气候变化文献数据挖掘

员司雨 17307110448

摘要：气候与经济密切相关，为了更好的分析经济，我们需要阅读气候变化文献来了解世界地区气候变化。本项目通过 LDA 主题模型与文本分类模型，实现了气候变化文献自动化分类，并借助词云图清晰明了地分析了近二十年气候变化文献热点变迁，为经济研究提供参考。与此同时，本项目设计了 SQT-CNN 神经网络，以实现相似文献自动化判断，有效解决引用文献查找困难的问题，为经济金融乃至各学科领域的引用文献查找匹配提供方法。

关键词：词云图、LDA 主题模型、PCA、K-Means、SQT-CNN 神经网络

一. 项目意义与目的

气候与经济看似不相干实则联系紧密，因为天气影响所产生的的微小改变可能对整体产生巨大影响。举一个简单的例子：如果今年夏天的温度比去年夏天的温度要高，那么大约每十个德国人中就有一人会比去年多买一件短袖，如此一来，总共就额外多了八百多万件短袖，与此同时，到了冬天羽绒服的销售量就会减少[1]。气候变化虽然在短期时间内不会立刻影响我们的决策，但是从长期来看，将会在潜移默化中改变我们的消费行为。例如温室效应引发的全球变暖，使得更多家庭相较于之前更倾向购买大功率的空调来降温，而这种需求也激励制造商生成符合消费者要求的产品。

此外，值得注意的是，一个国家的气候对该国经济发展至关重要。首先气候直接决定农业的前提条件。显而易见，极地、高山和热带沙漠等极端气候地理环境下是无法产生蓬勃的经济的，因为土地的贫瘠会导致人口的匮乏，而经济发展的根本要素就是人口；其次季风性气候、海洋性气候与大陆性气候也决定了经济发展的周期和模式。季风性气候强调水资源的跨时空调用，对水库和水利设施要求较高，需要集体力量大、组织管理水平较高的经济模式，海洋性气候则比较容易受到气候大周期的影响，是最容易发展发展经济的，而大陆性气候则由于降雨较少，农业相对落后，很难发展人口，使得经济发展较为缓慢。由上述分析也可以得出，大部分发达国家都在温带海洋性气候区域和温带季风性气候区域的原因。英国灾难紧急救援委员会成员之一的蒂尔基金会近日发表一份报告说，受环境问题影响，全球已有 2500 万人被迫离开家园，如果不能有效解决这些问题，这一数字将在 50 年内增加到 2 亿。如果有大量百姓流离失所，势必会影响全球经济发展。

在施瓦茨的《气候经济学》中也提到[2]，气候变迁不会像海啸一样向我们袭来，或许它已经向我们发出一些警讯，而不是征兆。事实上，人类对自然的破坏也将使自然反扑而改变经济型态，不论这个力道大小，它正一点一滴的影响我们的消费。联合国近日发布《2019 年世界经济形势与展望》报告，列举了全球经济面临的一系列重大问题。报告指出，随着世界上出现越来越多极端天气事件，气候风险正在加剧。1998 年至 2017 年，气候相关灾害造成的损失高达 22450 亿美元，比 1978 年至 1997 年期间增长 151%。联合国首席经济学家艾略特·哈里斯在接受新华社记者专访时强调，气候变化不仅是环境问题，也是全球经济面临的重大风险。他呼吁，国际社会应减小经济发展对化石燃料的依赖，为世界构筑可持续发展的基础。

---

[1] Stern S N . What is the Economics of Climate Change?[J]. World Economics, 2006, 7(2):1-10..

[2] 施瓦茨郭晗聃. 气候经济学:影响全球 80%经济活动的决定性因素[M]. 气象出版社, 2012.

近些年，世界上有多位经济学家不约而同的发现并研究气候对经济的影响，甚至诞生了气候经济学这一经济学的边缘分支，由此可见气候对于经济研究的重要性。2018 年诺贝尔经济学奖授予美国经济学家威廉·诺德豪斯和保罗·罗默，正是表彰他们对探索技术创新和气候变化与经济增长关系的贡献。进入上世纪 70 年代，科学家们越来越担心化石燃料的大量使用将导致全球气候变暖，诺德豪斯开始从能源、环境角度研究气候变化的经济影响。为了定量考察经济和气候变化的关系，诺德豪斯带领耶鲁大学的一个团队，运用大量资料先后建立了两个分析经济对气候变化的“可计算一般均衡模型”——集成多地区经济系统、气候和地球物理系统研究气候变化经济影响的模型(RICE 模型)[3]和单一地区气候变化社会经济影响全面综合模型(DICE 模型)[4]。这两个模型可用来分析碳排放对气候变暖的影响，为减排的经济和环境效益分析提供实证依据。这也使他更坚定了以渐进式政策应对气候变化的观点。因为经济和气候变化应对具有某种耦合性，所以须特别注重两者之间的平衡关系。诺德豪斯尤其强调利用市场经济方式来应对气候变化，例如，给碳排放定价。值得一提的还有，诺德豪斯和 1981 年经济学诺奖得主詹姆斯·托宾在 46 年前就提出了净经济福利指标。他们主张把环境污染、国防开支和交通堵塞等经济行为所产生的社会成本从国内生产总值中扣除，同时加上一直被忽略的休闲、家政、社会义务劳动等经济活动。这对切实保障国民的福利，减少环境污染和城市交通堵塞，提高生活质量具有重要意义。

为了更好的分析经济，我们需要阅读气候变化文献来了解世界地区气候变化。然而，在这个大数据时代，文献整理归类工作量巨大，给文献管理员带来了很大的负担。因此本项目利用数据分析方法来解决上述问题，希望通过该项目的实验结果，一方面可以得出气候变化领域焦点问题，为经济研究提供参考，另一方为经济金融乃至各学科领域的引用文献查找匹配提供方法。

本项目的主要贡献：

- 通过 LDA 主题模型与文本分类，实现气候变化文献自动化分类，并结合词云图分析近二十年来气候变化文献的热点和主题变迁。
- 借鉴文本卷积神经网络模型，设计 SQT-CNN 神经网络，以实现相似文献自动化判断，有效解决引用文献查找困难的问题。

二. 数据来源与变量介绍

数据集由从 2000 年至 2020 年 5 月的 SCI 气候变化文献信息组成，主要信息介绍如表 1 所示，

英文缩写	解释含义
FN	文件名
VR	版本号
DT	文献类别
AU	作者缩写（英文）
AF	作者全名
BA	书籍作者
CA	团体作者

[3] William D. Nordhaus. Estimates of the Social Cost of Carbon: Background and Results from the RICE-2011 Model[J]. social science electronic publishing, 2011.

[4] Nordhaus, William. Evolution of modeling of the economics of global warming: changes in the DICE model, 1992 - 2017[J]. Climatic Change, 2018.

GP	书籍团体作者
AB	摘要
TI	标题
CR	被引用的参考文献
BE	编者
SC	学科类别
DA	生成此报告的日期
PY	出版年
DI	数字对象标识符(DOI)
ER	记录结束
EF	文件结束

表 1

为了防止过拟合，同时降低模型复杂度，本项目提取标题（TI）、摘要（AB）、文献类别（DT）、学科类别（SC）、作者全拼（AF）以及被引用的参考文献（CR）作为特征，并将 DOI 作为唯一索引定位文献。除去缺失上述特征的缺失数据，清洗得到 8526 条文献数据。整理如表 2 所示，

存储字典	索引	内容
context_dic	DOI	题目、摘要、文献类别、学科类别、作者全拼
refer_dic	DOI	引用文献的 DOI
time_dic	DOI	文献的出版年份

表 2

三. 描述性分析

在进行算法分析之前，我们首先需要对各变量进行描述性分析以了解数据构成，为后续算法分析做铺垫

3.1 文献类别

如图 1 所示，通过对数据集中各文献的类别进行统计，我们可以看出，大部分的文献类别为 article，而 review、proceedings paper 等其他文献类别仅占少部分。由此可以看出气候变化文献主要由 article 组成。

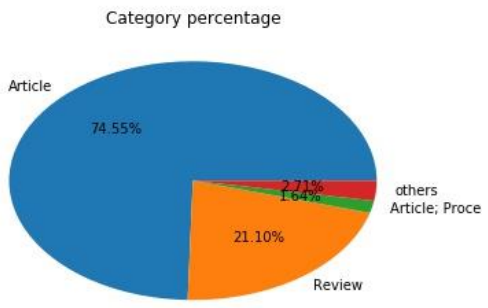

图 1：文献类别圆饼图

3.2 学科类别

如图 2 所示，通过对数据集中每个文献所属学科类别进行统计，并对数量最多的前五个类别进行可视化，我们可以看出，近二十年气候变化文献所属类别主要为“Science & Technology - Other Topics”、“Environmental Sciences & Ecology”、“Environmental Sciences & Ecology; Meteorology & Atmospheric Sciences”、“Biodiversity & Conservation; Environmental Sciences & Ecology”和“Meteorology & Atmospheric Sciences”。

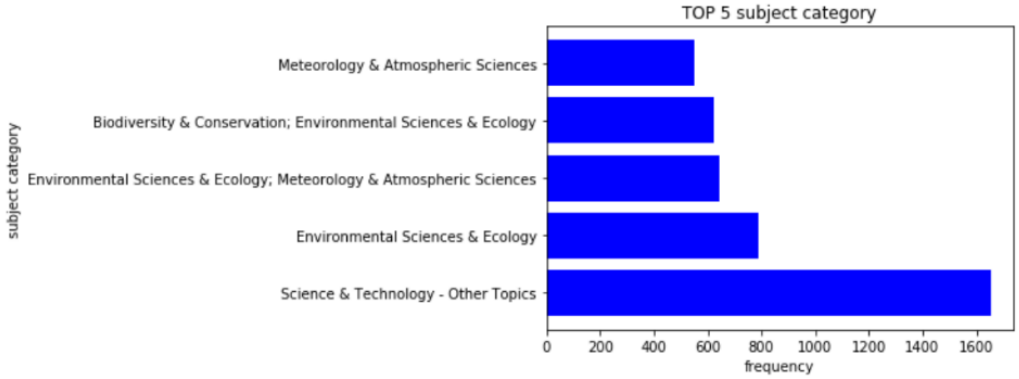

图 2：学科类别柱状图

3.3 引用文献

如图 3 所示，除去不存在数据集种的文献 DOI，对每个文献的引用文献数目进行统计，可以得出，气候文献引用文献数目基本不超过 20 篇，主要集中在 3 到 4 篇左右。

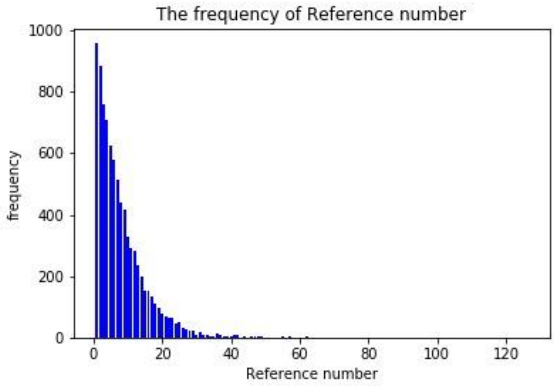

图 3：引用文献柱状图

3.4 词云图分析热点话题

使用 python 中的 wordcloud 对 2000 至 2020 年的文献摘要去除停用词并进行词云分析，最终以每五年为一簇得到词云图，如图 4 所示。

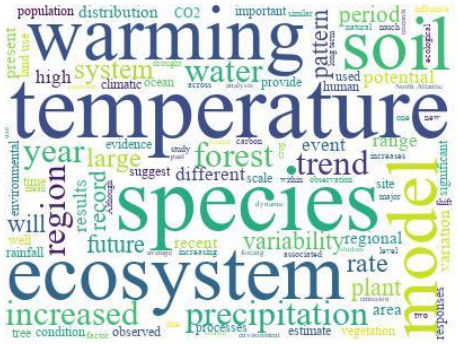

图 4(a)2000-2005 词云图

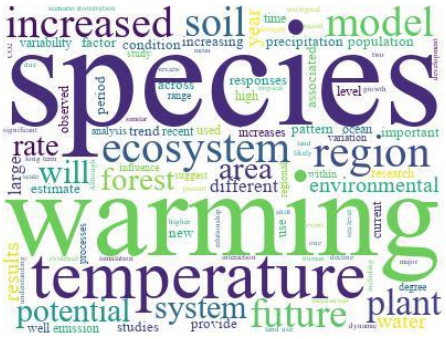

图 4(b)2006-2010 词云图

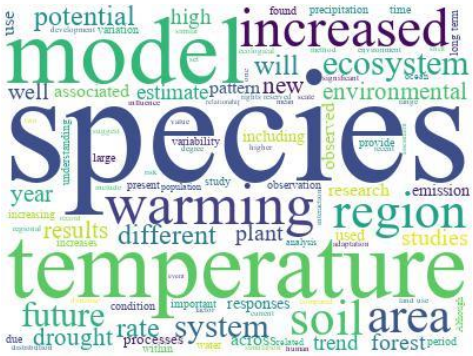

图 4(c)2011-2015 词云图

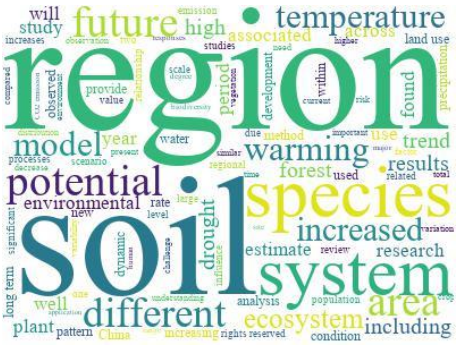

图 4(d)2016-2020 词云图

从上述词云图中可以看出，从 2000 年到 2020 年，这二十年间气候变化研究焦点呈现从温度到物种，再到地区土壤的变化。

上述变化其实与经济是密切相关的，进入千禧年之后，在全球变暖影响下，径流减少、蒸发增大，水资源系统不稳定性和水资源供需矛盾加剧。在干旱年份，气候变化产生的缺水量大大加剧各地区的缺水，并对社会经济产生严重影响。

在之后的十年里，由于人民对皮草、象牙等用品的追求，大量野生动物被猎杀，这严重导致了生态不平衡，从而极大影响了气候变化，而气候变化又进一步加剧了物种的灭绝。因此从 2006 年到 2015 年，气候变化文献大多数聚焦于气候与物种[5]。

而从 2016 年以后，文献开始更为关注地区土壤，更为准确详细的分析各个国家各个地区的气候对土壤的影响。土壤是农业的基础，而农业生产率的提高，是实行社会分工的基础。农业以外的一切事业的存在和发展，归根到底要取决于农业劳动生产率的提高而能够提供一定数量的剩余农产品。所以气候变化文献对土壤的关注其实也值得引起经济学家的重视。

四． LDA 主题模型

通过上述描述性分析，我们对该气候文献数据集已经有了初步的认识。为了跟进一步了解近 20 年来气候变化文献主题分布，我们需要使用 LDA 主题模型。

4.1 LDA 主题模型简介

LDA（Latent Dirichlet Allocation）是一种文档主题生成模型，也称为一个三层贝叶斯概率模型，包含词、主题和文档三层结构。所谓生成模型，就是说，我们认为一篇文章的

[5] Kardol P , Cregger M A , Campany C E , et al. Soil ecosystem functioning under climate change: plant species and community effects[J]. Ecology, 2010, 91(3):767-781.

每个词都是通过“以一定概率选择了某个主题，并从这个主题中以一定概率选择某个词语”这样一个过程得到。文档到主题服从多项式分布，主题到词服从多项式分布[6]。图 5 简要示意了 LDA 主题模型算法过程。

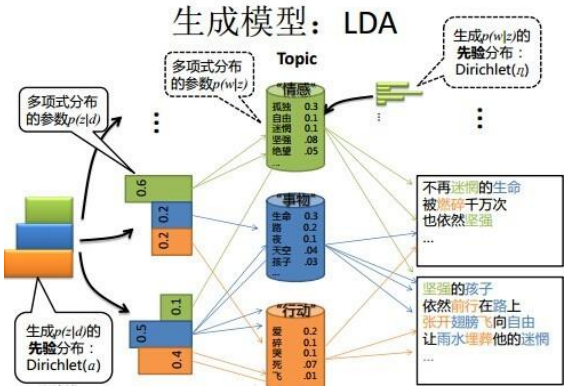

图 5：LDA 主题模型示意图[7]

4.2 气候变化文献主题分析

提取数据集中每条文献数据的题目与摘要生成文档集合，并使用 LDA 主题模型将文档分成 10 个主题，每个主题的 TOP5 关键词如表 3 所示，

主题序号	TOP5 关键词	主题
Topic 1	TC、LULC、AMOC、PV、use/land	土地利用
Topic 2	PM2.5、HAB、OC、buildings、NH3	有害物质
Topic 3	PCM、Baltic、indoor、livelihood、mycotoxins	海洋环境
Topic 4	species、plant、adaptation、sustainability、GHG	森林物种
Topic 5	haze、residential、CCS、equity、BC	空气污染
Topic 6	load、straw、fruit、deltas、inundation	农作分析
Topic 7	UHI、Anthropocene、green、cyanobacterial、cyanobacteria	水质分析
Topic 8	energy、water、change、climate、carbon	温室效应
Topic 9	flood、green、financial、SPEI、adsorption	暴雨与干旱
Topic 10	soil、biochar、SOC、innovation、farmers’	土壤污染

表 3

可以看出数据集中的气候变化文献最主要聚焦于土地利用，海洋环境，温室效应等主题，这也与词云图的结果相对应。

五. 文本分类模型

通过上述的 LDA 主题模型，我们得到了近 20 年气候变化文献的主题分布。实现气候变化文献自动化分类，我们需要用到 PCA 主成分分析算法和 K-means 聚类算法。

---

[6] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

[7] https://zhuanlan.zhihu.com/p/31470216

5.1 PCA 主成分分析算法介绍

在用统计分析方法研究多变量的课题时，变量个数太多就会增加课题的复杂性。人们自然希望变量个数较少而得到的信息较多。在很多情形，变量之间是有一定的相关关系的，当两个变量之间有一定相关关系时，可以解释为这两个变量反映此课题的信息有一定的重叠。主成分分析是对于原先提出的所有变量，将重复的变量（关系紧密的变量）删去多余，建立尽可能少的新变量，使得这些新变量是两两不相关的，而且这些新变量在反映课题的信息方面尽可能保持原有的信息。

设法将原来变量重新组合成一组新的互相无关的几个综合变量，同时根据实际需要从中可以取出几个较少的综合变量尽可能多地反映原来变量的信息的统计方法叫做主成分分析或称主分量分析，也是数学上用来降维的一种方法[8]。降维示意图如图 6 所示，

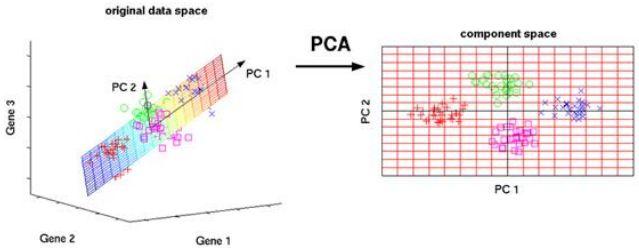

图 6：PCA 降维示意图[9]

5.2 K-Means 聚类算法介绍

k 均值聚类算法（k-means clustering algorithm）是一种迭代求解的聚类分析算法，其步骤是，预将数据分为 K 组，则随机选取 K 个对象作为初始的聚类中心，然后计算每个对象与各个种子聚类中心之间的距离，把每个对象分配给距离它最近的聚类中心。聚类中心以及分配给它们的对象就代表一个聚类。每分配一个样本，聚类的聚类中心会根据聚类中现有的对象被重新计算。这个过程将不断重复直到满足某个终止条件。终止条件可以是没有（或最小数目）对象被重新分配给不同的聚类，没有（或最小数目）聚类中心再发生变化，误差平方和局部最小[10]。

5.3 气候变化文献聚类

根据 LDA 主题模型结果，获取每个文档对每个主题的相关度，如文献 5 关于 10 个主题的相关度如表 4 所示，

Topic	与文献 5 的相关度
1	0.0084
2	0.0083
3	0.0084
4	0.0715
5	0.0100
6	0.0094
7	0.0131

[8] Yang J , Yang J Y . Why can LDA be performed in PCA transformed space?[J]. Pattern Recognition, 2003, 36(2):563-566.

[9] http://blog.csdn.net/zhongkelee/article/details/44064401

[10] by G Babu, M Murty. A near optimal initial seed value selection in kmeans algorithm using a genetic algorithm[C]// Pattern Recognition Letters. 1993.

8	0.8422
9	0.0122
10	0.0162

表 4

因此可以将该相关度作为代表文献 5 的 10 维向量，同理可得到每个文献的向量表示。将 10 维向量使用 PCA 降维得到 2 维向量，并使用 K-Means 聚类将数据集分为 10 类，分类结果可视化如图 7 所示，

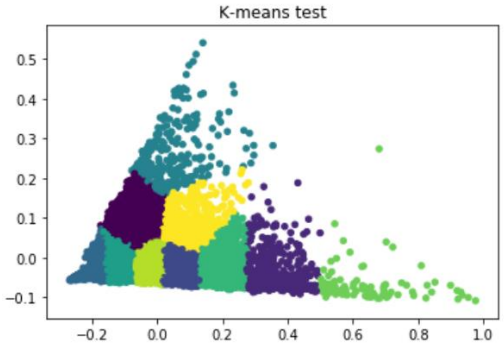

图 7：K-Mean 聚类结果

六. SQT-CNN 神经网络

通过文本分类，我们可以得到每个文献相似的文献集合。但是该方法仅仅用到文献的主题而忽略了学科类别、文献类型等信息。与此同时，虽然主题是根据文献题目与摘要获得的，但是在提取过程中也会丢失一些信息。为了更好的实现相似文献查阅，解决引用文献查找困难的问题，需要引入深度学习的方法。在本项目中，我基于文本卷积神经网络，设计了 SQT-CNN 神经网络。

6.1 文本卷积神经网络介绍

卷积神经网络（CNN）在图像处理领域取得了很大的成绩，它的卷积和池化结构能很好提取图像的信息，而在 NLP 领域循环神经网络（RNN）则使用的更多，RNN 及其各种变种因拥有记忆功能使得它们更擅长处理上下文。但 NLP 领域很多方面使用 CNN 取得了出色的效果，比如语义分析、查询检索、文本分类等任务，自从 2014 年 Kim.Y 提出 TxetCNN[11] 之后，使用 CNN 来做自然语言处理任务的工作越来越多了。

TxetCNN 神经网络算法的主要流程是通过使用 kernel_sizes 的卷积核对文本 embedding 二维向量进行卷积操作，每一种 kernel_sizes 的卷积核有多个，这样就可以获得类似 n-gram 的句法特征，最后经过 max_pooling，在进行一次拼接，即可以得到文档向量，经过全连接和 softmax 即可进行下游任务。具体算法示意图如图 8 所示，

[11] Kim Y . Convolutional Neural Networks for Sentence Classification[J]. Eprint Arxiv, 2014.

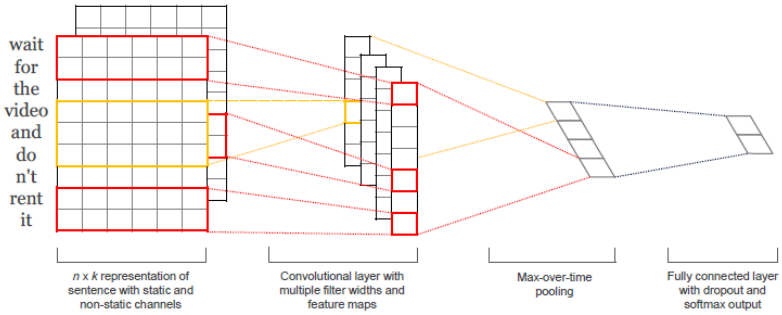

图 8：TextCNN 神经网络

6.2 SQT-CNN 神经网络

为了实现相似文献自动化判断，以解决引用文献查找困难的问题，我基于上述 TextCNN 神经网络和 Severyn A 提出的文本配对网络[12]，设计可实现相似文献查找的 SQT-CNN 神经网络（Similarity Query Text CNN）。具体算法实现过程示意图如图 9 所示。

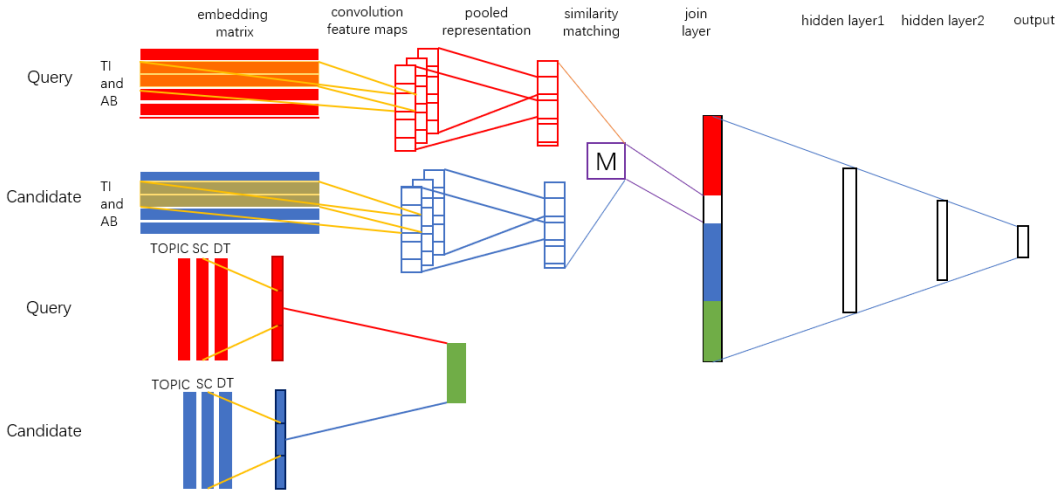

图 9：SQT-CNN 神经网络图示

通过对文献数据进行 embedding，我们得到词向量矩阵。接着，对题目和摘要共同构成的 embedding matrix 分别使用 kernel_sizes 为 3、4、5 的卷积核进行文本卷积，获得类似 n-gram 的句法特征，再经过 max-pooling 得到文献特征向量。经过卷积和池化后，实质上两个句子已经转为了两个句向量，句向量就可以开始进行相似度衡量了。我们构造一个相似矩阵 M，用于计算两者的相似度。

$$sim(x_q, x_c) = x_q^T M x_c$$

然后，将计算得到的相似度、两个句向量、以及学科类别、文章类别、主题这些额外特征进行拼接组合，得到一个向量，这个向量内涵盖了相似度、query 句向量、document 句向量以及额外特征 4 各方面信息，经过多层感知机，最终到达输出层。

[12] Severyn A , Moschitti A . Learning to Rank Short Text Pairs with Convolutional Deep Neural Networks[C]. the 38th International ACM SIGIR Conference. ACM, 2015.

6.3 实验过程

此部分主要介绍实验样本生成、baseline 构建以及实验结果。

6.3.1 Glove 词向量[13]

Glove 的全称叫 Global Vectors for Word Representation，它是一个基于全局词频统计（count-based & overall statistics）的词表征（word representation）工具，它可以把一个单词表达成一个由实数组成的向量，这些向量捕捉到了单词之间一些语义特性，比如相似性（similarity）、类比性（analogy）等。我们通过对向量的运算，比如欧几里得距离或者 cosine 相似度，可以计算出两个单词之间的语义相似性。

本项目通过使用 Glove 词向量将单词映射为 embedding matrix，有效解决语义难题。

6.3.2 负采样构建训练集与测试集

根据数据集的结构，我们可以将已经引用的文献作为查询文献的正样本，但为了平衡正负样本的大小，我通过随机抽取与正样本数量相等的数据集中的其他文献作为负样本。正样本标签为 1，负样本标签为 0。将得到的样本随机按比例划分为训练集和测试集。训练集和测试集的大小如表 5 所示，

	Train	Test
Sample Size	96000	32352

表 5

6.3.3 损失函数

此任务本质上是一个二分类问题，故而选择使用 BCEWithLogitsLoss，这种损失函数将 Sigmoid 层和 BCELoss 合并为一个类别。BCEWithLogitsLoss 比使用普通的 Sigmoid 和 BCELoss 的组合在数值上更稳定，因为通过将操作合并到一层中，利用了 log-sum-exp 技巧来实现数值稳定性。具体公式如下，

$$loss(z, y) = mean(l_0, \dots, l_{N-1}\}$$

$$l_n = -\big(y_n * \log\big(\delta(z_n)\big) + (1 - y_n) * \log\big(1 - \delta(z_n)\big)\big)$$

用 N 表示样本数量， z_n表示预测第 n 个样本为正例的得分， y_n表示第 n 个样本的标签， δ 表示 sigmoid 函数。

6.3.4 Baseline

本项目使用两个 Baseline 作为对比。

Baseline1：直接使用 LDA 主题模型结果，如果主题相同则判定标签为 1，若主题不同，则判断标签为 0。

Baseline2：使用类似于 SQT-CNN 的神经网络，但不加上学科类别、文章类别、主题这些额外特征。

---

[13] Pennington J , Socher R , Manning C . Glove: Global Vectors for Word Representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014.

6.3.5 实验结果

实验结果如表 6 所示

	Test Accuracy
Baseline 1	58.1%
Baseline 2	80.3%
SQT-CNN	86.6%

表 6

由上述结果可以得出，SQT-CNN 神经网络可以有效判断当前查询文献与候选引用文献的是否匹配，故而可以有效解决引用文献查找困难的问题。

七. 分析与改进

本项目通过 LDA 主题模型和文本分类模型，有效分析了 2000 年到 2020 年近二十年气候变化文献的热点问题，并提出 SQT-CNN 神经网络，在查找相似引用文献问题上达到了 86.6%的高准确率。

从词云图和 LDA 主题模型我们可以看出，当今大气环境污染问题日益加剧。无论是全球变暖、水资源污染，还是土壤恶化都会影响全世界人民的经济活动。实际上，这些气候环境问题的主要源头还是人类活动，因此从经济学角度来看，是一种负外部性的体现。外部性指的是私人收益与社会收益、私人成本与社会成本不一-致的经济现象。当个体的经济决策经过非市场的价格手段直接地或不可避免地影响了其他个体的生产函数或成本函数，并成为后者自己所不能控制的变量时，那么对前者来说就有了外部性[14]。我们可以通过污染外部性的内部化来有效治理气候环境问题。所谓污染外部性的内部化.就是使生产者或消费者产生的外部性效应进入它们的生产和消费决策，由它们自己承担，从而解决污染外部性问题。

污染外部性内部化的主要两种途径:命令与控制政策和以市场为基础的政策。命令与控制政策是指政府以规章制度对环境污染外部性进行直接干预，包括命令和控制。其中，最有代表性的就是实施排污标准控制，即由政府管制部门制定并依法强制实施的某一污染源特定污染物排放的限度。但此方法一方面对社会环境变化缺乏适应性，存在政策时滞；另一方面命令控制方法很难考虑企业间的技术差异和污染物处理的边际费用差异。以市场为基础的政策是指从影响成本和收益人手，利用价格机制，采取鼓励性或限制性措施促使污染者减少甚至消除污染。从而使污染外部性内部化，以便最终解决环境污染负外部性问题的一种手段。对于污染外部性的内部化来说，这是更符合经济学原理的一种手段。

以市场为基础的政策主要包括矫正税和可交易污染许可证。本质上矫正税规定了污染权的价格。正如市场把物品分配给那些 对物品评价最高的买者一样，矫正税把污染权分配给那些减少污染成本最高的工厂，并且激励工厂去开发更为环保的技术，因此既可以增加政府收入，又减少了大气环境污染，提高了经济效率。可交易污染许可证类似于一个市场，这种市场为供求力量所支配。看不见的手将保证这种新市场有效地配置污染权。

[14] Marshall A . The Principles of Economics[J]. History of Economic Thought Books, 1992.

SQT-CNN 神经网络可以为有效解决引用文献查找困难的问题，帮助科研人员减轻一些对文献归纳总结的负担。

在今后的进一步研究中，可以增大样本量一方面使得主题更具有代表性，另一方面可以有效提高 SQT-CNN 神经网络的准确性。

八. 参考文献

[1] Stern S N . What is the Economics of Climate Change?[J]. World Economics, 2006, 7(2):1-10.

[2]施瓦茨郭晗聃. 气候经济学:影响全球 80%经济活动的决定性因素[M]. 气象出版社, 2012.

[3] William D. Nordhaus. Estimates of the Social Cost of Carbon: Background and Results from the RICE-2011 Model[J]. social science electronic publishing, 2011.

[4] Nordhaus, William. Evolution of modeling of the economics of global warming: changes in the DICE model, 1992–2017[J]. Climatic Change, 2018.

[5] Kardol P , Cregger M A , Campany C E , et al. Soil ecosystem functioning under climate change: plant species and community effects[J]. Ecology, 2010, 91(3):767-781.

[6] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

[7] https://zhuanlan.zhihu.com/p/31470216.

[8] Yang J , Yang J Y . Why can LDA be performed in PCA transformed space?[J]. Pattern Recognition, 2003, 36(2):563-566.

[9] http://blog.csdn.net/zhongkelee/article/details/44064401.

[10] by G Babu, M Murty. A near optimal initial seed value selection in kmeans algorithm using a genetic algorithm[C]// Pattern Recognition Letters. 1993.

[11] Kim Y . Convolutional Neural Networks for Sentence Classification[J]. Eprint Arxiv, 2014.

[12] Severyn A , Moschitti A . Learning to Rank Short Text Pairs with Convolutional Deep Neural Networks[C]. the 38th International ACM SIGIR Conference. ACM, 2015.

[13] Pennington J , Socher R , Manning C . Glove: Global Vectors for Word Representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014.

[14] Marshall A . The Principles of Economics[J]. History of Economic Thought Books, 1992.