NLP 关键字提取技术对比

2017211106 班 2017212116 杨诺诚

摘要：关键词是代表文章重要内容的一组词汇，对文本聚类、分类、自动摘要等功能有着极为重要的作用。如果可以快捷的获得文章的关键词，还可以方便浏览和保存、识记信息。为了研究 TF-IDF、TextRank、LSA/LSI 以及 LDA 四种经典的 NLP 关键词提取算法和国产集成库 JiaGu 在多主题论文摘要文本关键词提取的场景下对比准确性、精确度以及运行速度等参数，进行理论学习和实验重现。可以得出 Jiagu、TextRank、TFIDF 算法在小型文本的场景下表现较好。

关键词：关键词、文本摘要、算法、对比

NLP keyword extraction technology comparison

Abstract: Keywords are a group of words representing the important content of an article. They play an important role in text clustering, classification, automatic summary and other functions.If you can quickly get the key words of the article, but also easy to browse and save, remember information.In order to study the four classic NLP keyword extraction algorithms of TF-IDF, TextRank, LSA/LSI and LDA and the parameters such as the accuracy, accuracy and running speed of the domestic integrated library JiaGu in the scene of multi-topic abstract text keyword extraction, research and experimental reproduction were carried out.It can be concluded that the Jiagu, TextRank and TFIDF algorithms perform better in small text scenarios.

Keywords： Keywords text abstract algorithm comparison

目录

引言

文章的关键词是最能表达文档主旨的几个词语，可以将文本关键词抽取问题转化为词语重要性排序问题，选取排名靠前的数个词语作为文本关键词。当下，主流的文本关键词抽取方法主要分为有监督（supervised learning）和无监督（unsupervised learning）两类。有监督学习通过分类方式，需要已有的较为丰富的词表，可以通过判断文章中词语和词表的匹配相关程度，通过打标签的方式来获得关键词，这种做法一般具有更高的精确度，但由于人类语言有创新性和多义性，在不同历史时期意思、词性等都可能有较大变化，需要相关人员长期维护更新。而无监督学习要求很低，它不使用人工维护的词表和标准语料支持就可以实现关键词提取，在当下使用范围更广，一般较为轻量级，但结果只针对文本词汇出现的频次、不去考虑文本内容逻辑以及潜在的主题。

算法介绍

在这个部分介绍了之后实验要用到的各种算法的逻辑，对提取的结果进行提前的预估。

TF-IDF 算法

TF-IDF 算法（Term Frequency-Inverse Document Frequency）是基于统计计算方法来评估单个词语对文章的重要性，需要语料支撑，属于典型的有监督学习方法。

TF-IDF 可以分为两个部分，即负责统计本文词频的 TF 部分和对比语料库中词频的 IDF 算法。如果单个词语在文章中出现次数足够多，可以认为它是本文的关键词汇，代表着中心意思，而当单个词语在语料库中出现次数较少时，则认为它可以有效的将文章和其他语料分开，区分能力更强。算法从词频和逆文档频次两个角度对词的重要性进行衡量。

以以下文本为例：“植树节当天，学校单位、植树志愿者可到植树规划地点查看之前的树苗情况，我们会进行公示，同时对树苗价格和成活率等也进行公示。”

高频词汇	本文中的词频	出现次数/1000 篇
植树	2	10
树苗	2	15
公示	2	100
进行	2	50

表 1-1 TF 模拟

文本中的“植树”、“树苗”以及“公示”、“进行”词汇的词频都是 2 次，从 TF 算法角度他们的重要性相当，而在 IDF 算法的角度显然植树、树苗更能凸显文章主题。

$$tf_{ij} = \frac{n_{ij}}{\sum_{k} n_{kj}}$$

公式 1-1

$$\mathrm{idf}_i = \log\left(\frac{|D|}{|1+D_i|}\right)$$

公式 1-2

公式 1-1 与 1-2 为 TF-IDF 的主要算法，n 表示词频数，k 则表示对文章中词频的统计，|D|指代总文章数，$|D_i|$则代表着文档集合中出现词语 i 的文章数。为了平衡两种算法的影响力，使用公式计算最终的重要性。

$$\mathrm{tf} \times idf_{(i,j)} = \frac{n_{ij}}{\sum_k n_{kj}} \times \log(\frac{|D|}{|1+D_i|})$$

公式 1-3

可以计算得出

高频词汇	Tf	Idf	tf×idf
植树	0.067	Log(1000/11)≈2	0.134
树苗	0.067	1.8	0.1206
公示	0.067	1.0	0.067
进行	0.067	1.3	0.0871

表 1-2 TF-IDF 模拟

结果显示“植树”最能代表这段文章的中心意思。

只靠词语的两类统计信息做关键词提取的传统 TF-IDF 算法不能很好的发挥出人工智能的优势，我们最终都要对文本中信息量的密集程度做出判断，包含信息量大的词汇，应该以更高的系数参与这场竞赛，比如定义了实体的名词词汇比虚词动词包含更多的信息；在文章首部的语句往往比段落中间的语句更具概括性。这也关系到某种特定语言的使用习惯、作者的行文逻辑等更包含智能的问题。在未来随着旧词新用和用语加速变化，必须要考虑这些方面来改进 TF-IDF 算法来获得更高的准确度和效率。

实验用到的文本一般内容直白简单，不涉及隐晦的文章主题问题，使用词频也可以取得不错的效果。

TextRank 算法

TextRank 算法思想来源于 Google 的网页排序 PageRank 算法，不依赖语料，可以仅对单篇文档进行分析提取关键词。

TextRank 在分析文档时，把文本分割成若干组成单元并建立完全有向图模型，利用投票机制对文本中的重要成分进行排序，

$$\mathrm{WS}_{(\mathrm{V_i})} = (1-d) + d \times \sum_{V_j \in \ln(V_i)} \left(\frac{\mathrm{w_{ji}}}{\sum_{V_k \in Out(V_j)} w_{jk}} \times WS_{(V_j)} \right)$$

公式 1-4

公式中 V 为点集合 E 为边集合，任两点 Vi , Vj 之间边的权重为 wji , 对于一个给定的点 Vi, In(Vi) 为指向该点的点集合 , Out(Vi) 为点 Vi 指向的点集合，此外为了保证收敛，加入阻尼系数 d。

TextRank 算法是利用局部词汇之间关系（共现窗口）对后续关键词进行排序，直接从文本本身抽取。还是使用这段文本：

“植树节当天，学校单位、植树志愿者可到植树规划地点查看之前的树苗情况，我们会进行公示，同时对树苗价格和成活率等也进行公示。”

首先可以对完成文章分句，完整句子分词得到 S=[‘植树节’，‘当天’，‘学校’，‘单位’，‘志愿者’……‘公示’]的词组集合，保留一部分候选词 S_i 构建关键词图，以共现关系（co-occurrence）构造边集合，在这里为了人工干预共现关系的强度，提出了“窗口”的概念，设立窗口值之后，最多共现 k 个词语，而不会出现泛化的很多词共同包含在一个共现关系中。可用例子获得以下几个窗口（窗口值 k=3）：[植树节、学校、单位]和[植树、志愿者、树苗]的两个，各点初始权值相同，根据公示 2-1 迭代至收敛后再对权值做排序，取靠前的词语获得候选词，为了消除邻近词语间意思重复，再将靠近的词汇合并，作为一个词出现。

目前一些学者针对 TextRank 做出很多的改进，将标题、段落、特殊句子、句子位置和长度等信息引入到网络图的构造中，对中文语法和一般写作的方法做出改进，获得了不错的效果。

LSA/LSI 算法

LSA(Latent Semantic Analysis，潜在语义分析)和 LSI(Latent Semantic Index，潜在语义索引)通常认为是同一种算法，应用场景略有不同。《Using Latent Semantic Analysis to Improve Access to Textual Information》。算法也是用之前介绍的 BOW 模型，经历以下步骤：

1)将现有语料的文本词向量拼接成“词-文”矩阵。

2)对“词-文”矩阵进行奇异值分解操作（SVD）

3)将分解后的结果映射到低维度 k 上

经过这一系列的操作后，文本和词都将近似成为 k 个主题空间中向量，就可以方便的计算出词、文档间的相似关系，相似度最高的词便是文章的关键词。SVD 操作可以将词和句映射到低维，充分的展现了文本的本质表达。但 SVD 本身计算复杂，且每次有新的文本进入特征空间都需要重新计算，非常不适合在本地计算机中使用，LSA 经过映射降维之后，物理解释性变差，词的频率分布不敏感，容易发生错误。

在发展过程中出现过 pLSA 的改进算法，它使用 EM 算法拟合 SVD 来实现降维，一定程度上优化了算法消耗，带还有很多不足。在 pLSA 中引入贝叶斯模型的 LDA（Latent Dirichlet Allocation 隐含 Dirichlet 分布）算法已经成为主流。

LDA 算法

LDA(Latent Dirichlet Allocation，隐含狄利克雷分布)融合了基础贝叶斯理论分布和 LSA 潜在语义分析，根据词共现的频率来拟合出词-文档-主题的语义空间。LDA 假设文档的主题，词汇出现都满足 Dirichlet 分布规律，使用语料中的分布情况得出贝叶斯理论中语料主题的“先验分布”，之后就可以通过实验观测得到也满足 Dirichlet 分布的后验概率和先验概率来获得一组 Dirichlet-multi 共轭，据此来推断文档中主题的后验分布和关键词的后验分布来

实现关键词提取任务。吉布斯 Gibbs 采样是求解 LDA 的经典方法。LDA 的目的是将文档集中每篇文档的主题以概率分布的形式给出，抽取出主题分布后，可以根据主题分布进行主题聚类或文本分类，最终根据给定的一篇文档，反推其主题分布。

LDA 算法是这些算法中计算量最大的一个，且抽取关键词的步骤需要先对文本主题进行推测，之后再对符合主题的关键词作抽取，需要大量的语料支撑才能达到较好的效果，实验中的语料并不丰富（只有 100 余篇文章），会给分词带来很多困难。

集成的 JiaGu 库

Jiagu 是一款国产集成的深度学习自然语言处理工具，提前使用了大规模语料训练，功能包含了知识图谱、关系抽取、中文分词、关键词和文本摘要等多种功能，参考了各大工具优缺点制作（https://github.com/ownthink/Jiagu）。

Jiagu 是目前算法中最为轻量级的自然语言处理包，因为它已经经过了大量训练，对新加入的语料有较好覆盖，可以在低性能终端上使用，得到不错的效果。

实验复现

实验所用的数据来自知网空间(http://www.cnki.com.cn/)中搜索“教育”、“汽车”、“通信”、“饮食”、“文学”、“历史”、“国家”、“政治”、“新闻”、“电子”共 10 个主题，135 篇文章。使用爬虫获取相关所有文章的题目、摘要文本以及作者给出的关键词。

在实验之前，先对数据集中摘要文本特征进行分析，分析图如下。

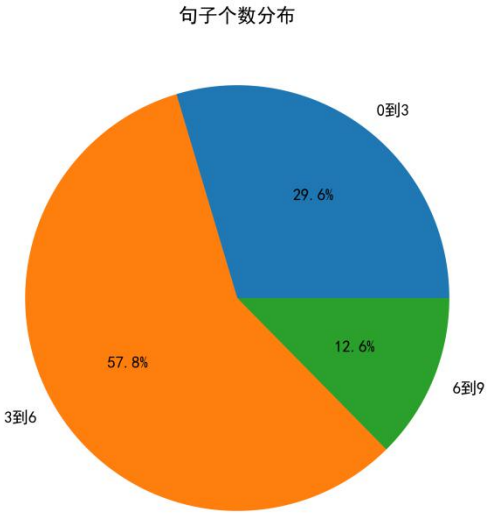

图 2-1 摘要数据集中句子长度的分布情况

先对文本的摘要句子数进行分析可知，这 135 篇文章的摘要长度一般集中在 6 句以下，数据量不大，没有足够的语料对 LDA、LSI 主题算法可能会有不利，但可以有效的体现 Jiagu、TextRank 不需要语料的算法的优势。

再对词数进行分析可得：

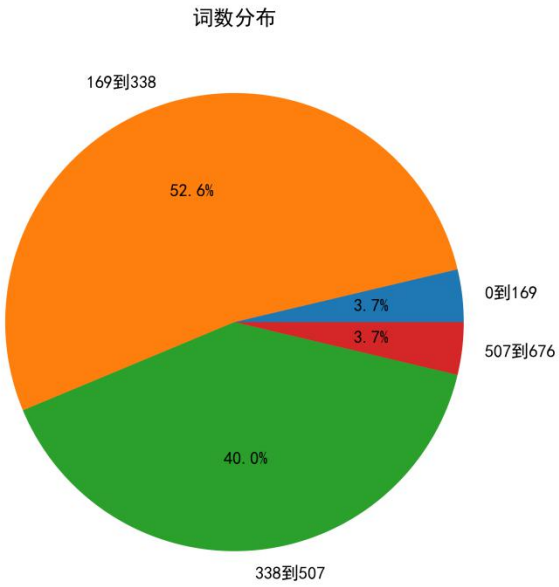

图 2-2 摘要数据集中词数的分布情况

摘要文本的字数主要分布在 169 词到 507 词间。图中的分类中的边界都取最大文本字数的几个四等分点。

考虑到摘要文本的篇幅和作者给出关键词数的限制，本次实验每种算法都选取四个关键词作比较。

模型	用时(s)
jiagu	3.1047
tfidf	92.5611
TextRank	2.5840
LSI	944.8247
LDA	1922.0688

表 2-1 模型运行时间一览

比较对这语料进行四个关键词提取的运行时长可以发现 LDA 算法的运行时间最长，TextRank 算法运行时间最短，因为每有新的语料加入，都需要重新计算概率才能确定关键词的分布情况，所有运行方法中 TextRank 的运行时间最短。

运行结果的关键词结果将和文章作者给出的关键词相比对,计算之后得到以下五个主要指标:精确匹配，关键词占比，模糊匹配数，最佳性能文章及其精确匹配值。

精确匹配占比（Accuracy）

精确匹配（Accuracy）指标是计算模型给出关键词可以精准匹配的词数占模型给出总关键词的比例，衡量模型给出结果的有效性，所有输出词汇中占 Accuracy 比例的结果可以精准匹配作者给出的关键词。

关键词占比（Precision）

关键词占比（Precision ）指标是计算模型给出的关键词中和作者给出关键词完全匹配的

词数占作者关键词的比例大小。这个指标可以较好的衡量模型给出关键词命中的精准的概率，是模型的最佳性能参数，作者标准关键词中占 Precision 比例的词汇可以被模型输出覆盖到。但由于作者关键词鼻根部仅仅依靠摘要给出，还会总览整篇文章，可能摘要对关键词的体现并不完整，只能作为参考。

模糊匹配数（dim）

模糊匹配数（dim）指所有模型输出中可以模糊匹配对应文章关键词的字数，这个指标设立之初，为了衡量模型给出同义词或包含相近意思字的能力。在文章中为了丰富语料会有一个意思多种表达的现象，模糊匹配可以大致反映出模型对文章的把握，仅看关键词是否能较全面的了解文章大意。

最佳性能文章及其模糊匹配值

最佳性能文章及其模糊匹配值的指标意在寻找最能符合模型特征的文段，回看文章摘要部分和作者给出关键词可以很好的了解模型的适用环境，对模型都能给出很好的精准匹配的文章还可以作为模型的样本。

model	accuracy	precision	dim	Best_performance	id
Jiagu	0.677	0.712	906	11	8
LDA	0.104	0.109	183	5	59
LSI	0.152	0.16	212	6	72
TextRank	0.665	0.698	850	11	82
TFIDF	0.676	0.71	907	11	8

表 2-2 模型的各个关键指标

对结果粗略观察可以知道 LDA、LSI 模型表现的并不好，其他模型的 Accuracy 平均到了 66%左右并且在表现最佳的文章中均有 11 字的覆盖，而 LDA、LSI 的 Accuracy 指标只到 15%，最佳表现文章只覆盖了 9 个字。且 Jiagu 和 TFIDF 结果非常相近，只在模糊匹配 dim 一栏中有略微的差异。

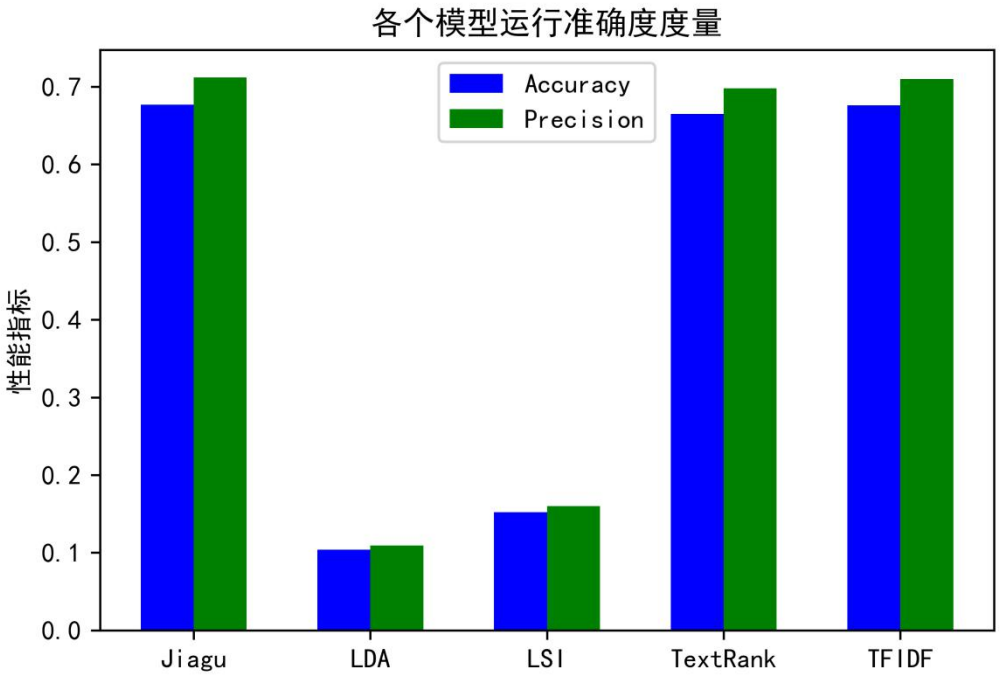

图 2-3 Accuracy-Precision 绘图

首先是 Accuracy 和 Precision 的联合绘图，五个模型中 Jiagu、TextRank、TFIDF 方法均取得了很高的准确度率，而 LDA、LSI 结果相近且效果不佳。Precision 总是略比 Accuracy 高，应该是样本给出的关键词均数比模型给出关键词均数略低导致的。

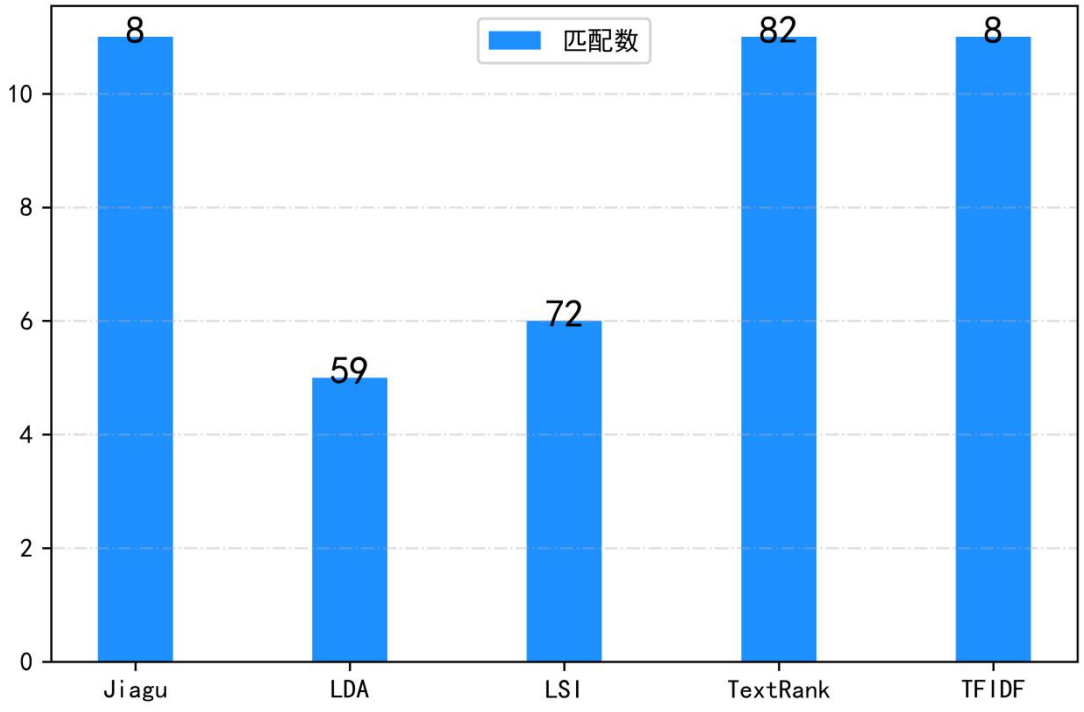

图 2-4 最佳文本和其表现值

Jiagu 最佳表现的第 8 篇文章中作者给出关键词是：“电动汽车 政策 商业模式创新”

模型	模型给出	模糊匹配数	有效字数占比
Jiagu	电动汽车 产业 创新 商业模式	11	91.7%
LDA	指标 全球 数据 手段	0	0

LSI	中国 全球 项目 城市	0	0
TextRank	电动汽车 创新 发展 产业	8	80%
TFIDF	电动汽车 产业 创新 商业模式	11	91.7%

表 2-3 Jiagu 的最佳表现文章

这篇文章是“汽车”主题下的一篇文章，研究电动汽车的创新和运营历史。各个算法都可以很好的获得电动汽车这一关键词，它是摘要中最多出现的词语，创新和运营则是文章行文的主要问题。Jiagu、TFIDF 这种直接分析文章分布的算法较有优势。

LDA 表现最佳的是第 59 篇文章，作者给出的关键词是：“‘‘互联网+” 中国远程教育 教育新生态 公共服务模式”。

模型	模型给出	模糊匹配数	有效结果占比
Jiagu	互联网 远程教育 教育 变革	9	81%
LDA	水平 模式 技术 互联网	5	45%
LSI	中国 方式 时代 环境	3	37.5%
TextRank	教育 远程教育 模式 发展	8	80%
TFIDF	互联网 远程教育 教育 变革	9	81.8%

表 2-4 LDA 的最佳表现文章

这是一篇教育主题的文章，围绕“互联网+教育”的展开，中途讨论远程教育的模式改革，深层意思也在文章中表现出来，LDA 这种以主题分类为重心的算法可以表现出优势。

LSI 表现最佳的是第 72 篇文章，作者给出的关键词是：哲学观 唯物史观 世界历史理论 经济全球化。

模型	模型给出	模糊匹配数	有效结果占比
Jiagu	全球化 经济 历史 世界	6	66.7%
LDA	全球 科学 根源 趋势	2	25%
LSI	世界 全球 社会 经济	6	75%
TextRank	全球化 经济 历史 世界	6	66.7%
TFIDF	全球化 经济 历史 世界	6	66.7%

表 2-5 LSI 的最佳表现文章

这是一篇历史类的文章，注重介绍了马克思理论下的经济学规律，阐述经济全球化的发展理论。

TextRank 表现最佳的文章是第 82 篇，作者给出关键词是：自然资源 国家所有权 公权 私权

模型	模型给出	模糊匹配数	有效结果占比
Jiagu	宪法 所有权 属于 自然资源	7	63.6%
LDA	行政 财产 宪法 条款	0	0
LSI	中国 社会 规定 国家	3	37.5%
TextRank	所有权 宪法 国家 自然资源	9	81.8%
TFIDF	宪法 所有权 属于 自然资源	7	63.6%

表 2-6 TextRank 的最佳表现文章

这是一篇法律类文章，分析了国内法律中“国家”所有的含义。

TFIDF 的最佳表现文章为第 8 篇，与 Jiagu 一致。

这些文章中第 8 篇和第 82 篇文本关键词出现频数明显更多，而 72 和 82 篇中的一些关键词并不是频数最高的内容。

结论

纵观这些文章和各个模型给出的结果，再加上之前模糊匹配和精准匹配的能力参数，可以发现在本次小规模文本环境中实验的 Jiagu、TextRank、TFIDF 性能相近，且较为优秀。Jiagu 和 TFIDF 甚至再一百三十多文章中只有一个模糊匹配的差别，而 LDA 和 LSI 实验效果则是远远落后的。它们并不适合较小规模的关键词提取。原因可能是 LDA、LSI 模型从分类上都属于主题模型，旨在较长的文章中分析得到出现频次不高且最为重要的隐含主题，再加上实验所用的数据集较小（result/corpus.txt），没有更完整全面的语料用于关键词提取，这也显现出主题类模型并不适合小文本抽取关键词的弊端。在接下来的实验中，可以将论文的正文作为输入，再把输出的关键词和作者给出的关键词做对比，在挖掘深层主题上 LDA、LSI 可能会更胜一筹。

关键词提取技术只是自然语言处理（Natural Language Processing，NLP）的一个中间环节，在它背后有着中文分词、词性标注、命名实体的一些基础工作，而在它的应用之上，又构建起了语音语义理解、机器等新的应用场景。NLP 是机器学习基础上的应用，尝试理解人类语言中旧词新用、递归嵌套的从句语法、知识偏差带来主观性判断差异等一系列问题，用数学中的概率、向量改变了“语言空间”的维度和广度，有了许多新的突破。希望不久的一天机器可以拥有“智能”，理解文字背后的含义。

参考文献

余珊珊，苏锦钿，李鹏飞. 基于改进的 TextRank 的自动摘要提取方法[J]. 计算机科学, 2016, 43(6):240-247
吴军.数学之美（第二版）
Python 自然语言处理实战：核心技术与算法 徐铭，刘祥，刘树春
夏天. 词向量聚类加权 TextRank 的关键词抽取[J]. 现代图书情报技术, 2017, 1(2):28-34
Python 程序设计 [美]戴维 I.施耐德 著 车万翔 译