之间存在一条无向有权边,上述句子相似度的计算式子中,*S_i*、*S_j*两个句子,*w_k*代表句子中的单词,那么分子代表同时出现在两个句子中的单词的个数,分母是对句子中单词个数求对数之和。

4. 如权利要求1所述的基于主题词的文本相似度匹配方法,其特征在于,通过tf-idf算法对输入文本进行关键词提取,取tf-idf权重值最大的15个关键词代表文本语义,使用textRang算法进行关键词提取计算拿到关键词及其权重值,并计算其tf-idf、textRank的均值权重,通过关键词综合均值权重乘以关键词的词向量获得关键词的加权向量,使用加权向量用于相似度计算。

5. 如权利要求1所述的基于主题词的文本相似度匹配方法,其特征在于,所述步骤G中:

余弦值越接近1,就表明夹角越接近0度,也就是两个向量越相似,夹角等于0,即两个向量相等;其两两文本之间通过余弦相似度算法计算公式如下:

$$\cos(\theta) = \frac{\sum_{i=1}^n (x_i \times y_i)}{\sqrt{\sum_{i=1}^n (x_i)^2} \times \sqrt{\sum_{i=1}^n (y_i)^2}}$$

其中 $x_i y_i$ 分别代表向量 x 和 y 的分量, θ 代表两个向量的夹角。

6.如权利要求1所述的基于主题词的文本相似度匹配方法,其特征在于,所述方法号包括:对两篇相似文档,分别计算句子向量,并计算两篇文档的两两句子相似度。

一种基于主题词的文本相似度匹配方法

技术领域

[0001] 本发明涉及文本数据挖掘和计算信息处理技术领域,尤其涉及一种基于主题词的从大规模语料数据库中快速检索相似文章的文本相似度匹配方法。

背景技术

[0002] 随着计算机文本信息挖掘等各种自然语言处理应用的普及,当今社会对基于文本相似度的文档检索系统需求日益增加,同时人们对计算机文本处理也提出了更高的要求。在自然语言处理过程中,经常会涉及到如何度量两个文本之间的相似性,我们都知道文本是一种高维的语义空间,如何对其进行抽象分解,从而能够站在数学角度去量化其相似性,是此方法的重点。在相似度检索领域,现有的相似度检索方法,要么在检索效率上存在不足,要么在准确度方面不能令人满意。为解决此问题,本文研制了一套基于主题词的相似度检索查询方法。本方法首先基于倒排索引进行文档筛选,从大批量文档中精确过滤出需要的相关文档进行相似度比较,极大提升了检索效率。然后基于关键词的tf-idf、textRank综合权重值给词向量加权,用加权词向量进行文档向量计算并进行余弦相似度比较。最后对两篇相似文档,分别计算句子向量,并计算两篇文档的两两句子相似度,设置阈值以判别句子是否相似,对于句子相似度高于阈值的即判断相似并标红。

发明内容

[0003] 为解决上述技术问题,本发明的目的是提供一种基于主题词的文本相似度匹配方法。该方法首先基于倒排索引进行文档筛选,从大批量文档中精确过滤出需要的相关文档进行相似度比较,极大提升了检索效率。然后基于关键词的tf-idf、textRank综合权重值给词向量加权,用加权词向量进行文档向量计算并进行余弦相似度比较。最后对两篇相似文档,分别计算句子向量,并计算两篇文档的两两句子相似度,设置阈值以判别句子是否相似,对于句子相似度高于阈值的即判断相似并标红。

[0004] 本发明的目的通过以下的技术方案来实现:

[0005] 一种基于主题词的文本相似度匹配方法,包括:

[0006] A将文本进行碎片化处理,把各种格式的文本统一入库,并进行数据清洗形成统一格式的文本;

[0007] B对文本进行分词、去除停用词操作,并将文档id和分词结果存储于数据库中;

[0008] C使用倒排索引算法对数据库中所有已分词处理的文本进行统计计算,形成单词-文档列表矩阵,并将结果存储到数据库中;

[0009] D通过tf-idf算法提取每篇文本的关键词并计算文本字词的tf-idf值,使用textRank算法提取文本摘要,并对文本的字词进行权重值统计计算,计算每个词的tf-idf、textRang综合均值权重;

[0010] E加载word2vec模型,使用tf-idf提取倒排索引筛选出的多篇文档的关键词,使用word2vec模型查询出输入文本和筛选出的多篇文档关键词的词向量;

[0011] F关键词权重按标题、摘要、正文的计算比例与关键词的综合权重值再进行进行加权计算,得到关键词的最终权重值,然后再乘关键词的词向量,得到加权的词向量,并通过线性相加求均值及自归一化处理得到最终的文本向量;

[0012] G通过余弦相似度算法得到两两文本之间的相似度。

[0013] 与现有技术相比,本发明的一个或多个实施例可以具有如下优点:

[0014] 用于在各领域语料系统的查相似的工作,提高系统查重效率及准确率,减少人力资源浪费。

附图说明

[0015] 图1是基于主题词的文本相似度匹配方法流程图;

[0016] 图2是文本碎片化后入库示意;

[0017] 图3是tf-idf、textRank及其均值权重计算结果示意图;

[0018] 图4是word2ve词向量示意图;

[0019] 图5是基于主题词的文本相似度匹配方法原理示意图;

[0020] 图6是相似文本集展示图;

[0021] 图7是相似文本中相似语句标红显示示意图。

具体实施方式

[0022] 为使本发明的目的、技术方案和优点更加清楚,下面将结合实施例及附图对本发明作进一步详细的描述。

[0023] 如图1所示,为基于主题词的文本相似度匹配方法流程,包括以下步骤:

[0024] 步骤10将文本进行碎片化处理,把各种格式的文本统一入库,并进行数据清洗形成统一格式的文本;

[0025] 步骤20对文本进行分词、去除停用词操作,并将文档id和分词结果存储于数据库中;

[0026] 步骤30使用倒排索引算法对数据库中所有已分词处理的文本进行统计计算,形成单词-文档列表矩阵,并将结果存储到数据库中;

[0027] 步骤40通过tf-idf算法提取每篇文本的关键词并计算文本字词的tf-idf值,使用textRank算法提取文本摘要,并对文本的字词进行权重值统计计算,计算每个词的tf-idf、textRang综合均值权重即二者相加除2,将提取的关键词、文本摘要和三个权重值的计算结果存入数据库中;

[0028] 步骤50加载word2vec模型,使用tf-idf提取倒排索引筛选出的多篇文档的关键词,使用word2vec模型查询出输入文本和筛选出的多篇文档关键词的词向量;

[0029] 步骤60关键词权重按标题、摘要、正文的计算比例与关键词的综合权重值再进行进行加权计算,得到关键词的最终权重值,然后再乘关键词的词向量,得到加权的词向量,并通过线性相加求均值及自归一化处理得到最终的文本向量;

[0030] 步骤70通过余弦相似度算法得到两两文本之间的相似度。

[0031] 上述步骤10中将文本进行碎片化处理,把各种格式的文本统一入库,并进行数据清洗形成统一格式的文本。将word,pdf等格式的文本通过程序识别,统一格式并保存在数

数据库内。下表1为数据库统一结构,其中属性f_article_title为每篇文本标题,f_after_content为去除html标签的文本全文,本算法主要使用属性f_after_content的全文信息。

[0032] 表1

	id	int	11	0
	f_page_url	text	0	0
	f_page_title	text	0	0
	f_fetchtime	varchar	255	0
	f_job_name	varchar	255	0
	f_content	longtext	0	0
	f_article_title	varchar	255	0
	f_article_content	longtext	0	0
	f_article_author	varchar	255	0
	f_article_publisher	varchar	255	0
	f_article_time	varchar	20	0
[0033]	f_article_introduction	text	0	0
	f_article_subtitle	varchar	255	0
	f_article_tags	varchar	2500	0
	f_res_site	varchar	255	0
	f_res_column	varchar	255	0
	f_language	varchar	20	0
	f_location	varchar	255	0
	f_file	varchar	255	0
	f_after_content	longtext	0	0
	readNum	int	11	0
	SimHash	char	64	0
	after content	lonatext	0	0

[0034] 文本格式化后入库后内容如图2所示。

[0035] 上述步骤20,使用分词器(jieba分词、Hanlp分词等)对文本进行分词、去除停用词(分词使用通用的分词算法即可,用户词典及停用词词典根据用户自己的需求选择专业领域词典或者通用词典),用户可以指定自己自定义的词典,以便包含分词器词库里没有的词,虽然分词器有新词识别能力,但是自行添加新词可以保证更高的正确率,以保证将一些专业领域词汇准确切分出来。停用词是指在信息检索中,为节省存储空间和提高搜索效率,在处理自然语言数据(或文本)之前或之后会自动过滤掉某些没有实际意义的并且常用的名词、动词、语气助词等。

[0036] 上述步骤30基于倒排索引原理对分词进行倒排索引操作,倒排索引是实现“单词-文档矩阵”的一种具体存储形式,通过倒排索引,可以根据单词快速获取包含这个单词的文档列表。倒排索引主要由两个部分组成:“单词词典”和“倒排文件”。使用倒排索引查找相关文档,对于大规模语料数据库检索效率有极大的提升。基于倒排索引原来编写程序将倒排索引表存入数据库,在使用系统查找相似文本时,直接读取所有关键词的所在翁当的id列表。新增语料时需同步更新倒排索引表图下表2。

[0037] 表2

走失	4412,
死档	2531,2294,
得胜	3354,
无精打彩	405,
电传	1436,1421,1465,998,1033,1044,1088,1709,1073,
报告书	1375,1472,
实话	18,549,
难以言状	2313,
篮球队	43,
复发性	4608,4543,
百奥维达	4502,
车行	2985,
坐车	3649,4690,4660,2980,3994,3441,3045,
叶黄素	4045,
同盟国	1557,1554,1375,1472,1759,
[0038] 实收	2405,
优惠	3244,3242,2029,3116,752,2300,3476,3874,4601,3875,3754,2707,3634,3877,3635,3911,3757,3518,3879,3092,4344,2044,2043,;
虔诚	2889,
当特雷泽盖	109,
把脉	2473,
扰	3547,3502,4045,3808,1561,3483,
计算机	4172,4452,3761,3088,1183,1182,2427,2701,2028,2424,1697,2026,2388,997,1210,1739,2707,2828,2948,4847,1339,1856,2945,;
栗子	4721,3962,
接考	2454,2453,2715,
犹豫	4110,3189,2396,3063,2150,273,2009,2107,2645,2149,315,2365,316,2807,4888,4627,3170,3130,2480,2236,2313,2599,1520,21;
拉票	2840,
淮海	1046,
车船税	4601,
杭州华夏医院	4241,
拍腿	794,4712,
商业局	4954,

[0039] 上述tf-idf是一种用于信息检索 (information retrieval) 与文本挖掘 (text mining) 的常用加权技术,用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加,但同时会随着它在语料库中出现的频率成反比下降。如图3为tf-idf、textRank及其两者均值的计算结果,并存入数据库表中;其核心思想是:如果某个单词在一篇文章中出现的频率TF高,并且在其他文章中很少出现,则认为此词或者短语具有很好的类别区分能力,适合用来分类。

$$[0040] \quad \text{词频}(TF) = \frac{\text{某个词在文章中的出现词数}}{\text{文章的总词数}}$$

$$[0041] \quad \text{逆文档频率}(IDF) = \log\left(\frac{\text{语料库的文档总数}}{\text{包含该词的文档数} + 1}\right)$$

$$[0042] \quad \text{TF-IDF} = \text{词频}(TF) \times \text{逆文档频率}(IDF)$$

[0043] TextRank基于算法PageRank而来,用TextRank提取来提取关键词,用PageRank的思想来解释它:如果一个单词出现在很多单词后面的话,那么说明这个单词比较重要一个TextRank值很高的单词后面跟着的一个单词,那么这个单词的TextRank值会相应地因此而提高。这样TextRank的公式就可以由PageRank公式改写为:

$$[0044] \quad S(v_i) = (1-d) + d \sum_{(j,i) \in \mathcal{E}} \frac{w_{ji}}{\sum_{v_k \in \text{out}} w_{jk}} S(v_j)$$

[0045] TextRank中一个单词的权*i*重取决于与在前面的各个点*j*组成的(*j*, *i*)这条边的权重,以及*j*这个点到其他其他边的权重之和。使用tf-idf和textRank综合权重是为了平衡两个算法统计的误差,提高关键词权重的准确性,不会因为某个词的tf-idf或textRank值过大或过小从而影响这个词的真实权重,从而提高词向量的准确性。将文本中的每个句子看做一个节点,如果两个句子有相似性,则认为两个句子对应的节点之间存在一条无向有权边。句子相似度的计算式子如上所示,*S_i*、*S_j*两个句子,*w_k*代表句子中的单词,那么分子代表同时出现在两个句子中的单词的个数,分母是对句子中单词个数求对数之和。分母使用对

数可以抵消长句子在相似度计算上的优势(长句子包含相同单词的可能性更高)。根据以上相似度公式循环迭代计算得到任意两个节点之间的相似度,构建节点连接图,最后计算PR值,经排序选出PR值最高的的节点对应的句子作为摘要。

[0046] 基于word2vec算法,使用大规模语料进行词向量模型训练,得到word2vec模型,将模型放到服务器指定路径保存。一个单词的词向量可以表示单词的语义,本文使用从文本中提取的关键字的词向量代表文本语义,使用基于加权线性求均值的方法得到文本向量,使用文本向量去计算相似度。

[0047] 对输入文本使用tf-idf算法进行关键词提取,取其tf-idf值最大的15个词,本文主要针对短文本相似度计算,所以这里取15个权重值最大的关键词来代表文本语义。并使用textRang算法进行关键词提取计算拿到关键词及其权重值,并计算其tf-idf、textRank的均值权重(关键词的textRank值和tfidf值相加除2)。此步骤是为了后续计算文本向量做准备,这里计算的关键词综合均值权重乘以关键词的词向量就得到了关键词的加权向量,使用加权向量用于相似度计算。

[0048] 使用主题词表以及哈工大同义词词林扩展版对属于主题词同义、近义范畴内的关键词替换为主题词,通过对所有文本的关键词与主题词进行统一,达到词语消歧、词语统一的作用,进而实现文本语义消歧的作用,以提高文本相似度的准确性。针对专业领域的语料库,使用专业领域的主题词,对属于专业主题词表的这专业名词的词向量权重赋予更高的值,会提升系统对专业名词、术语的语义理解。

[0049] 使用输入文本提取的15个关键词通过倒排索引表拿到包含这样关键词的文档id,并进行取交集操作,拿到包含关键词最多的50篇文档用于相似度比较。对拿到的50篇文档使用主题词表进行主题词替换,从而使得输入文本的关键词和比较文本同义或近义的关键词统一为同样的主题词,从而提高相似度计算的准确性。

[0050] 使用word2vec算法对大规模通用语料(如百度百科、维基百科、知乎、微博等语料)进行模型的训练,然后用已入库的业务语料对word2vec模型进行增量训练,这样便可以兼顾通用性及专业性的需求。一般认为词向量维度为200或者300维即可,训练完毕,保存模型。词向量的概念是将word映射到一个新的空间中,并以多维的连续实数向量进行表示叫做“Word Representation”或“Word Embedding”。自从21世纪以来,人们逐渐从原始的词向量稀疏表示法过渡到现在的低维空间中的密集表示。用稀疏表示法在解决实际问题时经常会遇到维数灾难,并且语义信息无法表示,无法揭示word之间的潜在联系。而采用低维空间表示法,不但解决了维数灾难问题,并且挖掘了word之间的关联属性,从而提高了向量语义上的准确度。Distributed representation的关键点在于,将高维空间中的词汇映射到一个低维的向量空间中,并且让相关或者相似的词,在距离上更接近。我们这里说的词向量是在词粒度进行分析,当然我们也可以在字粒度的字向量、句子粒度的句向量以及文档粒度的文档向量进行表示分析。如图4所示为tf-idf、textRank及其两者均值的计算结果,并存入数据库表中。

[0051] 加载word2vec模型,使用tf-idf提取倒排索引筛选出的50篇文档的关键词,使用word2vec模型查询出输入文本和筛选出的50篇文档关键词的词向量。统计关键字是否属于标题、摘要,若关键词属于标题说明它在文本中重要性更高,若属于摘要,重要性次等重要,只属于正文重要性一般,关键词权重按(标题0.5、摘要0.3、正文0.2)的计算比例与关键词

的综合权重值再进行进行加权计算,得到关键词的最终权重值,然后再乘关键词的词向量,得到加权的词向量,并通过线性相加求均值及自归一化处理得到最终的文本向量。通过余弦相似度算法得到两两文本之间的相似度。余弦值越接近1,就表明夹角越接近0度,也就是两个向量越相似,夹角等于0,即两个向量相等,这就叫“余弦相似性”。余弦距离使用两个向量夹角的余弦值作为衡量两个个体间差异的大小。相比欧氏距离,余弦距离更加注重两个向量在方向上的差异。

$$[0052] \quad \cos(\theta) = \frac{\sum_{i=1}^n (x_i \times y_i)}{\sqrt{\sum_{i=1}^n (x_i)^2} \times \sqrt{\sum_{i=1}^n (y_i)^2}}$$

[0053] 其中 $x_i y_i$ 分别代表向量 x 和 y 的分量, θ 代表两个向量的夹角。

[0054] 对文本以标点符号(。|,|:|;|!|?)切分得到句子,对句子进行分词,并对字词使用主题词表进行同义词、近义词替换。并通过词向量结合综合权重值得到加权句向量,对相似的两篇文本分别计算其句向量的余弦相似度。设置阈值(一般设为85%即可),句子相似度高于阈值即判定这两个句子是相似的,予以标红显示。

[0055] 图5是基于主题词的文本相似度匹配方法原理示意图;图6是相似文本集展示图;图7是相似文本中相似语句标红显示示意图。

[0056] 虽然本发明所揭露的实施方式如上,但所述的内容只是为了便于理解本发明而采用的实施方式,并非用以限定本发明。任何本发明所属技术领域内的技术人员,在不脱离本发明所揭露的精神和范围的前提下,可以在实施的形式上及细节上作任何的修改与变化,但本发明的专利保护范围,仍须以所附的权利要求书所界定的范围为准。

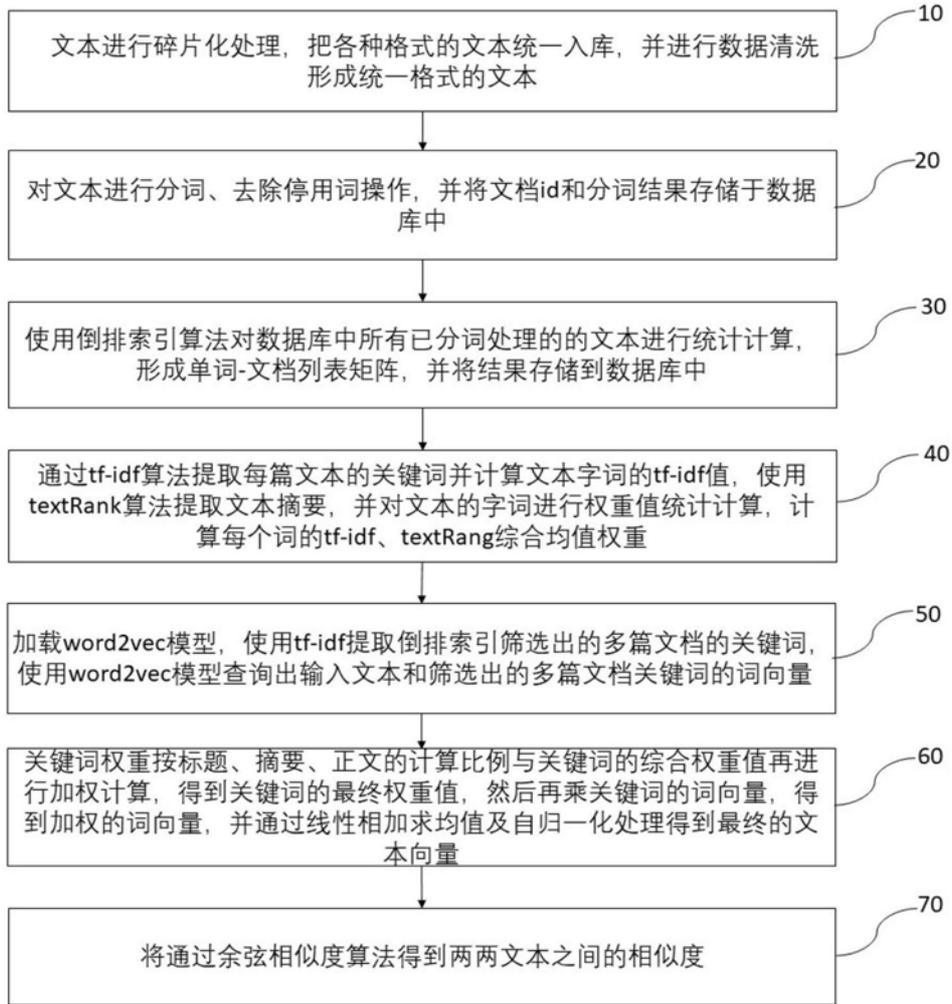


图1

id	f_page_url	f_page_title	f_fetchtime	f_job_name	f_content	f_article_title	f_article_content	f_article_author	f_article_publisher
10052	http://www.jsnc.com	(Null)	2017-05-25 18:29:57	江苏海洋与渔业	(Null)	高淳区农业与青松水产合作社一行<!--ZJEG_RSS.content.begin--><meta name="ContentStart"><p			(db:来源)
9940	http://www.jsnc.com	(Null)	2017-05-25 18:25:48	江苏海洋与渔业	(Null)	高邮第一起非法捕捞水产品罪案判决书<!--ZJEG_RSS.content.begin--><meta name="ContentStart"><p			(db:来源)
9939	http://www.jsnc.com	(Null)	2017-05-25 18:25:47	江苏海洋与渔业	(Null)	曹海海洋渔业协会参加南通市农业<!--ZJEG_RSS.content.begin--><meta name="ContentStart"><p			(db:来源)
3	http://www.jsnc.com	(Null)	2017-05-25 09:18:20	江苏农业网	(Null)	我省组织参加首届中国茶叶博览会5月18-21日,首届中国国际茶叶博览会在杭州国际博览中心举办,本			(db:来源)
5	http://www.jsnc.com	(Null)	2017-05-25 09:18:20	江苏农业网	(Null)	南京市江宁区汤山街道“渔天芳歌”特色小镇建设,绿树掩映,美丽乡村,一个个小花篮点缀村中,元		江宁新闻	(db:来源)
7	http://www.jsnc.com	(Null)	2017-05-25 09:18:21	江苏农业网	(Null)	扬州市召开小冀新品种新技术及水稻推广为了及早谋划秋播小冀新品种应用,新技术推广,加		扬州农委	(db:来源)
9938	http://www.jsnc.com	(Null)	2017-05-25 18:25:47	江苏海洋与渔业	(Null)	宝应县水产局开展“5.10”爱鱼爱水宣传<!--ZJEG_RSS.content.begin--><meta name="ContentStart"><p			(db:来源)
10	http://www.jsnc.com	(Null)	2017-05-25 09:18:21	江苏农业网	(Null)	舒颜县防汛抗旱指挥部综合科5月23日下午,舒颜县出台了防汛抗旱暨春耕备耕暨综合科利用沈少		舒颜县农委	(db:来源)
13	http://www.jsnc.com	(Null)	2017-05-25 09:18:22	江苏农业网	(Null)	邳州市四户镇立足生态农业创建建群5年以来,邳州市四户镇在农委包挂帮扶工作队的指导和建		邳州市农委	(db:来源)
15	http://www.jsnc.com	(Null)	2017-05-25 09:18:22	江苏农业网	(Null)	邳州大港通过国家生态原产地保护认证 近日,国家生态原产地保护专家一行来我市,开展邳州大		邳州日报	(db:来源)
17	http://www.jsnc.com	(Null)	2017-05-25 09:18:23	江苏农业网	(Null)	邳州市沭河镇大力发展畜牧特色产业 玫瑰蟹与美于一身,受到人们的青睐。在邳州市沭河镇310		胡敬	(db:来源)
19	http://www.jsnc.com	(Null)	2017-05-25 09:18:23	江苏农业网	(Null)	金坛创新休闲农业特色鲜明 5月21—22日,全省休闲农业工作推进会在常州市金坛区召开,曹		金坛区农委	(db:来源)
24	http://www.jsnc.com	(Null)	2017-05-25 09:18:24	江苏农业网	(Null)	淮安野龙虾小镇奠基开工 23日上午,野龙虾小镇奠基开工仪式在野龙虾小镇隆重举行,标		淮安日报	(db:来源)
26	http://www.jsnc.com	(Null)	2017-05-25 09:18:25	江苏农业网	(Null)	淮安市淮安区部署春耕备耕“两禁”应用 5月22日,淮安市淮安区及贾汪镇市政府关于夏季		淮安政府网站	(db:来源)
10056	http://www.jsnc.com	(Null)	2017-05-25 18:29:57	江苏海洋与渔业	(Null)	省厅对海头镇头岛良种进行水产养殖<!--ZJEG_RSS.content.begin--><meta name="ContentStart"><p			(db:来源)
28	http://www.jsnc.com	(Null)	2017-05-25 09:18:26	江苏农业网	(Null)	南京江宁区启动现代农业插上科技的“农业要发展,要有持续的生命力,必须和科技相结合,只		江宁新闻	(db:来源)
30	http://www.jsnc.com	(Null)	2017-05-25 09:18:27	江苏农业网	(Null)	南京江宁区“大数据+”走出现代农业新路子 日前,全国人大常委、省政协副主席、省政		江宁新闻	(db:来源)
32	http://www.jsnc.com	(Null)	2017-05-25 09:18:27	江苏农业网	(Null)	常州市农委组织参加首届中国国际茶叶博览会 5月18日,首届中国国际茶叶博览会(以下简		常州市农委	(db:来源)
35	http://www.jsnc.com	(Null)	2017-05-25 09:18:28	江苏农业网	(Null)	泰州市海陵区创新农业经营主体建设 按照泰州市委市史市市长曹海陵在城西种养合作社提出		泰州市农委	(db:来源)
38	http://www.jsnc.com	(Null)	2017-05-25 09:18:29	江苏农业网	(Null)	泰州市海陵区农民创业人数稳步提高 打造优良创业环境,激发全民创业热情,在全市城乡形		泰州市农委	(db:来源)
40	http://www.jsnc.com	(Null)	2017-05-25 09:18:30	江苏农业网	(Null)	常州市武进区推进农业供给侧结构性改革 根据常州市武进区农业局《2017年武进区农业系统		常州市农委	(db:来源)
41	http://www.jsnc.com	(Null)	2017-05-25 09:18:31	江苏农业网	(Null)	徐州市全面启动夏季秸秆综合利用工作 记者从徐州市秸秆综合利用办公室获悉,5月19日,徐		中国徐州网	(db:来源)
10054	http://www.jsnc.com	(Null)	2017-05-25 18:29:54	江苏海洋与渔业	(Null)	浮标设置到位 南海海洋牧场建设推进<!--ZJEG_RSS.content.begin--><meta name="ContentStart"><p			(db:来源)
10055	http://www.jsnc.com	(Null)	2017-05-25 18:29:57	江苏海洋与渔业	(Null)	南通海洋牧场建设工作启动<!--ZJEG_RSS.content.begin--><meta name="ContentStart"><p			(db:来源)
44	http://www.jsnc.com	(Null)	2017-05-25 09:18:31	江苏农业网	(Null)	常州市武进区推进“三夏”生产 暨正值三夏大忙季节,为切实做好今年常州市武进区		武进日报	(db:来源)
9937	http://www.jsnc.com	(Null)	2017-05-25 18:29:47	江苏海洋与渔业	(Null)	省厅召开2017年度“一村一品一业”暨<!--ZJEG_RSS.content.begin--><meta name="ContentStart"><p			(db:来源)
47	http://www.jsnc.com	(Null)	2017-05-25 09:18:32	江苏农业网	(Null)	沭阳县召开2017年度“一村一品一业”暨<!--ZJEG_RSS.content.begin--><meta name="ContentStart"><p			(db:来源)
9936	http://www.jsnc.com	(Null)	2017-05-25 18:25:46	江苏海洋与渔业	(Null)	曹海海洋渔业协会参加南通市农业<!--ZJEG_RSS.content.begin--><meta name="ContentStart"><p			(db:来源)
9933	http://www.jsnc.com	(Null)	2017-05-25 18:25:44	江苏海洋与渔业	(Null)	加强自身建设,做好太湖网围堰建设<!--ZJEG_RSS.content.begin--><meta name="ContentStart"><p			(db:来源)
9934	http://www.jsnc.com	(Null)	2017-05-25 18:25:45	江苏海洋与渔业	(Null)	中国海监211舰开展休渔执法检查<!--ZJEG_RSS.content.begin--><meta name="ContentStart"><p			(db:来源)
9935	http://www.jsnc.com	(Null)	2017-05-25 18:25:46	江苏海洋与渔业	(Null)	“两沙”2017年度海域使用征收有序<!--ZJEG_RSS.content.begin--><meta name="ContentStart"><p			(db:来源)
70	http://www.jsnc.com	(Null)	2017-05-25 09:18:39	江苏农业网	(Null)	江苏2017年农产品出口额 农产品出口额<!--ZJEG_RSS.content.begin--><meta name="ContentStart"><p			(db:来源)

图2

id	textRank	tf-idf	vsm
4	[100=0.5395906, 体育=0.52069634]	[进球=0.0061892685, 阅兵=0.007184829, 物体=0.0055075907, 功开=0.008138911, 小男孩=0.00669	
5	[体育=0.52069634, 1000=0.671211]	[23=0.02510621, 郑身=0.037650246, 决赛=0.020171294, 法国=0.0237239, 阿森纳=0.069	
6	[101=1.2449793, 体育=0.52069634]	[带走=0.015824689, 赛后=0.008788655, 状态=0.005609255, 素质=0.015885139, 08=0.01]	
7	[102=1.2074866, 体育=0.52069634]	[带走=0.0058840066, 正式=0.0066113803, 决赛=0.0028416836, 皮雷=0.006919339, 记忆]	
8	[103=1.2319465, 体育=0.52069634]	[体能=0.0046210294, 巴西=0.004893921, 达成=0.0038524568, 打击=0.003096256, 巴西]	
9	[104=1.4980369, 体育=0.52069634]	[赛后=0.008843584, 专家=0.009726891, 比赛=0.014590337, 打出=0.009873647, 队员=0]	
10	[体育=0.52069634, 105=1.1569337]	[进球=0.013101142, 滋味=0.011398542, 南京站=0.03266748, 激情=0.008667869, 0]	
11	[体育=0.52069634, 106=1.3074486]	[决赛=0.020593287, 赛后=0.017761173, 见状=0.012535873, 弄走=0.010011284, 状态=0]	
12	[体育=0.52069634, 107=1.1963484]	[失望=0.012769786, 决赛=0.008544785, 付海峰=0.03057496, 提高=0.0041809725, 女队]	
13	[体育=0.52069634, 108=1.185339]	[林斌=0.005478394, 郭晶晶=0.022191358, 最终=0.02914131, 规则=0.027987057, 10分]	
14	[体育=0.52069634, 109=1.3473026]	[大本钟=0.015696326, 正式=0.0046548415, 决赛=0.00600219, 2003=0.012475605, 后]	
15	[11=0.5, 体育=0.52069634]	[教练员=0.053410817, 聚焦=0.042314976, 护卫队=0.06372936, 记者=0.011095843, 驻地]	
16	[110=0.881714, 体育=0.52069634]	[赛前=0.02710429, 记者=0.015146708, 状态=0.0143347625, 教练=0.022720061, 汤杯=0]	
17	[111=1.3473026, 体育=0.52069634]	[赛后=0.01413419, 赛后=0.024737297, 12分=0.015702887, 24分=0.007910668, 赛后=0]	
18	[112=1.3266063, 体育=0.52069634]	[犯规=0.025772674, 筛选=0.017698763, 之间=0.007042484, 履约=0.024967283, 伊朗纳什]	
19	[113=1.1760913, 体育=0.52069634]	[链子=0.02476094, 决赛=0.020337999, 状态=0.022390662, 素质=0.010568212, 大半=0.01]	
20	[114=1.2319465, 体育=0.52069634]	[参赛者=0.009407915, 老说=0.0103798285, 招募=0.007464088, 突出=0.0084351525, 新]	
21	[115=1.1760913, 体育=0.52069634]	[网址=0.03515148, 本场=0.029169083, 状态=0.032253217, 过往=0.0446355, 比赛=0.01]	
22	[体育=0.52069634, 116=1.2449793]	[尤伯=0.05904123, 决赛=0.011288522, 残缺=0.053512912, 1948年=0.025140941, 1956年=0.02367	
23	[体育=0.52069634, 117=1.259257]	[带走=0.01507559, 决赛=0.0072807632, 付海峰=0.019539028, 创纪录=0.0072169607, 状态=0.0080	
24	[体育=0.52069634, 118=1.2738874]	[助理=0.017374836, 意识到=0.009989614, 状态=0.0050735394, 比赛=0.013114909, 安慰]	
25	[体育=0.52069634, 119=1.3473026]	[田间=0.009405964, 盾牌=0.010953679, 道理=0.0131120635, 纵队=0.013080272, 外教=0.0118555	
26	[12=0.5, 体育=0.52069634]	[教练员=0.054682504, 聚焦=0.043322477, 护卫队=0.06524673, 记者=0.011360031, 驻地]	
27	[体育=0.52069634, 120=0.8266063]	[决赛=0.005444643, 判断=0.0059399237, 梅开二度=0.010238351, 终结=0.009151265, 仍]	
28	[121=1.259257, 体育=0.52069634]	[绝大多数=0.0056821303, 履约=0.009920774, 直白=0.009074048, 无所事事=0.009920774]	
29	[122=1.2904625, 体育=0.52069634]	[天型=0.012725599, 随便=0.00681136, 带兵=0.00965983, 专访=0.007063643, 没事=0.00]	
30	[123=1.3958453, 体育=0.52069634]	[赛后=0.01513204, 马丁=0.05069202, 赛后=0.048269354, 阿森纳特=0.020760324, 不耳]	
31	[124=1.3266063, 体育=0.52069634]	[即便是=0.0029939355, 天气=0.0027559411, 持续=0.001945888, 哨音=0.004837724, 不耳]	
32	[125=1.0449525, 体育=0.52069634]	[国奥=0.013199483, 吉尔伯特=0.021934189, 后卫线=0.007606967, 封闭=0.018245095, 算]	
33	[126=1.1662192, 体育=0.52069634]	[网址=0.033364117, 本场=0.027685907, 47=0.032347206, 卡马亚=0.054538086, 成功=0.0]	
34	[体育=0.52069634, 127=1.1760913]	[巴西=0.012324324, 处于=0.0032901864, 赛程=0.013423153, 不幸=0.005834798, 赛后=0.0086278	
35	[体育=0.52069634, 128=1.3266063]	[国奥=0.013199483, 吉尔伯特=0.021934189, 后卫线=0.007606967, 封闭=0.018245095, 算]	
36	[体育=0.52069634, 129=1.3074486]	[赛后=0.056934908, 失望=0.012949641, 1996年=0.012859682, 赛后=0.009964601, 24分]	
37	[13=0.573064, 体育=0.52069634]	[梦想=0.0083229765, 急于=0.0097449655, 充足=0.008453318, 激烈=0.0060913316, 时机]	
38	[体育=0.52069634, 130=0.9888618]	[体育=0.008398969, 之间=0.0028842934, 2003年=0.008728693, 三国演义=0.02196412, 占扁=0.00	

图3

```

365113 300
-0.233957 0.112669 -0.056303 0.787765 0.794357 -0.111734 -0.291528 0.255560 0.055221 -0.234847 0.439171 -0.712020 0.695978 -0.239523 -0.447078 -0.492325 -0.829695 -0.550252 0.037133 0.146304 -0.180787
-0.230083 0.559056 0.132101 0.287553 -0.243016 0.667487 -0.459343 0.346680 -0.145659 -0.062363 0.421874 -0.093149 -0.045562 -0.105993 -0.441210 -0.482140 -0.346894 -0.066152 0.019354 -0.537485 -0.230617
-0.776394 0.565979 -0.139537 0.175162 -0.225444 0.307338 0.718184 0.067950 0.059152 -0.061645 0.424034 0.163301 0.759910 -0.546200 -0.590429 0.682789 -0.183301 -0.879561 -0.515166 0.292980 -0.547408 0.
-0.159396 0.072122 0.373923 0.026290 -0.794223 -0.454989 -0.043291 -0.302263 0.247455 0.104558 0.412677 0.011923 0.574996 -0.395614 -0.425831 -0.449187 -0.466827 -0.337915 -0.159531 0.062118 0.020787 0.
0.176112 0.749392 -0.030075 -0.249354 -0.773159 -0.237183 0.135704 -1.129595 0.381755 -0.104459 0.710104 -0.710271 0.576238 -0.373094 -0.207124 -0.745922 -0.546462 -0.752746 -0.259303 -0.421441 -0.12396
-0.409497 0.301003 -0.249235 0.494741 -0.447097 0.301540 -0.149536 -0.409910 -0.013073 0.164311 0.321193 -0.119622 0.459977 0.770213 0.290179 -0.105410 -0.023979 -0.021449 0.302102 0.042466 0.46703
0.196104 0.590305 -0.189702 -0.363516 -0.239428 -0.136193 0.161653 0.026041 0.159490 -0.094414 -0.134712 -0.137669 0.306456 -0.030307 -0.185748 -0.030649 -0.301433 0.233715 0.021022 0.042466 0.46703
在 -0.351035 0.526678 0.002980 0.131613 0.418120 0.110507 -0.371679 -0.626415 0.038959 -0.786214 0.188354 -0.648206 0.399797 0.081853 -0.214532 -0.066159 -0.019796 0.027964 0.064746 -0.154216 0.152984 0
-0.411530 -0.126723 0.359474 0.140570 -0.500443 0.526175 -0.221960 -0.009701 -0.374612 -0.435706 0.115933 0.374939 0.134706 0.663465 -0.392194 -0.119749 -0.147765 -0.320180 -0.049322 0.049746 -0.345966
了 -0.699956 0.497795 0.314994 -0.102235 0.325907 0.629793 -0.109796 0.464907 -0.255462 -0.444937 -0.463309 -0.202990 0.514936 -0.030105 -0.263702 -0.247945 0.609425 -0.027617 0.516126 0.702091 0.130491 0.
* 0.395920 0.376994 0.738009 0.171753 0.187023 0.553465 -0.749711 0.760063 -0.310221 -0.604302 0.461717 -0.164797 0.695096 -0.574702 -0.127112 0.060495 0.051273 -0.202959 -0.220459 -0.055064 -0.059435 -0.1
0.395861 0.516972 0.323793 0.319462 0.302128 0.239077 -0.595303 0.518743 -0.189419 -0.497428 0.285749 -0.094018 -0.095779 -0.224677 -0.232477 -0.010972 -0.156619 -0.414219 0.283127 -0.153426 -0.485393 0.
-0.229118 0.149180 0.618693 0.506356 0.081285 0.558373 -0.102360 -0.220726 -0.013814 0.100432 -0.947711 -0.493934 -0.377036 -0.533715 -0.047406 0.045339 -0.623771 -0.735599 0.169595 0.139533 0.044263 0.
) -0.526997 0.493909 -0.539595 -0.263516 0.740491 0.453365 -0.269416 -0.006060 0.004232 -0.514244 -0.127158 -0.129637 0.413762 -0.209459 -0.144337 -0.701965 -0.1218717 -0.3494204 0.251245 -0.413024 -0.4339
(- -0.149397 0.695665 -0.145102 -0.295301 0.078746 0.023700 0.469291 -0.221025 -0.161547 -0.704949 0.399453 -0.336930 -0.235515 -0.114459 0.007716 -0.929563 -0.710490 -0.374750 0.302306 -0.121459 -0.17329
) -0.103397 0.253959 -0.541258 0.007243 0.232400 0.594627 0.410339 0.376903 -0.520273 -0.510069 -0.103210 0.103100 0.702247 0.450236 0.029550 -0.251000 -0.321379 -0.103944 -0.324140 -0.413024 -0.404907 0.
中国人类基因组 -0.554552 -0.117153 0.154736 -0.213345 -0.804052 0.259949 -0.280345 -0.260343 -0.572241 -0.851183 0.265322 0.196331 0.524735 0.021048 0.249394 0.295832 -1.804102 -0.437710 -0.256446 -0.76237
-0.338393 0.391520 0.066054 -0.107587 0.345048 0.188811 -0.048263 -0.128328 -0.080536 -0.317617 0.190479 0.165699 0.204866 0.410687 -0.632168 -0.256174 -0.310790 -0.575797 0.207894 -0.254746 0.162035
而 -0.683616 -0.139710 -0.283748 -0.248002 -0.410549 0.435339 -0.382991 0.285230 -0.347524 -0.372399 0.417284 0.173145 -0.400059 -0.072022 -0.370026 0.239835 -0.424012 0.531262 -0.242244 0.390584 0.100448
年 -0.007939 0.193132 0.037028 -0.463608 -0.155337 -0.207249 -0.432929 0.000967 -0.259462 -0.742979 -0.149109 -0.248516 0.098218 -0.544625 -0.435960 -0.119896 0.051500 -0.250010 0.127341 -0.493449 -0.4949
记 -0.327351 0.315930 0.105710 -0.455545 0.104671 -0.113224 -0.305036 0.122645 0.280216 -0.400794 -0.917457 0.322979 -0.142556 -0.555251 0.446439 0.104832 -1.157476 0.445310 0.009596 -0.432420 0.497362 0.
地 -0.139489 0.366709 -0.537243 0.538955 0.065546 -0.593978 -0.479537 -0.198422 -0.542359 -0.451418 -0.136948 -0.401302 -0.671460 -0.643932 0.104481 -0.639339 -1.504565 0.467690 0.573733 0.231293 0.4
与 -0.200055 0.481881 0.423410 -1.124764 -0.584877 0.738011 0.051111 0.349428 -0.282263 -0.484454 -0.357406 0.317410 0.049094 0.521468 0.216489 -0.104739 -0.530940 -0.173491 0.426115 -0.029393 -0.155196 0.
0 -0.584948 0.475662 0.060132 -0.444824 -0.168810 0.201031 -0.356272 0.186458 0.020977 -0.947499 0.244937 0.126569 -0.010646 0.603938 -0.222244 -0.040273 0.493378 -0.401415 0.801448 0.738314 -0.280321 0.
一个 -0.023847 0.116393 0.174876 0.407787 -0.441499 0.834712 -0.273033 -0.038076 -0.302474 -0.321432 0.127442 -0.378922 0.009149 -0.174637 0.357346 0.487470 -0.740794 -0.468971 -0.289700 0.219471 -0.05154
也 -0.413869 0.302570 0.104742 -0.049221 0.400444 0.346644 0.306530 -0.369633 0.272059 -0.218155 0.236779 0.711720 -0.181742 0.212496 0.070444 0.344442 -0.809499 0.006462 0.024244 -0.402977 -0.145459 -0.4
-0.507936 0.217076 -0.079622 -0.093724 -0.187017 -0.240519 -0.199379 0.356639 0.124073 -0.044034 0.990678 -0.659393 0.159964 -0.450500 -0.007416 -0.170354 -0.433394 0.230771 0.119432 -0.141121 0.0262
0.055435 0.012091 -0.242028 -0.061104 -0.184341 -0.423979 0.117708 -0.017418 0.010751 -0.238471 0.210337 0.313647 -0.301891 0.748448 0.709898 -0.317890 -0.034702 0.016072 0.056976 0.643397 0.469138 0.
者 -0.681358 -0.209425 0.209185 0.134171 -0.081344 0.171429 0.521251 -0.875270 -0.041718 -0.164330 -0.404734 0.358560 0.169762 0.160314 -0.283354 0.194334 -0.739423 -0.081036 0.506442 0.455449 0.203964 0.
-0.449996 0.314245 -0.337955 -0.649337 0.227015 -0.221813 -0.347227 -0.110313 -0.449454 -0.250464 -0.305642 -0.190223 0.689203 0.185000 -0.205469 -0.602812 -0.275401 0.171016 0.303844 -0.054644 -0.30301
- 0.502389 0.103976 0.669356 0.115554 -0.470230 0.077622 -0.301181 0.359306 -0.290000 -0.541434 -0.017033 -0.307323 0.109569 -0.552220 0.007315 -0.395620 -0.734200 -0.702617 0.675557 -0.409365 -0.261505
0.102200 0.357103 0.364280 -0.763375 0.129512 0.246706 -0.599393 0.126952 -0.394633 0.524454 0.407453 0.939704 -0.652240 0.242432 0.447300 -0.309395 0.461246 0.457941 -0.100425 0.119420 0.74
日 -0.491561 -0.376857 -0.119016 -0.550716 -0.087828 0.012346 -0.345129 0.386118 -0.044075 -0.415806 0.049334 0.119106 0.346989 -0.564738 0.847828 -0.351479 -0.052697 0.201492 0.075100 0.023031 -0.183980
我们 -0.028360 0.157122 -0.079340 0.107241 0.417036 0.175543 0.069358 -0.179503 -0.325453 0.093194 -0.044320 0.117105 -0.116712 -0.159460 0.129234 0.524418 -0.182777 -0.685230 0.054875 0.047178 -0.312035
公司 -0.218759 -0.499209 -0.601310 -0.184789 -0.149749 0.530166 0.255370 0.229822 -0.143204 -0.126259 -0.393253 -0.065171 -0.395056 0.246942 -0.044480 0.462039 0.159492 0.194982 0.170374 -0.210009 0.01039
等 -0.712253 0.599465 0.216296 0.387513 -0.174298 0.545460 0.065770 -0.189580 0.024345 0.403301 0.015903 -0.034310 0.677002 0.702663 0.303742 0.048487 -0.359370 -0.365094 0.139297 0.035210 0.374465 0.597
( -0.164956 0.410369 0.020755 -0.646032 -0.005647 0.459483 -0.306494 0.231711 -0.415261 -0.355339 0.072331 0.354036 -0.100716 0.336078 0.049507 -0.370032 -0.379040 -0.564199 -0.021742 -0.514390 -0.090211
进行 -1.081092 0.656891 -0.504659 -0.222100 -0.323952 0.773391 0.502098 0.710663 0.117400 -0.312078 0.349494 1.341781 -0.123873 -0.302620 0.152341 -0.569366 -0.794360 -0.456364 0.477327 0.058077 -0.42574
改善 -0.410488 0.501241 -0.667087 -0.314803 0.364829 0.437170 0.138364 -0.530041 0.081441 -0.718211 -0.174149 0.338393 -0.580274 0.483510 -0.334202 0.558467 -0.717850 -0.127417 0.854467 -0.128420 -0.558162
人 -0.037749 0.154949 -0.008900 0.050701 -0.012394 -0.531354 -0.195364 -0.096399 -0.162447 -0.102121 0.899489 0.029511 0.117281 -0.310247 0.355455 0.001938 0.001938 -0.229644 0.327428 0.178179 0.339523 0.
-0.335303 -0.050389 -0.466206 -0.704453 -0.443489 1.359935 -0.063768 -0.000258 -0.203338 0.055402 0.301432 0.097306 0.049837 0.672516 0.232277 0.000954 -0.043094 -0.229563 -0.064745 0.294780 0.18115
上 -0.480478 0.601425 -0.152125 -0.194683 0.186910 -0.065225 -0.113017 -0.084612 0.266405 0.184754 0.511079 0.214147 -0.464762 0.081328 -0.330357 -0.149491 0.120432 -0.229448 0.478032 0.562470 -0.018287 0.
地 -0.193795 -0.237145 0.268055 0.251335 0.294627 0.374871 0.034799 -0.428749 -0.456871 -0.030469 0.195557 -0.261607 0.080317 -0.480370 -0.184694 0.494946 0.501335 -0.494864 0.379156 -0.494864 0.583494 -0.
自己 -0.324914 -0.110824 -0.346858 0.352417 0.108833 0.181172 0.139638 -0.072928 -0.454456 -0.774331 -0.015782 -0.035774 -0.004787 -0.104862 -0.140464 0.444499 -0.111510 -0.762955 -0.043860 0.078173 0.809
不 -0.355303 -0.110824 -0.346858 0.352417 0.108833 0.181172 0.139638 -0.072928 -0.454456 -0.774331 -0.015782 -0.035774 -0.004787 -0.104862 -0.140464 0.444499 -0.111510 -0.762955 -0.043860 0.078173 0.809
或 -0.281117 0.116479 -0.173922 0.175253 0.694454 -0.054645 -0.204530 -0.420597 -0.604135 0.292424 0.390247 -0.340494 -0.142852 -0.305195 -0.003037 0.421169 -0.203490 -0.157245 0.184627 -0.233349 -0.103660
被 -0.827108 -0.194611 -0.188813 0.015546 -0.501452 0.230778 -0.189091 -0.124894 -0.241985 -0.246392 0.011030 0.246317 -0.005462 0.820255 0.001195 0.460132 -0.149780 -0.280813 0.184721 -0.223349 -0.33652
三 -0.404770 -0.110313 -0.500296 0.512924 0.135480 -0.460885 -0.177730 0.277495 -0.405821 -0.104974 -0.105712 0.510136 0.400536 -0.349103 0.091195 -0.255416 0.208907 -0.516017 -0.562000 0.475227 -0.79447
工作 -0.365694 0.781449 0.785651 -0.113182 -0.304008 0.432649 0.278255 -0.011127 -0.647755 -0.307455 -0.893450 0.198980 0.402232 -0.175244 0.506556 -0.210445 -0.116740 -0.554422 -0.144474 -0.238176 -0.442
被 -0.062320 0.420708 -0.470424 -0.274172 -0.052497 0.116739 -0.593794 -0.463056 -0.055019 -0.203393 0.162210 0.284697 0.256405 -0.020953 -0.272327 0.014819 -0.281600 -0.189296 -0.025449 -0.223503 -0.14977
三 -0.481189 1.079541 -0.301899 -0.246954 -0.475414 0.492064 -0.172775 -0.202661 -0.750221 -0.017328 0.096765 -0.568992 0.200739 0.164918 -0.315997 0.147159 -0.705429 0.002592 -0.011521 0.510136 -0.4065
个 -0.433957 -0.131154 0.669495 0.663257 -0.621851 0.535761 -0.514277 -0.116794 -0.214454 -0.023159 0.581734 0.252246 0.519075 0.280366 0.431128 0.492488 -0.329783 -0.767437 0.561447 0.220052 0.733771 -0.
这 -0.355947 -0.156463 0.551252 -0.150815 0.627899 0.194219 -0.337338 -0.443099 -0.149201 0.002884 -0.339455 -0.166191 0.577797 0.553957 -0.484613 0.615870 -0.249594 -0.430000 1.149037 -0.217680 0.031444

```

图4

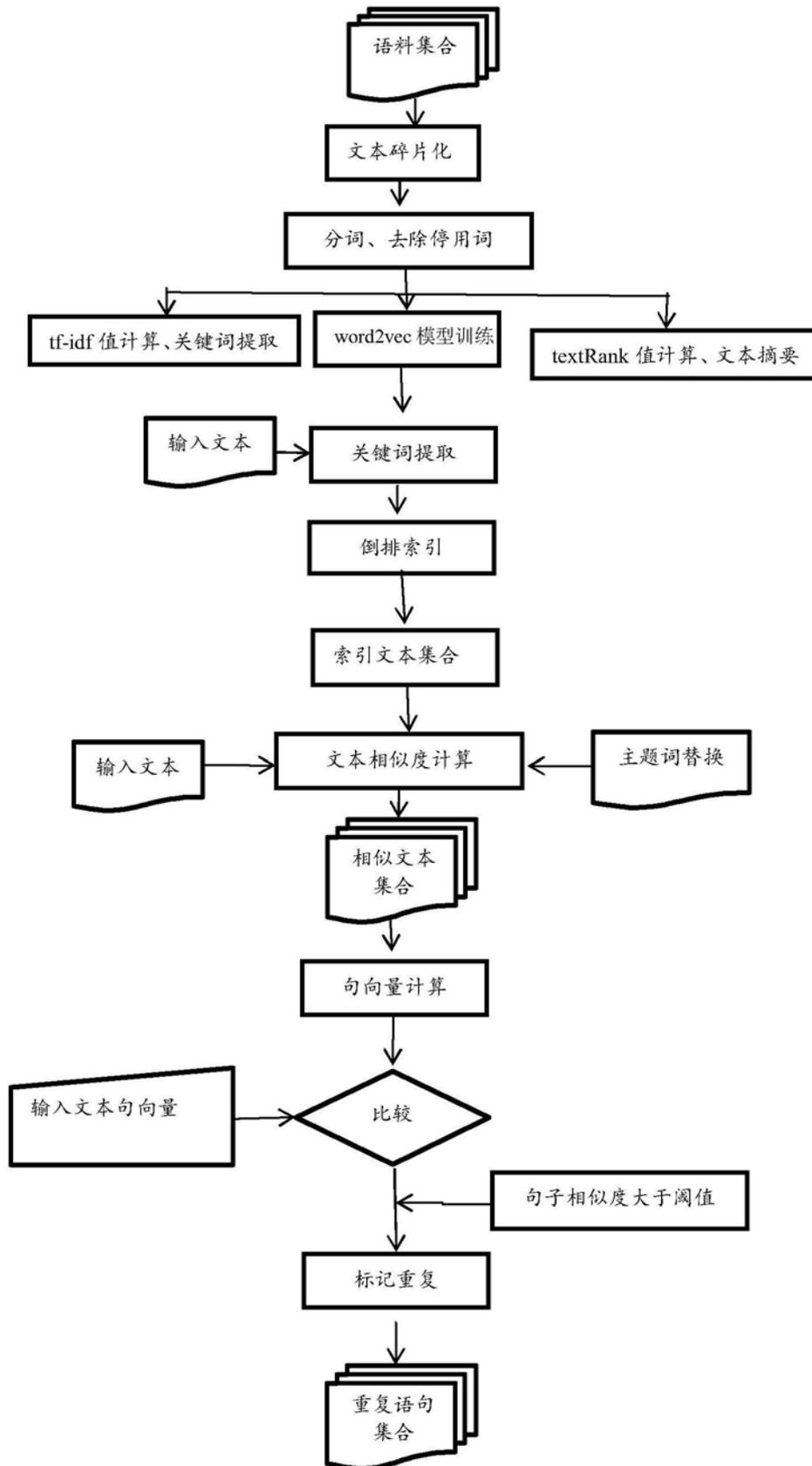


图5

军事 (557) 的相似文本:

标题	关键词	全文	相似度	操作
军事 (120)	伊朗, 潜艇, 水域, 海军, 波斯湾, 国产, 推演, 以色列, 瓦伊迪, ...	新华网消息 伊朗国家电视台日前报道称, 伊朗武装部队已在波斯湾水域部署了新型国产潜艇。这艘“豹”号潜艇可以携带多用途 ...	59.9%	相似度分析
军事 (356)	伊朗, 民兵, 导弹, 精英, 特种部队, 游击战, 美国, 分子, 美军, ...	伊朗以“游击战”为主的“不对称作战”方式是对美国这样的强国实施有效打击的“捷径” 国际先驱导报特约记者朱伟报道2月20日 ...	59.9%	相似度分析
军事 (570)	伊朗, 导弹, 朝鲜, 战机, 西方, 飞毛腿, 幻影, 卫队, 140, 武器, ...	谁给了伊朗“超级武器”? 本报专稿 吴义福 据美联社报道, 伊朗近来通过演习炫耀其“王牌武器”, 让原本因伊朗核 ...	59.9%	相似度分析
军事 (615)	伊朗, -140, 导弹, 海军, 中队, 驱逐舰, 米萨格, 潜艇, 装甲, 驱逐舰, ...	根据伊朗广播电台2月6日的报道, 伊朗国防部开始正式生产一种名为“米萨格”-2的新型导弹。报道指出, 伊朗国防部研制成功 ...	59.9%	相似度分析
军事 (663)	伊朗, 防空, 战争, 美军, 空袭, 东非西澳, 空军, 体系, 武器, 赫, ...	如果将传统的消耗战比成流血流汗而死, 那么现代战争则是一上来就对手的脑袋开一枪, 让其来不及流血就成为“ ...	59.9%	相似度分析
军事 (669)	伊朗, 教练机, -23, 坦克, 狙击, 博物馆, 突击, 士兵, 且, ...	“猎鹰”展翅——中国 15 新型高级教练机研制成功 “猎鹰”的首飞成功, 对于完善中国初、中、高级教练机的研制生产体系 ...	59.9%	相似度分析
军事 (670)	伊朗, 展出, 航空, 马来西亚, 法吉尔, 莫哈杰尔, 武器展, 沙海, ...	中新网4月27日电 据俄罗斯塔斯社报道, 目前正在马来西亚举行的第10届“OSA-2006”国际武器展览会上, 伊朗展出了其在军 ...	59.9%	相似度分析
军事 (737)	伊朗, 社会, 国际, 中东, 核武器, 利益, 困境, 沈德福, 核大国, ...	伊朗核问题折射出“一国家安全”和“全球整体安全”相矛盾的安全困境。这给国际社会提出新的挑战。解决这一困境需要国际社会 ...	59.9%	相似度分析
军事 (738)	伊朗, 中东, 布什, 总统, 核武器, 美国, 伊拉克, 地区, 政权, 萨, ...	美国之所以在伊朗核问题上喋喋不休, 就是为了掩盖自己在中东战略问题上束手无策的尴尬处境 在布什总统向美国国会 ...	59.9%	相似度分析
军事 (776)	伊朗, 安理会, 美国, 动武, 制裁, 制裁, 原子能, 以色列, 冲突, ...	“打还是不打, 这是个问题!” 伊朗核问题提交安理会, 可謂一石激起千层浪, 国际社会空前关注, 国内媒体也纷纷猜测: ...	59.9%	相似度分析
军事 (785)	伊朗, 防空, 道尔, 导弹, 德米特里耶夫, 合同, 履行, 俄罗斯, ...	新华网莫斯科2月9日电 (记者 卢涛) 俄罗斯对外军事技术合作局局长德米特里耶夫9日对此间媒体说, 俄罗斯将履行向伊朗 ...	59.9%	相似度分析
军事 (794)	伊朗, 核武器, 约瑟夫, 伊拉克, 美国, 石油, 谴责, 西方, 军事, ...	近日, 美国一些政要已公开讨论对伊朗开战的问题; 而针对美国如此举动, 伊朗方面表示, 有石油能操作后盾, 伊方不惧怕 ...	59.9%	相似度分析

图6

相似文本: 体育 (922)
相似度: 31.41%
相似字数: 311

国奥主帅 动态 本报讯 中国足协准备好了与特鲁西埃谈判的合同文本, 也在北京给他预订好了房间, 但“白巫师”特鲁西埃爽约了! 昨天, 正在摩洛哥的特鲁西埃通过律师致函中国足协, 称他本人和律师因故无法如约在5月6日抵达北京与中国足协谈判, 并对此表示歉意。这令中国足协措手不及, 并质疑特鲁西埃这种出尔反尔的行为[1]。 中国足协确定特鲁西埃为第一候选之后, 随即委托盈方公司与特鲁西埃及其律师进行联系[2]。特鲁西埃多次向中国足协表达执掌中国国奥队的强烈愿望。中国足协也对特鲁西埃“志在必得”, 所以双方约定5月6日在中国进行面谈。与此同时, 中国足协积极准备与特鲁西埃谈判的各种文本材料[3]。前晚, 特鲁西埃在与记者通话时也证实, 他将与律师谈议一起到北京, 并开玩笑说让记者“请他们吃北京烤鸭”[4]。然而昨天, 特鲁西埃通过律师发来的一封信致歉信令中国足协恼火不已, 仅仅两天时间, 特鲁西埃突然变卦, 昨晚, 记者通过电话找到正在摩洛哥的特鲁西埃, 他明确告诉记者, 这次不会来中国谈判了。“但我可以告诉你, 我绝不是因为有了其他球队的邀请或者有其他原因, 完全是因为我自己的家庭[5]。我马上就要领养第二个女儿, 两个月之内我无法离开摩洛哥[6]。” 国奥主帅 隐情 特帅为何有了二心? 沙特豪门重金诱惑 本报讯 国奥主帅热门人选特鲁西埃为什么突然爽约不来中国足协进行谈判呢? 一个很重要的原因就是特鲁西埃现在和亚洲豪门伊蒂哈德越走越近。据昨天出版的沙特主流媒体《祖国报》报道, 特鲁西埃将成为伊蒂哈德的主教练, 率领向亚冠“三连冠”发起冲击。该报援引伊蒂哈德高层官员的话报道, “俱乐部已经和特鲁西埃初步达成了一致, 如果不出意外的话, 这名法国教练很快将与俱乐部签约。” 据沙特记者透露, 伊蒂哈德给特鲁西埃开出了100万欧元的年薪, 还不包括各种比赛奖金。 特鲁西埃在此前接受记者采访时表示: “2008年北京奥运会举世瞩目, 如果能以中国队主教练身份参加, 对我个人而言, 无疑是巨大的荣誉。” 虽然如此, 但对特鲁西埃来说, 薪金也是重要的考虑因素。中国足协虽然有盈方资本介入, 但在国奥队主教练待遇问题上, 上限价位也只是100万美元, 而且一直希望能以180万美元确定。据足协内部人士透露, 在此前足协和特鲁西埃经纪人的非正式接触中, 双方在薪水问题上存在一定分歧, 足协报价仅为50万美元左右。但在和伊蒂哈德的谈判中, 沙特人很快就开出了100万欧元的价格, 这让特鲁西埃很难以拒绝。

时报讯 中国足协准备好了与特鲁西埃谈判的合同文本, 也在北京给他预订好了房间, 但“白巫师”特鲁西埃爽约了! 昨天, 正在摩洛哥的特鲁西埃通过律师致函中国足协, 称他本人和律师因故无法如约在5月6日抵达北京与中国足协谈判, 并对此表示歉意。这令中国足协措手不及, 并质疑特鲁西埃这种出尔反尔的行为[1]。 中国足协确定特鲁西埃为第一候选之后, 随即委托盈方公司与特鲁西埃及其律师进行联系[2]。特鲁西埃多次向中国足协表达执掌中国国奥队的强烈愿望。中国足协也对特鲁西埃“志在必得”, 所以双方约定5月6日来华。与此同时, 中国足协积极准备与特鲁西埃谈判的各种文本材料[3]。前晚, 特鲁西埃在与记者通话时也证实, 他将与律师谈议一起到北京, 并开玩笑说让记者“请他们吃北京烤鸭”[4]。然而昨天, 特鲁西埃通过律师发来的一封信致歉信令中国足协恼火不已, 仅仅两天时间, 特鲁西埃为何突然变卦[5]? 昨晚, 记者通过电话找到正在摩洛哥的特鲁西埃, 他明确告诉记者, 这次不会来中国谈判了。“我已经致函中国足协, 在信中向他表示了歉意。但我可以告诉你, 我绝不是因为有了其他球队的邀请或者有其他原因, 完全是因为我自己的家庭。我马上就要领养第二个女儿, 按这里的规矩, 未来4个月之内我无法离开摩洛哥。” 特鲁西埃说: “我不想给中国足协压力, 而且我已经做好了准备, 争取在土伦杯赛期间去看中国队的比赛, 中国国家队到法国、瑞士参加的两场热身赛, 我也将努力争取去看至少一场。我只是因为家庭的原因才向中国足协写了这封致歉信, 绝对没有其他意思。在世界杯赛期间, 我也有很多事情, 恐怕无法像此前中国足协所要求的那样, 在世界杯赛期间就去中国上任。” 特鲁西埃突然变卦, 足协高层非常恼火。“他这么做是不是太不靠谱了? 说好了来, 我们筹备工作也全部做好了, 甚至连酒店、会议场所也准备好了, 怎么能够说来就不来了呢? 而且还抛出一个家庭原因, 如果是家庭问题, 之前主动要求参加候选时怎么不说呢? 而现在我们真心诚意邀请他来, 怎么又找出这么个理由?” 虽然特鲁西埃只是表示这次不能来中国谈判, “今后来不来还未确定”, 但鉴于他的态度, 中国足协不会在特鲁西埃身上下太多工夫, 而将目标转移给其他人。(李 岩)

图7