《数据科学与工程算法基础》实践报告

报告题目：基于爬山算法的抽取式文本摘要提取

姓　　名：郑佳辰

学　　号：10182100359

完成日期：2021.1.4

摘要：

爬山算法是一种基于启发式的局部择优搜索算法，广泛用于求解各种优化问题，但是传统的爬山算法存在着很难得到近似全局最优结构和搜索次数多的缺陷。本文主要探索基于爬山算法的改进研究用于解决文本摘要生成问题的技巧和实践方法。首先介绍了爬山算法和文本摘要生成问题，然后研究了面向文本摘要生成问题的爬山算法改进研究，最后设计实验进行算法验证。本实践项目证明了改进的爬山算法用于文本摘要自动生成的可行性和有效性。

Abstract:

As a heuristic based search algorithm, the hill climbing l algorithm is widely used to solve various local optimization problems. However, the traditional hill climbing algorithm suffers from its incapability of obtaining approximate global optimal structure and spends too much time on searching. In this report, the improved hill-climbing algorithms are explored and the techniques and practical methods and studied to solve the problem of text summary generation. Firstly, the hill climbing algorithm is introduced and the problem of text summarization generation is described. And then the improvements of hill climbing algorithms are studied on the problem of text summary generation. Finally, experiments are designed and the feasibility and effectiveness of the improved hill-climbing algorithms are verified on the automatic generation of text summarization.

一、 项目概述

在搜索引擎中，广为使用的是 PageRank 算法，该算法会根据指向某一页面的其他页面数决定该网页在搜索结果中的排名。该排名有时可通过向搜索引擎公司交一笔价格不菲的搜索费来提高，然而一些小公司无法在本就微薄的利润中再抽出一部分用于向搜索引擎缴费来获得并不显而易见的推广。为提高网页排名，小公司们转而把目光投向了 PageRank 算法本身，即提高指向自家页面的链接数。为此，他们需要在各大平台发内容不同的软文并附上链接以推广自家公司，提高 PageRank，可软文的生成又谈何容易。本次实验需编写的的摘要程序即是为了生成对某一事件的不同摘要，使这些不同的内容通过平台的审核程序。这样既可以提高其网站的 PageRank，又可为企业节约大笔资金。

文本摘要算法是在给定一个或多个文本数据源后，抽取关于文档内容的关键信息。它可让用户不必阅读整个文档，而从文本摘要中获取其需要的所有关键信息点，从而节约用户的的时间和精力。本次实验中的任务同样需要提取出核心信息。按照输入类型划分，文本摘要类型可以分为两种。单文档摘要针对单个文档进行内容抽取和总结从而生成摘要，多文档摘要根据包含多个文档的集合生成一份概括文档中心内容的摘要形式。按照摘要的输出结果划分，文本摘要可分为抽取式和生成式两种。抽取式摘要的内容是将原文档的关键词和句子选择并重组产生。生成式摘要是在原文档的基础上，允许生成新的词和短语组成摘要。在使用机器学习方法时，文本摘要算法可按是否包含类别标签数据分为有无监督两种。有监督摘要在句子层面本质上可以看作一个二分类问题。它

会选取文本中的主要内容作为训练数据，其中属于摘要的句子看作正样本，不属于摘要的句子作为负样本。这种方法需要大量的注释和标签数据参与模型训练。无监督摘要方法在训练过程中不使用现成的摘要，而是通过对文档进行检索直接生成摘要。在本次实验中，程序在无监督情况下生成多文档抽取式摘要。

二、 问题描述

本次实践活动要解决的主要问题是抽取式文本摘要问题。其生成方法大致是识别出文本中的重要部分并且抽取出来作为文本的摘要。本次实验中由于实践项目背景的设置，故增设以下要求。首先，本次实验是对同一特定话题生成摘要，所以需要大量语料，每次生成语料时，从中选取多个作为提取摘要的原文档。其次，由于需要采用爬山算法进行最终实现，所以需要进行关键词的提取。在这种情况下生成的文本摘要需要在满足覆盖所有关键词的条件下，最小化选取的句子数量。再次，该摘要属于无监督摘要，最终的摘要效果没有明确指标，需要自己构建指标并检验。最后，由于生成出的摘要主要作为软文发布至论坛上，即其阅读对象主要是机器，所以对于其可读性要求不高，不必符合复杂的语法规范并适合人类的阅读习惯。换句话说，本次实验生成的摘要只需让机器识别，不需做到日常摘要的概括准确，符合语法，无歧义等特性。

利用爬山算法进行文本摘要，大致可以分为以下几个步骤。先抽取文档中的句子并构建包含文本主要信息的文本表征。在本次实验中选取的文本表征信息是其所含的关键词。然后基于构建的文本表征对句子进行评分。本步骤会进行多轮，每轮选取目前的最优值组成最终的摘要。

三、 方法

使用爬山算法进行文本摘要的流程大概包含以下几个主要步骤：文本数据获取、文本分句、文本选择、关键词提取、文本摘要生成、文本摘要的调整及输出。下面依次对这些步骤进行分析，并将部分源代码附在相应段后便于对照。

本次实验采用的原始数据存放在文件 content.txt 中，其来自之前自然语言实验中的语料库（非开源）。数据集中的内容为商业新闻，爬取自搜狐新闻等多个门户网站，并经过了简单的预处理。数据集中每行为一条商业新闻，经过筛选共有五千多条长度合适，比较干净的语料。这些数据虽全部为商业新闻，但其中提到的内容多有不同。需要对这些数据进行简单聚类或分类，将有相似提及内容的新闻提取出来。

进行简单分类的代码位于 classify.py 中。首先调用 load_content_data 函数加载数据。然后分别对每条语料提取一定数量的关键词并记录。然后排序找出出现频率最高的几个关键词，将含有这些关键词的语料写入新文件中。即只要任意两篇新闻含有相同的关键词，就认为它们是可以被分到同一类。其中，除去一些对于聚类没有意义的关键词，如%、11月、中等词外，出现频率较高的关键词包括中国、公司、市场、投资等。根据任意一个关键词可以生成相应主题的摘要数篇，以下以关键词中国为例演示生成摘要的算法，此时的数据保存在文件 content 中国.txt 中。

```
if __name__ == '__main__':
    text = load_content_data()
    keywords = []
    dict = {}
    k = 20
    for i in range(len(text)):
        document = text[i]
        keyword = HanLP.extractKeyword(document, k)
        for word in keyword:
            if word not in dict.keys():
                dict[word] = []
            dict[word].append(i)
        keywords.append(keyword)
        if i % 200 == 0:
            print(i)
            print(keyword)

    dict = sorted(dict.items(),key=lambda x:len(x[1]),reverse=True)
    print(dict)

    for i in range(10):
        entry = dict[i]
        output = open('content'+entry[0]+'.txt', 'w', encoding='utf-8')
        for j in entry[1]:
            output.write(text[j])
        output.close()
```

选择出合适的语料之后，还需对这些新闻进行分句，这样才可抽取出含关键字的句子。进行分句的文件为 getsentence.py，生成的文件为 sentences 中国.txt。分句时使用了 NLP 开源工具 LTP 库。LTP 是由哈工大研发的语言技术平台，它提供了一系列中文自然语言处理工具，用户可以使用这些工具对于中文文本进行分词、词性标注、句法分析等工作。本文件中使用的是其分句函数。LTP 中自带的分句功能在商业新闻表现并不尽如人意。商业新闻中含有大量数据，而 LTP 的分句功能会将数据中的小数点也当成句号，所以需要在分句完成后进行判断。若末尾是英文句号，则需将其与后一句话连接，并输出至 txt 中。段尾加入空行用以分隔。

```
ltp = LTP()
input = open('content中国.txt', 'r', encoding='utf-8')
sentences = open('sentences中国.txt', 'w', encoding='utf-8')
text = []
for eachline in input:
    sents = ltp.sent_split([eachline])
    for sent in sents:
        if sent[len(sent) - 1] == '.':
            sentences.write(sent)
        else:
            sentences.write(sent + '\n')
    sentences.write('\n')
```

生成摘要的主体程序在 summary.py 中。变量 N 为程序生成的摘要个数，首先调用 load_data 函数加载好刚刚分好句的数据。然后对于每一篇摘要，依次选择用于生成摘要的文本并生成摘要。选择文本即为从所有新闻中随机选出 10 篇用于生成摘要。

```
def load_data():
    input = open('sentences中国.txt', 'r', encoding='utf-8')
    texts = []
    text = []
    for eachline in input:
        if eachline == '\n':
            texts.append(text)
            text = []
        else:
            eachline = eachline[:-1]
            text.append(eachline)
    return texts
```

然后进行生成摘要之前的预处理。将加载出的新闻整理成字符串的格式，利用 hanlp 中的返回文本关键字的函数求得这 10 篇新闻的共 20 个关键词。由清华大学开发的 pyhanlp 是面向生产环境的多语种自然语言处理工具包，基于 TensorFlow 2.x，目标是普及落地最前沿的 NLP 技术。HanLP 具备功能完善、性能高效、架构清晰、语料时新、可自定义的特点。在获得关键词之后，需要找出每个句子包含的关键词并存成矩阵。利用 LTP 进行分词并进行匹配，若该关键词出现在该句中，则记为 1，否则记为 0。若一句话中没有任何关键词，则这句话在以下抽取式

摘要的生成过程中被认为是没有用的，不必存储这句话。这样就减少了已有的数据规模。该矩阵存放在 appear 列表中。

```
def summary(passage):
    k = 20
    ltp = LTP()
    document = ''
    sentences = []
    for eachtext in passage:
        for eachline in eachtext:
            document = document + eachline
            sentences.append(eachline)
    keywords = HanLP.extractKeyword(document, k)
    summary_test = HanLP.extractSummary(document, 10)
    # print('\n')
    # print(document)
    print(keywords)
    non_list = [0 for i in range(k)]
    appear = []
    sents = []
    for sentence in sentences:
        appear_line = []
        seg, hidden = ltp.seg([sentence])
        for keyword in keywords:
            if keyword in seg[0]:
                appear_line.append(1)
            else:
                appear_line.append(0)
        if appear_line != non_list:
            appear.append(appear_line)
            sents.append(sentence)
```

本次实验中生成摘要的思路是从原文中抽取若干句子，使得这些句子能够覆盖原文中所有关键词，并且抽取出的句子数尽量少。依此思路编写爬山算法。算法执行过程如下：用 cover 列表记录关键词的覆盖情况，delta_si 记录在加入这一句之后，关键词数量的增加数。算法迭代多轮，每一轮中先遍历 appear 矩阵中的每行，更新 delta_si 并找出本轮 delta_si 最大的一个句子。然后将此句话加入已选择的列表并更新关键词覆盖的 cover 列表。如果最大的 delta_si 为 0，说明所有句子都不能使覆盖的关键词增加，此时爬山算法执行完毕。之后调整语序，将选出的句子按照原始顺序排列，计算抽取率并输出。

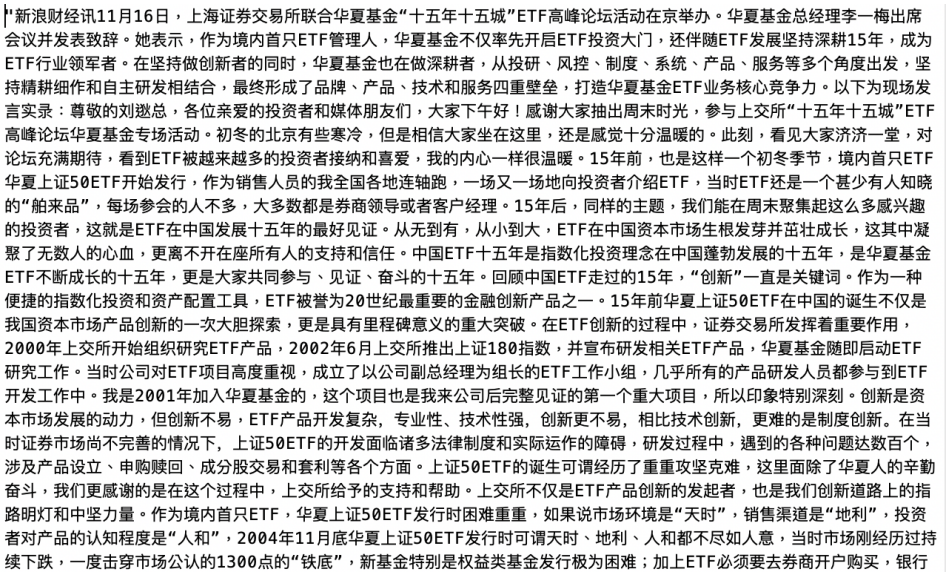
```
cover = [0 for i in range(k)]
delta_si = [-1 for i in range(len(sentences))]
summarynum = []
maxi = 0
while True:
    for i in range(len(appear)):
        if delta_si[i] == 0:
            continue
        s = 0
        for j in range(k):
            if (cover[j] == 0) and (appear[i][j] == 1):
                s += 1
        delta_si[i] = s
        if delta_si[i] > delta_si[maxi]:
            maxi = i
    if delta_si[maxi] == 0:
        break
    summarynum.append(maxi)
    for j in range(k):
        if appear[maxi][j] == 1:
            cover[j] = 1

summarynum.sort()
print(summarynum)
summary = ''
for i in summarynum:
    summary = summary + sents[i]
print(summary)
```

以上步骤执行完成后，程序就成功生成了一篇摘要。要满足实验要求，还需生成更多摘要以备发表。在主程序中，设置为生成 10 篇摘要。该数值可增加至选用产品的用户满意为止。对于算法的有效性，分析和具体性能放在下一小节讲解。

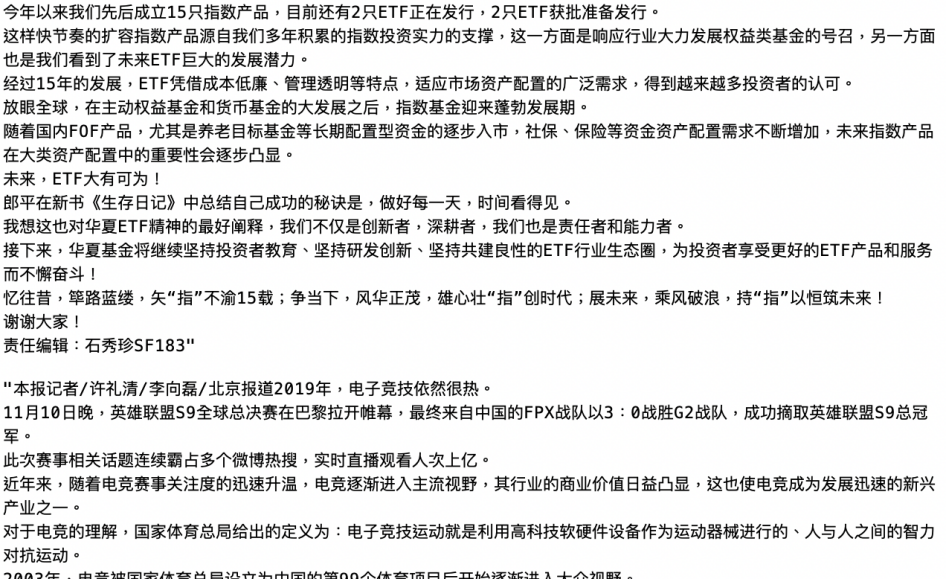
四、 实验结果

实验中使用的原始数据及其分类后结果如下图所示，该图片截取自 content 中国.txt。可以看出，此时的语料内容较长，不方便直接进行处理，需要进行分句。



"新浪财经讯11月16日，上海证券交易所联合华夏基金“十五年十五城”ETF高峰论坛活动在京举办。华夏基金总经理李一梅出席会议并发表致辞。她表示，作为境内首只ETF管理人，华夏基金不仅率先开启ETF投资大门，还伴随ETF发展坚持深耕15年，成为ETF行业领军者。在坚持做创新者的同时，华夏基金也在做深耕者，从投研、风控、制度、系统、产品、服务等多个角度出发，坚持精耕细作和自主研发相结合，最终形成了品牌、产品、技术和服务四重壁垒，打造华夏基金ETF业务核心竞争力。以下为现场发言实录：尊敬的刘逖总，各位亲爱的投资者和媒体朋友们，大家下午好！感谢大家抽出周末时光，参与上交所“十五年十五城”ETF高峰论坛华夏基金专场活动。初冬的北京有些寒冷，但是相信大家坐在这里，还是感觉十分温暖的。此刻，看见大家济济一堂，对论坛充满期待，看到ETF被越来越多的投资者接纳和喜爱，我的内心一样很温暖。15年前，也是这样一个初冬季节，境内首只ETF华夏上证50ETF开始发行，作为销售人员的我全国各地连轴跑，一场又一场地向投资者介绍ETF，当时ETF还是一个甚少有人知晓的“舶来品”，每场参会的人不多，大多数都是券商领导或者客户经理。15年后，同样的主题，我们能在周末聚集起这么多感兴趣的投资者，这就是ETF在中国发展十五年的最好见证。从无到有，从小到大，ETF在中国资本市场生根发芽并茁壮成长，这其中凝聚了无数人的心血，更离不开在座所有人的支持和信任。中国ETF十五年是指数化投资理念在中国蓬勃发展的十五年，是华夏基金ETF不断成长的十五年，更是大家共同参与、见证、奋斗的十五年。回顾中国ETF走过的15年，“创新”一直是关键词。作为一种便捷的指数化投资和资产配置工具，ETF被誉为20世纪最重要的金融创新产品之一。15年前华夏上证50ETF在中国的诞生不仅是我国资本市场产品创新的一次大胆探索，更是具有里程碑意义的重大突破。在ETF创新的过程中，证券交易所发挥着重要作用，2000年上交所开始组织研究ETF产品，2002年6月上交所推出上证180指数，并宣布研发相关ETF产品，华夏基金随即启动ETF研究工作。当时公司对ETF项目高度重视，成立了以公司副总经理为组长的ETF工作小组，几乎所有的产品研发人员都参与到ETF开发工作中。我是2001年加入华夏基金的，这个项目也是我来公司后完整见证的第一个重大项目，所以印象特别深刻。创新是资本市场发展的动力，但创新不易，ETF产品开发复杂，专业性、技术性强，创新更不易，相比技术创新，更难的是制度创新。在当时证券市场尚不完善的情况下，上证50ETF的开发面临诸多法律制度和实际运作的障碍，研发过程中，遇到的各种问题达数百个，涉及产品设立、申购赎回、成分股交易和套利等各个方面。上证50ETF的诞生可谓经历了重重攻坚克难，这里面除了华夏人的辛勤奋斗，我们更感谢的是在这个过程中，上交所给予的支持和帮助。上交所不仅是ETF产品创新的发起者，也是我们创新道路上的指路明灯和中坚力量。作为境内首只ETF，华夏上证50ETF发行时困难重重，如果说市场环境是“天时”，销售渠道是“地利”，投资者对产品的认知程度是“人和”，2004年11月底华夏上证50ETF发行时可谓天时、地利、人和都不尽如人意，当时市场刚经历过持续下跌，一度击穿市场公认的1300点的“铁底”，新基金特别是权益类基金发行极为困难；加上ETF必须要去券商开户购买，银行

而 sentence 中国.txt 文件中的数据格式如下图所示，每篇不同的新闻用空行分隔，已经分句完毕，便于处理。



今年以来我们先后成立15只指数产品，目前还有2只ETF正在发行，2只ETF获批准备发行。
这样快节奏的扩容指数产品源自我们多年积累的指数投资实力的支撑，这一方面是响应行业大力发展权益类基金的号召，另一方面也是我们看到了未来ETF巨大的发展潜力。
经过15年的发展，ETF凭借成本低廉、管理透明等特点，适应市场资产配置的广泛需求，得到越来越多投资者的认可。
放眼全球，在主动权益基金和货币基金的大发展之后，指数基金迎来蓬勃发展期。
随着国内FOF产品，尤其是养老目标基金等长期配置型资金的逐步入市，社保、保险等资金资产配置需求不断增加，未来指数产品在大类资产配置中的重要性会逐步凸显。
未来，ETF大有可为！
郎平在新书《生存日记》中总结自己成功的秘诀是，做好每一天，时间看得见。
我想这也对华夏ETF精神的最好阐释，我们不仅是创新者，深耕者，我们也是责任者和能力者。
接下来，华夏基金将继续坚持投资者教育、坚持研发创新、坚持共建良性的ETF行业生态圈，为投资者享受更好的ETF产品和服务而不懈奋斗！
忆往昔，筚路蓝缕，矢“指”不渝15载；争当下，风华正茂，雄心壮“指”创时代；展未来，乘风破浪，持“指”以恒筑未来！
谢谢大家！
责任编辑：石秀珍SF183"

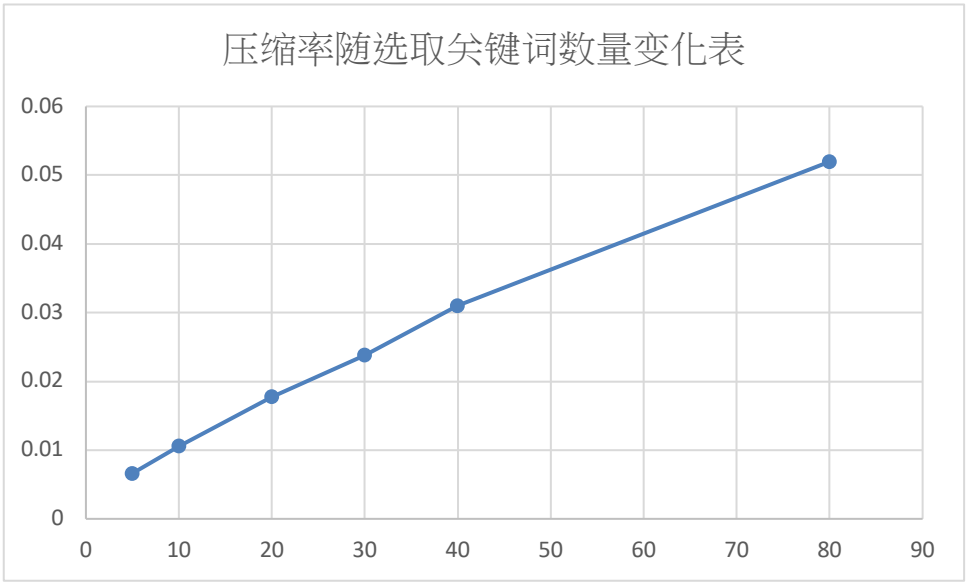
"本报记者/许礼清/李向磊/北京报道2019年，电子竞技依然很热。
11月10日晚，英雄联盟S9全球总决赛在巴黎拉开帷幕，最终来自中国的FPX战队以3：0战胜G2战队，成功摘取英雄联盟S9总冠军。
此次赛事相关话题连续霸占多个微博热搜，实时直播观看人次上亿。
近年来，随着电竞赛事关注度的迅速升温，电竞逐渐进入主流视野，其行业的商业价值日益凸显，这也使电竞成为发展迅速的新兴产业之一。
对于电竞的理解，国家体育总局给出的定义为：电子竞技运动就是利用高科技软硬件设备作为运动器械进行的、人与人之间的智力对抗运动。
2003年，电竞被国家体育总局设立为中国的第99个体育项目后开始逐渐进入大众视野。

运行程序所得结果如下图所示。对于每条抽取的摘要，会生成三行数据，其中第一行为组合后文段的关键词，第二行为抽取的句子的编号。

第三行为生成的摘要本身。最后输出的是所有压缩率及其平均值。本实验中，压缩率用摘要选取的句子数量除以文段中的句子总数得到。

```
[中国, 酒店, 公司, 中, 市场, 投资, 11月, %, 产品, 汽车, 称, 目前, 全球, 企业, 表示, 显示, 马拉松, 连锁酒店, 约, 发展]
[110, 161, 253, 256, 257, 261, 272]
市场份额被同类酒店挤占中国饭店协会联合盈蝶咨询发布的《2019中国酒店连锁发展与投资报告》显示，截至2019年1月1日，中国连锁酒店品牌规模最大的是如家酒店，拥有门店2253家，客房233226间；其次是汉庭酒店，拥
[中国, 华为, 表示, 市场, 发展, 汽车, 任正非, 11月, 公司, 中, 德国, 美国, 服务, 说, %, 欧洲, 手机, 生态, 国家, 目前]
[0, 32, 37, 40, 52, 146, 175, 180]
"MateX今天再次开卖，最高被炒卖至近10万元 华为：正全力生产11月18日消息，华为官方给出最新公告称，MateX5G折叠屏手机今天继续开卖，售价是16999元（8+256GB版本），购买的用户可以获得半价换屏服务，至于本
[网络安全, 机场, 中国, 电竞, 产业, 发展, 行业, 俱乐部, 全球, %, 安全, 国内, 我国, 公司, 会, 企业, 泰国, 已经, 目前, 投资]
[13, 18, 28, 49, 121, 173, 192]
然而，事实却是，泰国机场不仅是全球资本市场的明星标的，它还是全世界市值最大的机场上市公司，其在泰国证券交易所上市，目前市值约2300亿元！2018年，入境泰国的外国游客再创历史新高，达到3800万人次，其中中国
[中国, 发展, 市场, 经济, %, Burberry, 动力电池, 行业, 企业, 金融, 表示, 领域, 材料, 方面, 电池, 产品, 汽车, 开放, 我国, 合作]
[11, 26, 32, 73, 86, 181]
新能源汽车销量下降，在汽车零部件中首当其冲的便是动力电池企业，中国化学与物理电源行业协会动力电池应用分会数据显示，2019年10月我国动力电池装机量约4.07GWh，同比下降31.35%，动力电池装机量连续第三个月同
[中国, 市场, ETF, 经济, 机场, %, 中, 表示, 投资, 基金, 我国, 发展, 行业, 投资者, 消费, 改革, 政策, 风险, 资本, 三星]
[7, 42, 166, 231, 238, 330]
高培勇：决不能把稳增长的全部希望寄托于逆周期调节（中国社科院副院长、学部委员高培勇）中国社科院副院长、学部委员高培勇表示，十九届四中全会要求，“坚持和完善社会主义基本经济制度，推动经济高质量发展”，“健全
[中国, 市场, 经济, 发展, 企业, 会, 行业, 公司, %, 产品, 康养, 说, 旅居, 项目, 目前, 利率, 投资, 旅游, 健康, 增长]
[101, 110, 126, 137, 181, 189, 210]
中国社科院保险与经济发展研究中心秘书长王向楠解释称，如果监管对此不加以限制，寿险行业恐将形成较大的利差损积累，因为对于国民经济所能产生的投资回报率，很多公司对预期的调整是滞后或缓慢的。尤其是在利率下行、
[中国, 驾驶, 自动, 美国, 发展, 科技, 市场, 表示, 企业, 特朗普, 中, 实现, 材料, 投资, 称, 院士, 创新, 产业, 未来, 广州]
[15, 18, 63, 112, 123, 168]
"作者 | 王瑞编辑 | 郝秋慧实现完全自动驾驶，为时尚早。自动化学会理事长、中国工程院院士郑南宁表示，辅助安全驾驶，结构化环境的无人驾驶，和一些应用背景明确的无人驾驶任务，目前已经得到实现，但如何实现完全自主
[中国, 网络安全, 市场, 发展, 制度, 三星, 中, 产品, 华为, 产业, 社会主义, 安全, 国家, ETF, 特色, 手机, 任正非, %, G, 我国]
[22, 64, 103, 191, 247]
正是因为中国特色社会主义制度具有显著优势，能够提供坚强保障，才使得一个“一穷二白”的经济文化落后的国家，在70年的发展进程中创造出令世界震惊的“中国奇迹”。据分析机构IDC公布的2019年第三季度全球智能手机出货
[中国, 市场, 经济, 表示, 发展, 三星, 金融, 投资, 企业, 中, 银行, 改革, 行业, 我国, 风险, 政策, 当前, 技术, 需要, 汇率]
[299, 336, 370]
崂山区政府党组成员许哲在致辞中表示：当前，崂山区正在实施“高端++虚拟现实”攻势行动方案，发布了《崂山区人工智能行动计划2019-2021》，出台了《青岛市崂山区关于加快新旧动能转换打造高端促进制造业发展的实施细
[0.011194029850746268, 0.015521064301552107, 0.02564102564102564, 0.02194357366771116, 0.020761245674740483, 0.011029411764705883, 0.022222222222222223, 0.025751072
0.017474078933856257
```

首先观察压缩率，上图中显示的是选取了 20 个关键词时，摘要的平均压缩率为 0.1747。调整选择的关键词个数并重复多次实验，研究在选取篇数相同时，选取的关键词个数对压缩率的影响。由于每篇语料的质量不同，得到的压缩率可能大不相同。所以采用在进行了 100 次文本摘要之后得到的平均值，用来代表平均水平。将不同关键词个数对应的压缩率情况制表如下。



从图中可以看出，在本实验所选取的区间内，压缩率随选取的关键

词数基本上是线性增加。即选取的关键词越多，抽取的摘要就越长，这和我们的一贯认知相符。可预见地，压缩率不会无限上升。随着关键词数量不断上升，压缩率最终会偏离线性分布。事实上，在取 80 个关键词时，已经略微出现斜率下降。另一方面，随着选取的关键词数量的增加，程序的运行速度没有明显下降。这说明该爬山算法运行时的复杂度随关键词数量变化较小，主要和语料本身的大小有关。

将经算法生成的摘要与 hanlp 自带的摘要功能对比，可以发现爬山算法与 hanlp 的抽取式摘要的区别主要体现在以下方面。爬山算法的抽取以整句为单位，而 hanlp 以分句为单位。爬山算法选取的摘要的句子数不固定，主要取决于关键词的覆盖情况，而 hanlp 选取的分句数由用户设置的固定值决定。并且 hanlp 抽取出的句子是多个文本中相关的部分，并会对选取出的句子进行重排，以适应正确的逻辑推理顺序。总体来说，hanlp 的抽取效果较好。下图是部分生成结果的对比。其中每组最下方在中括号内的是 hanlp 的生成结果，倒数第二行是爬山算法生成的结果。

[中国, 市场, ETF, 数字, 货币, 邮轮, 发展, 国际, 铁矿石, 全球]
[122, 193, 204, 265]
根据《邮轮绿皮书：中国邮轮产业发展报告（2019）》指出，2018年，上海、天津邮轮港口游客接待量呈现一定下降趋势，其中上海吴淞口国际邮轮港游客接待量下降6.84%，天津国际邮轮母港游客接待量同比下降27.49%，不
[国际邮轮公司看好华南市场第十四届中国邮轮发展大会暨国际邮轮博览会15日在广州开幕，国际邮轮公司重新加码中国市场，15年前华夏上证50ETF在中国的诞生不仅是我国资本市场产品创新的一次大胆探索，将有更多的优质

[中国, 中, 公司, %, 微, 美国, 刻蚀, 设备, 发展, 全球]
[9, 173, 174]
高纪凡在大会上发言高纪凡回忆说，全国工商联一直重视支持民营经济健康发展和民建企业健康成长，特别重视支持企业参与绿色发展，十几年前就建立了新能源商会，他本人荣幸当选常务会长，后来转任中国光伏行业协会理事长
[中微公司在全球介质刻蚀设备市场中占有2.5%的市场份额，中微公司还是国家集成电路产业基金成立后投资的第一家公司，在全球电容性刻蚀设备市场中占有1.4%的市场份额，中微7纳米介质刻蚀设备研制成功，开始投资支持

[中国, 垃圾, 市场, 分类, 公司, 汽车, 11月, 解说, 企业, 中]
[38, 192, 196]
"大事件10月全国汽车产销跌幅收窄，新能源市场大跌46%11月11日，中国汽车工业协会发布了上月中国汽车工业产销数据。李皓：但是很遗憾的就是，改革开放的时候，像这些再生资源回收公司全部市场化以后就解散了这个国有
[垃圾分类是中国最开始，随着外资酒店品牌OYO进入中国市场，中国城市生活垃圾年产生量为1.18亿吨，李皓博士是中国最早推动垃圾分类的学者之一，外资开始快速进入中国市场，消息称特斯拉计划在2019年底之前生产1

[中国, 陈启宗, 陈文博, 市场, %, 商场, 品牌, 恒隆, 全球, 公司]
[10, 14, 22, 30, 78]
而整场对话，也由此徐徐展开——中国消费者与中国品牌"对中国消费者的看法"，陈启宗开始向陈文博提出第一个问题。按照陈启宗提供的数据，目前，在世界部分大牌的全球消费中，中国消费者所占的份额大概介于36%-37%，而
[无论是关于中国消费者与中国品牌、实体与电商的讨论，陈启宗开始向陈文博提出第一个问题，也由此徐徐展开——中国消费者与中国品牌，关于中国消费者与中国品牌的讨论还停留在商业探索层面，陈文博一直以为到恒隆旗下

[中国, 发展, 公司, 城市, %, 市场, 物流, 空间, H, 地下]
[72, 132, 148]
图片来源：《中国城市地下空间发展蓝皮书2019》发展，亦将显著提升站域沿线价值，带来大规模人流、物流、流、，并带动基础设施、生活、商务办公的集聚开发。"一方面，对H股上市公司来说，实现全流通可以在资本市场上
[《中国城市地下空间发展蓝皮书2019》发展，指引中国城市地下空间未来发展，《中国城市地下空间发展蓝皮书2019》""原标题，《中国城市地下空间发展蓝皮书2019》发展综合实力强劲，其他9座城市均位于中国地下空间

本次实验采用的爬山算法是一种基于子模函数的近似算法，其近似

效率优于最优解的（1 - 1 / e）= 63%。总体来说，爬山算法通过多次选取局部最优进行迭代来接近总体最优。在通常情况下，爬山算法得到的结果非常接近最优解。在本次文本摘要的过程中，随着关键词被选取，每个句子包含的关键词变化数越来越少。这符合课上所讲的边际递减效应。从算法运行的结果来看，采用爬山算法的近似效果较好。选取出的句子数量不多，覆盖情况较好。

五、 结论

本次实验中存在以下不足。在数据方面，找寻的语料虽均为商业新闻，然主题过于分散，需要聚类。为保证数据量，聚类算法实现较为简单朴素，聚类情况并不是很好。并且为使语序连贯，分句时使用句号分隔，导致每句话较长。此外，关键字的提取效果与使用的 NLP 开源工具有关。不同的 NLP 开源工具可能会导致不一样的结果。仅仅要求摘要覆盖所有关键词而不顾其逻辑关系也会带来问题。

除以上原因之外，爬山算法本身也具有一些缺陷。爬山算法本质上是使用了贪心策略的一种近似算法。它求得的是这一类覆盖问题的近似解而非精确解。对于精确解的求解需要使用动态规划等算法，思考难度大，且时间复杂度相对较高。近似解求取时可以引入随机因素，即使用模拟退火算法。另外，对于抽取式摘要来说，仅仅覆盖关键词的做法有时并不能很好的总结出这一段文字的大意，其生成结果可能并不能让人理解。并且由于本次实验要求是多文本摘要，而这种方法无法找出文本之间的相通之处，更适合单文本摘要。同时在一些情况下，抽取式摘要的摘要效果要差于生成式摘要，这也是本次实验中摘要效果不好的其中一个原因。

针对本次实验中的不足，未来有以下改进方向。一是获取规模更大，相关程度更高的语料，或是改进聚类算法，用以提高文本的相关性。二是可以尝试换用更好的 NLP 开源工具，以达到更好的分句和关键词提取效果。三是在现有算法中进行改进，引入随机化因素，通过模拟退火算法求取更好的近似解。四是整体修改算法，改用 LSTM 神经网络等深度

学习算法来提取摘要，这样既可采用抽取式摘要，又可采用生成式摘要。五是使用多文本摘要算法，这类算法相比单文本摘要更符合本次实验的任务。

综上所述，使用爬山算法进行抽取式文本摘要有一定合理性，但不是最佳选择。爬山算法可以在给定关键词的情况下，对找出覆盖所有关键词的最小的句子集合这一问题求取它的的近似最优解。在实际的摘要生成问题中，要获取更加合理有效的摘要，还需使用更为复杂的 NLP 工具和算法。