**一、文本挖掘问题**

**数据集：** 2011-2020十年间网络论坛中与“公知”相关的讨论帖（以“公知”为关键词搜索），逐年下载整合，搜索源为慧科新闻搜索研究数据库，数据放在 “论坛.zip” 里面。

**问题：**

1.数据库下载后，文本中有较多诸如版权页、无效字符等冗余信息，需要剔除

--技术路径：文本过滤、去停用词、去重

2.文本数据量较大，需探索网络论坛10年间关于“公知”讨论的聚焦点变化

--技术路径：分词；使用TF-IDF算法选取重要词汇，并分类；使用LDA和DTM模型进行主题分析

**二、Jigsaw 多语言有毒评论分析(可用深度学习相关模型)**

对话 AI 团队是由 Jigsaw 和谷歌创立的一项研究计划，旨在构建技术来保护对话中的声音。一个主要关注的领域是采用机器学习和深度学习的方法，识别在线对话中的有毒评论，其中“有毒”被定义为任何粗鲁、不尊重或可能使某人离开讨论的事情。如果可以确定这些有害贡献，我们就可以拥有一个更安全、更具协作性的互联网。

**数据描述：**

Jigsaw的APIPerspective以越来越多的语言为毒性模型和其它模型提供服务。该领域一个新的探索是多语言转化能力。这次数据与以往数据的最大区别是，需要使用纯英语训练数据构建多语言模型。此次提供的训练数据是英文数据，测试数据是几种不同语言的维基百科讨论页评论。

**文件包括：**

1. 训练集：

jigsaw-toxic-comment-train.csv维基百科讨论页编辑的英文评论

jigsaw-unintended-bias-train.csv具有一系列附加标签的civilcomments数据集的扩展版

1. 验证集：

validation.csv来自不同非英语语言的维基百科讨论页的评论

1. 测试集

test.csv来自不同非英语语言的维基百科讨论页的评论

test\_lables.csv测试数据的真实标签

数据下载链接以及更详细的介绍见：https://www.kaggle.com/competitions/jigsaw-multilingual-toxic-comment-classification/data

**问题：**

1. 数据预处理和清洗

加载数据：从给定的 Kaggle 链接获取数据集，包括训练集、验证集和测试集。

文本清洗：对英文评论进行清洗，包括去除无关字符、标点符号、数字，以及文本规范化（如小写化）。

多语言处理：由于测试数据包含多种语言，需要确保模型能够处理不同语言的文本。

2. 构建模型并获取数据的 Embedding

选择模型：可以选用 LSTM, CNN, Transformer, Bert 等模型。

训练模型：使用训练数据集来训练模型。这个步骤可能需要调整模型的参数和结构，以达到最优性能。

生成 Embedding：用训练好的模型处理数据，得到每条评论的高维表示（embedding）。

3. 对测试集进行有毒性预测

应用模型：将训练好的模型应用于验证集和测试集，以预测评论的有毒性。

评分系统：将预测结果标记为有毒（1.0）或良性/无毒（0.0）。

评估模型性能：使用真实标签（test\_labels.csv）评估模型预测的准确性。