# 1 引言

网络舆情是随着网络和信息技术的快速发展普及产生的，是大众就各种社会问题在现代的网络空间中发表的不同意见和态度的信息。互联网时代，网络已成为社会舆论的思想文化信息的集中地。舆论信息反映了民众的思想状况，对于社会的和谐稳定有着重要的作用，因而有效的对网络上的舆情信息进行监控，实时掌握网络舆情信息的动态是非常有必要的。

移动互联网时代，中国创业热潮风生水起，不仅有大量的移动互联网企业赴美上市，更是诞生了无数个创业奇迹。移动互联网不仅为我们的生活带来了便利，更是把创业热潮推向了历史最高峰。“大众创业万众创新”的提出，各种互联网应用创新和商业模式创新不断涌现，引发传统行业生态的深刻变革。互联网移动化、产品化进程加快，互联网企业聚力转型，抢占未来产业高地。大数据时代的创业热点更是成为创业者、投资者的关注点，进而提出了对于创业热点的分析。

目前信息处理领域对舆情监控的研究和成型的产品很多，对网络舆情监控的研究也较为成熟，但还没有对创业热点的抓取和分析。如果能够对创业热点信息进行监控，获得较为准确的突发创业热点，那么不管对于想要创业的创业者还是风险投资的天使人都有很重要的参考价值和意义。针对近年创业热潮的分析监控的研究暂时为空白，本文将已有的对舆情监控的研究方法和分析，运用到对创业突发热点的分析上来，为众多创业者投资者提供有参考价值的创业热点数据，避免盲目创业，降低风险投资。

对于创业数据文本的挖掘，应用主题模型是一个很好的方法。主题模型相对于传统的文本挖掘方法，能够高效地完成一些基本工作，如发掘出文本的潜在关系、判断关联性、分类等。本文使用基于潜在语义分析的文本挖掘方法来进行创业热点的挖掘，主要使用LDA模型——一种基于潜在狄利克雷分布的主题生成模型。在这个主题模型中，一系列主题以服从多项式分布的形式生成每个文本，再从这些主题中以服从多项式分布的方式抽样出每个单词，由此构成该模型围绕主题生成文本的过程。

本文所做的主要工作如下：爬取创业大数据，根据数据本身类别标记进行数据处理，使用LDA模型分析各个类别下的创业热点。

# 2 创业热点分析相关技术

## 2.1 网络爬虫技术

创业数据的采集是创业热点分析处理的第一步。本文的创业数据来源是有关的创业网站如36氪、it桔子、创业邦等。通过对这些网站的网页结构、数据加载等方面分析发现，这些网站并不是简单的静态网页，网页中的数据都是通过JavaScript代码的执行动态加载再渲染出来的，这类基于Ajax脚本的动态网页带来了更好的用户体验，同时也使得对其进行获取和分析面临着新的问题和挑战。只是简单的借助Python的urllib模块已经无法获取到我们所需要的数据信息，本文采用动态爬虫工具Selenium2，它是通过模拟浏览器的各种操作来执行动态脚本程序。通过爬取动态网页的数据，将特征信息创业公司名称、简介、分类、投资信息、招聘职位月薪开支等数据存入MongoDB数据库中，构建自己的创业数据库。

## 2.2 数据预处理

本文已经在Selenium爬取数据时借助Beautiful Soup已经对需要的数据信息进行了提取，所以预处理这一环节就只需要去停用词和中文分词这两步处理。

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字或词，这些字或词即被称为Stop Words（停用词）。中文语言中的一些功能词语其他词相比没有特别的实际含义，如“的”、“啊”、“呀”等，这些词通常用于连接句子成分或表达感情、强调语调等用途，因此称为停用词。这些停用词都是人工输入、非自动化生成的，生成后的停用词会形成一个停用词表。停用词的存在不仅会降低实验处理结果的正确率，且由于停用词数量较多，导致文本矩阵存在严重的高维稀疏等问题，从而影响实验性能。

分词是将连续的字序列按照一定的规范重新组合成词序列的过程。由于中文文本处理的基本单位是词语，而中文文本并不像英文那样每一个单词之间会用空格分开，中文文本中只有句子单位之间才会有标点符号，因此中文分词也就比英文要复杂的多。中文分词是进行文本挖掘的基础，所以首先需要对中文文本进行分词操作。本文中采用Python中文分词组件jieba中文分词，提供的功能包括有分词、添加自定义词典、关键词提取、词性标注、并行分词等。

## 2.3 主题模型

传统判断两个文档相似性的方法是通过查看两个文档共同出现的单词的多少，如TF-IDF等，这种方法没有考虑到文字背后的语义关联，可能在两个文档共同出现的单词很少甚至没有，但两个文档是相似的。比如苹果和乔布斯这两个词一起出现时，可以设定属于苹果公司这同一主题。所以在判断文档相关性时需要考虑文档的语义，而语义挖掘的利器是主题模型，LDA就是其中一种很有效的模型。

Latent Dirichlet Allocation（LDA）模型LDA 由Blei, David M.、Ng, Andrew Y.、Jordan 于2003 年提出，是一种主题模型，它可以将文档集中每篇文档的主题以概率分布的形式给出，从而通过分析一些文档抽取出它们的主题（分布）出来后，便可以根据主题（分布）进行主题聚类或文本分类。同时，它是一种典型的词袋模型，即一篇文档是由一组词构成，词与词之间没有先后顺序的关系。此外，一篇文档可以包含多个主题，文档中每一个词都由其中的一个主题生成。

LDA模型以概率统计为基础，认为每一篇文章的每个次具有一定概率，假设作者在写一篇文章时是通过一定的概率在头脑中选择了某个主题，并基于这个主题在大脑中以一定的概率想到了某个词语。因此生成一篇文档的每个词语的概率为：

·················2.1

这个概率公式可以表示为图2.1的矩阵：

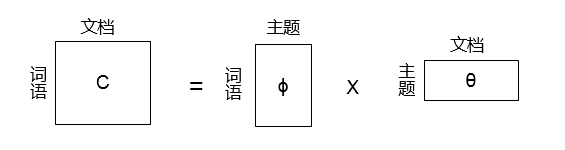


图2.1 LDA概率公式矩阵

其中“文档-词语”矩阵表示每个文档中每个单词的词频；“主题-词语”矩阵表示每个主题中每个单词出现的频率；“文档-主题”矩阵表示每个主题出现的频率。

在LDA模型中，一篇文档生成的方式如下：

（1）从狄利克雷分布中取样成文档i 的主题分布；

（2）从主题的多项式分布中取样生成文档i 第j 个词的主题

（3）从狄利克雷分布中取样生成主题对应的词语分布

（4）从词语的多项式分布中采样最终生成词语

将这个过程重复N次就产生了文档d，图模型如下：

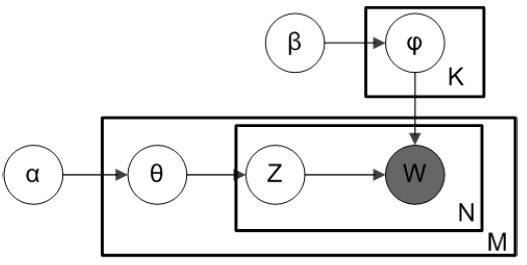


图2.2 LDA模型图

和表示预料级别的参数，也就是每个文档都一样，因此生成过程只采样一次。是文档级别的变量，每个文档对应一个，也就是每个文档产生各个主题z的概率是不同的，生成每个文档采样一次。z和w都是单词级别变量，z由生成，w由z和共同生成，一个单词w对应一个主题z。

从上面的 LDA生成模型的描述中我们知道，在LDA主题模型在生成过程需要两个控制参数，分别是和，有了这两个参数就能够用于生成新的文本，其中表示文档集中各个潜在主题之间的相对强弱，则刻画了各个潜在主题本身的概率分布，即潜在主题中各特征词的概率。

通过LDA建模后就可以知道“主题-词语”的分布情况，表示单词在主题中的概率，概率值越大则该词越能表示该主题，因此可以从每个主题中选择前n个概率最大的词作为特征词，这样基于LDA的主题模型的特征词的选取很好的克服了传统的特征词选取中存在的丢失语义的不足。

# 3 实验

## 3.1 创业数据准备

本文实验数据通过动态爬虫工具主要采集的是招聘网站拉勾网上的创业公司数据，保留拉勾网网站本身对创业公司的分类标记，在剔除一些空白噪声数据后，保留12万条创业公司的简介信息作为实验数据分析。在使用LDA模型分析之前必须进行数据预处理工作：去除停用词、中文文本分词。

## 3.2 实验步骤

在创业热点分析的LDA模型实验中，各个参数的设置参考了文献[2]中的说明，Topic=10,α=0.1,β=0.01，循环迭代次数设为50。实验步骤如图所示：

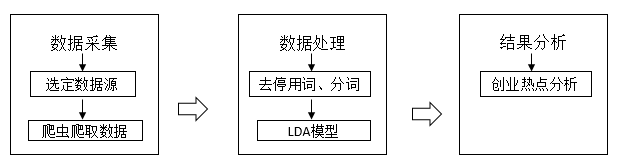


图3.1 实验步骤

## 3.3 实验结果

通过对带有标记的创业公司的简介文本运用LDA模型得到的主题分布显示了各个类别下各主题的热词，并按照各个词的热度概率降序排列。每个类别下的主题挖掘都会有相关员工待遇、企业文化等Topic，本文仅展示该类别下细粒度的热点主题。如图3.2中展示了硬件类别下其中3个主题，根据各个主题对应的关键词可以发现，Topic1是机器人相关的主题，Topic2是智能家居相关的主题，Topic3是3D打印相关的主题。

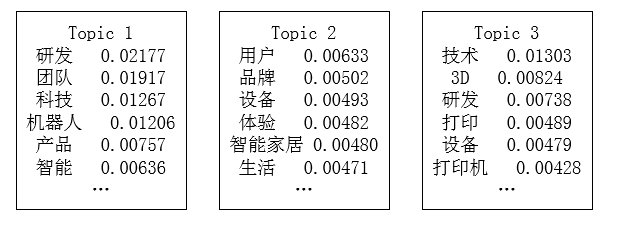


图3.2 LDA模型的硬件类别下主题挖掘

图3.3中展示了教育类别下的3个细粒度主题，其中Topic1是海外留学相关的主题，Topic2是课程辅导相关的主题，Topic3是儿童早教相关的主题。说明模型挖掘到的主题，主题对应的关键词准确性较高，且主题之间的独立性较强。



图3.3 LDA模型的教育类别下主题挖掘

再如图3.4中展示了游戏类别下的主题，其中Topic1是手游相关的主题，Topic2是VR相关的主题，Topic3是动漫相关的主题。



图3.4 LDA模型的游戏类别下主题挖掘

通过实验结果可以看出，主题模型处理分析创业数据可以得到较好的分类效果。

# 4 结束语

本文将对网络舆情的研究应用到对创业热点的抓取和分析上来，通过爬取创业网站网页数据，搭建了自己的创业数据库，根据数据本身类别标记进行数据处理，进而使用LDA模型分析各个类别下的创业热点，获得某个类别的创业领域下的细粒度主题。

本文的不足之处在于本实验的创业数据采集没有考虑时间因素的影响，网站本身的信息量是随着网站存在时间的增长而增长的。因此，为了尽量淡化时间因素对实验结果的影响，在采集数据时考虑只采集近几年内新成立的创业公司的信息或者近期获得投资的创业公司信息。本文在今后的研究中将引入对时间序列的考虑。

# 参考文献

1. https://zh.wikipedia.org/wiki/%E5%A4%A7%E4%BC%97%E5%88%9B%E4%B8%9A%E3%80%81%E4%B8%87%E4%BC%97%E5%88%9B%E6%96%B0
2. Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003, 3(Jan): 993-1022.
3. Wei X, Croft W B. LDA-based document models for ad-hoc retrieval[C]//Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2006: 178-185.
4. 唐晓波, 向坤. 基于 LDA 模型和微博热度的热点挖掘[J]. 图书情报工作, 2014, 58(5): 58-63.