# **题目：基于大语言模型给政策文件自动打分类标签**

# 成员：

侯国通（23秋大数据工程）

薛颉（231017000135）

庞恺（221017000229）

张琳（221017000222）

# 研究背景

随着管理型政府向服务型政府转变，各级政府出台的政策性文件也越来越多，越来越有针对性。由于政策文件众多，这些政策文件通常都包含大量的信息，涉及不同的主题和领域，企业及个人用户往往无法及时、准确从大量政策中获取自己想要的政策信息，政策无法发挥最大的效果。通过给政策文件打标签，标识文件中的特定信息或主题，不但有助于提高政策解读的效率、政策的管理和检索效率，还可以根据用户画像将对应政策文件精准推送给有需要的企业或个人，提升政策惠及面。

随着文件的增长，文件标签的管理变得越来越重要，而手动创建应用标签是一项繁琐且易出错的任务。为了提高效率和准确性我们可以使用标签自动化解决方案，通过大模型给政策自动打标签，不仅可以提高处理速度和准确性，还可以保证标签的一致性，降低成本，具有很高的应用价值。

本方案将介绍一个基于机器学习和自然语言处理的标签自动化解决方案，通过文本摘要和关键词等信息来自动生成文件标签。

# 基本思路：

1、数据准备及预处理，到政府网站爬取各类政策文件，并去除文件头等无关信息，对数据进行预处理；

2、梳理政策文件标签，形成标签库，并提取一定的特征；

3、模型训练，使用机器学习算法，对特征进行训练，以学习政策文件的主题和分类信息；

4、标签自动化，利用训练好的模型，对新的政策文件进行自动打标签；

5、评估和优化：对自动打标签的结果进行评估，包括准确率、召回率、F1值等指标。根据评估结果，对模型进行优化和调整，提高标签自动化的准确性和效率。

# **初步分工**

## 1 文本收集

工作内容：下载文件，集中下载10个品类的文件（10X10份），

负责人：薛颉 张琳

## 2 设计标签库

工作内容：

负责人：侯国通

|  |  |
| --- | --- |
| 标签名称 | 关键字 |
| 科技 | 传感器技术、智能家居、智能城市、大数据、云服务、云存储、云计算平台、分布式账本、加密货币、智能合约、基因编辑、细胞治疗、生物制药、虚拟现实、增强现实、自动驾驶、5G技术、机器学习、深度学习、自然语言处理 |
| 民生 | 教育公平、教育质量、教育改革、医疗保障、医疗服务、医疗改革、社会保险、社会福利、社会救助、就业创业、住房政策、住房保障、房地产市场、环境保护、交通设施、交通管理、公共交通、扶贫脱贫、就业率、失业率、就业政策、 |
| 经济 | 电子商务、移动支付、数字广告、数字创意产业、经济增长率、GDP增长、人均收入增长、物价指数、消费者价格指数、生产者价格指数、失业率、就业率、平均工资、出口贸易、贸易顺差、贸易逆差、利率、货币供应量、货币政策工具 |
| 其他 |  |

## 3 编码实现

工作内容：负责方案的模型搭建及训练，编写数据自动化处理统计等

负责人：庞恺

# **训练步骤**

1.数据整理：100份数据清洗部分不能解析后生成83份，其中66份作为训练集，17份作为测试集覆盖4大类别

2.数据扩增：基于样本过少，采用500字断文策略，共得到776份训练集和124份测试数据

3.重点词统计：对训练集得所有文章统计频率并取前5500个字作为目标

4.下载搜狐sgns.wiki.char.bz2词向量预训练库(https://pan.baidu.com/s/1ZBVVD4mUSUuXOxlZ3V71ZA),获取步骤4的目标词对应向量，避免训练样本少，自己生成得词向量不精确

5.设计网络模型，采用单层Transformer网络层作为预测网络，损失函数是交叉熵CrossEntropyLoss, 优化器是Adam，基于小样本学习，避免模型过拟合，dropout设置0.1

6.开始训练,训练超参如下：

EMBEDDING\_SIZE = 300

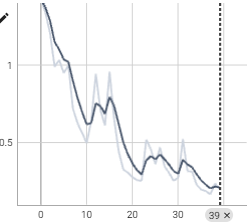
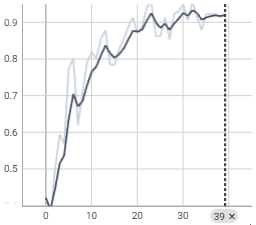
TRAIN\_BATCH\_SIZE = 32

TEST\_BATCH\_SIZE = 16

LEARNING\_RATE = 0.001

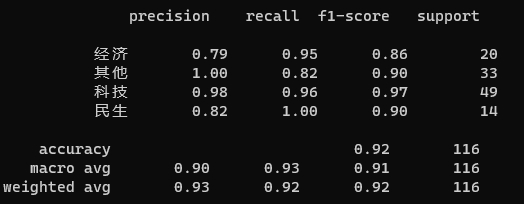
EPOCHS = 40

7.训练过程如下

损失值 准确率

8.效果



# 结论：

通过本次课题的研究，我们成功地利用机器学习和自然语言处理技术实现了文件标签自动化。通过数据预处理、模型训练、标签自动化、评估优化等步骤，我们构建了一个高效的标签自动化系统，可以自动对大量的政策文件进行分类和标签，该系统的性能表现良好，准确率、召回率和F1值等指标均达到了较高的水平。同时，该系统还具有可扩展性和灵活性，可以根据实际需求进行优化和调整。

通过本次研究，我们了解了机器学习和自然语言处理技术在标签自动化方面的应用，积累了相关的实践经验和理论知识。