2022

基于文本分类的诈骗案情分析方法研究

指导老师：孙承杰 报告人：徐浩铭 杜佳兴 樊宇宇 王雅斌

目录

CONTENTS

1

立项背景

疫情下的诈骗案频发

近年来，随着科学技术的快速发展，加之疫情下网络占据越来越重要的地位，诈骗给大家带来了极大的威胁，我们需要对反诈进行新的治理。

AI的发展带来新的挑战

机器学习的发展为电信诈骗带来了新的转机和挑战。网络上涌现大量高级的骗术很多都是基于AI的。

反诈人员也需技术支持

这段时间，国家反诈中心app很火，这就是机器学习大数据带来的新的技术红利，我们还需要更多这方面的研究，需要多部门联合治理。

吉林大学刘源

01

运用聚类的数据挖掘系统，通过Kohonen聚类算法分群后,分析出具有欺诈可能的群集,对群集中的不正常用户采取相应的控制措施,达到以减少运营商的收入流失的目的。

人大Yibo Wang

02

研究索赔中的文本信息来分析保险欺诈行为,提出一种新的深度学习方法

吉大李昊泉

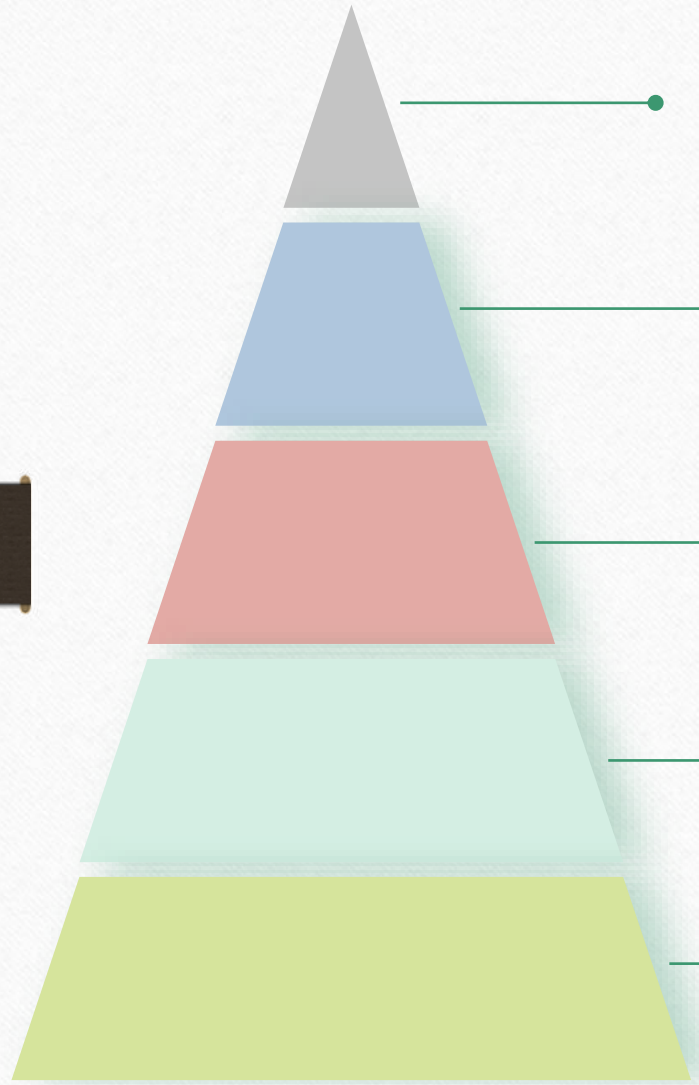
03

在权威网站爬取大量判决文书,通过对判决文书的学习, 将训练出的模型以文本分类的方式, 用于协助办案人员进行罪名预测。

- 好像做诈骗内容分类的还比较少，但是其实对于每天接受大量信息的警察，一个好的判别诈骗内容的工具还是很重要
- 想法：做出一个比较可靠的分析诈骗案情的文本分类器

2

研究内容



语料库构建
- 公安局提供

数据预处理
- 分词处理
- 去噪
- 向量化

模型
- 选取多种模型作为文本分类器

优化
- 提高精确度和召回率

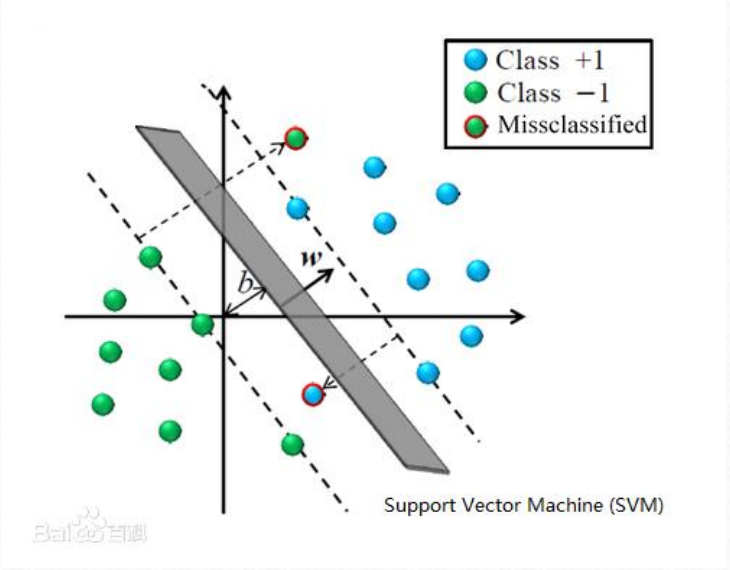
可视化和交互
- 图表词云等
- 交互式机器人
- 小程序公众号

分词处理	目前考虑用jieba或hanlp完成文本的分词处理
去噪	通过停用词表去除中文停用词，通过正则表达式去除数字
向量化	词袋模型如CountVectorizer或者分布式表示如Word2vec等

综合选取合适的模型

核心！！！

机器学习：由于我们是初学，我们将首先采取经典的机器学习算法，如SVM,贝叶斯，KNN等，以达到学习和练手的目的



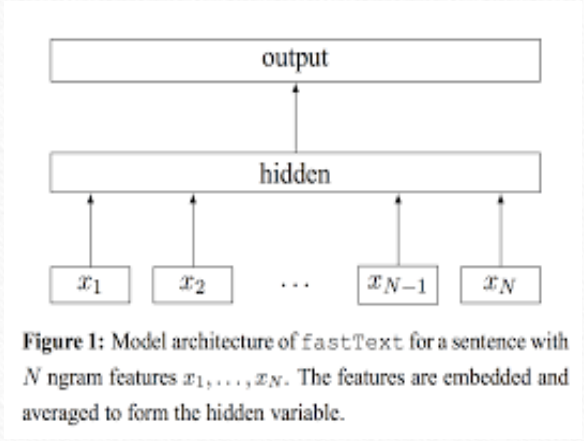
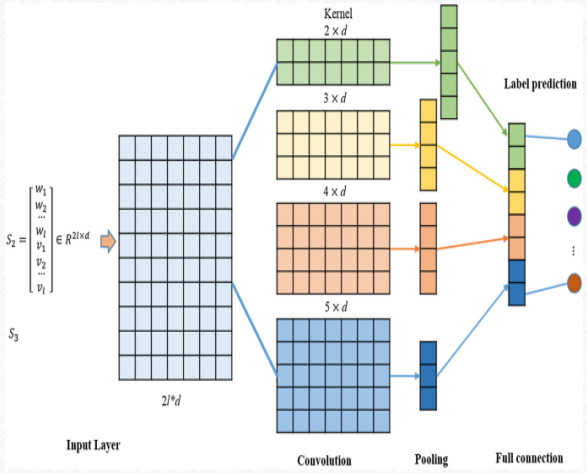


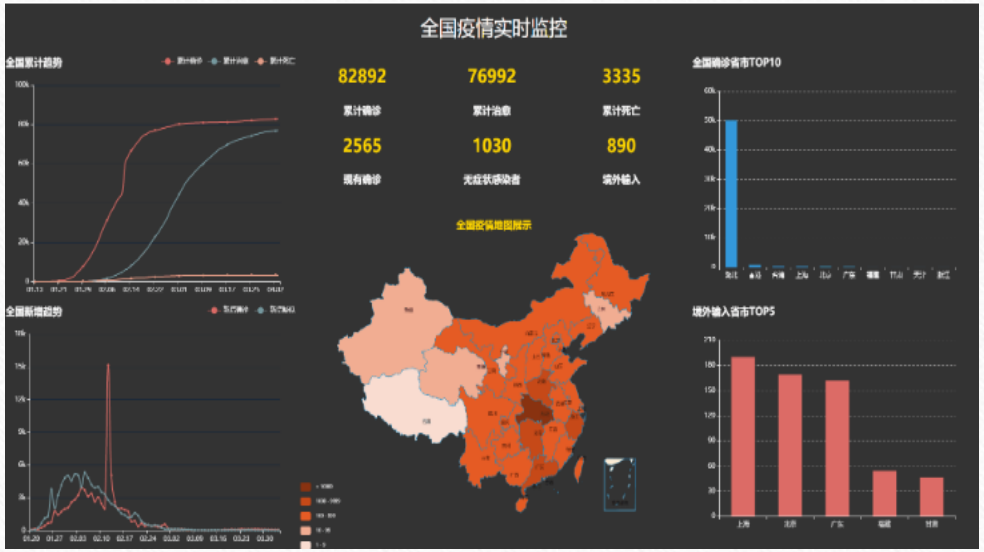
Figure 1: Model architecture of `fastText` for a sentence with N ngram features $x_1, \ldots, x_N$. The features are embedded and averaged to form the hidden variable.



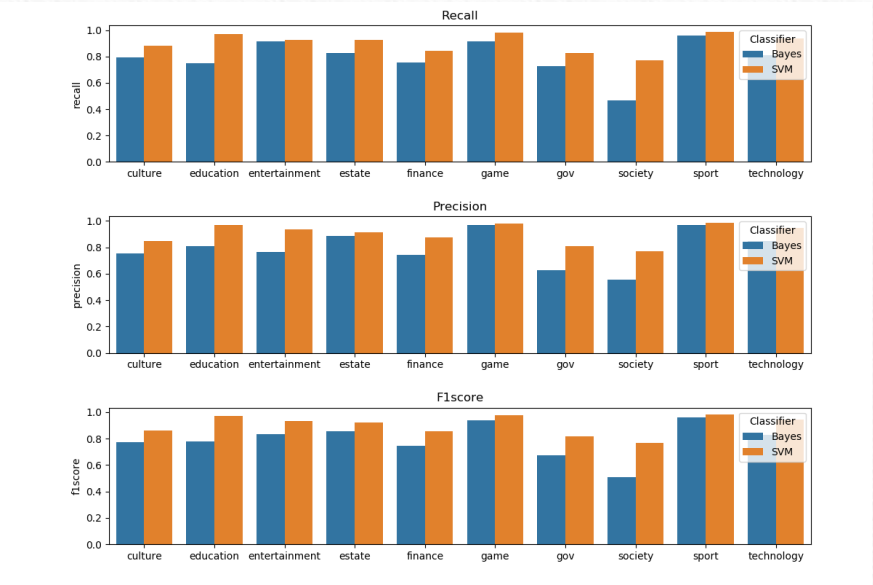
深度学习：采取更加复杂有效的模型，综合选取合适的模型如FastText, TextCNN, BERT并且最后达到理想的精确率和召回率，实现项目的有效性

第二部分　优化和可视化

- 结合k折交叉验证和网格搜索进行调参优化
- 完成较好的可视化，做好诈骗案情分析
  - 网页框架：考虑flask，同时最好将分类机器人嵌入，部署上服务器
  - 绘图模块：matplotlib
  - 或者：做成小程序，嵌入公众号



示例网页



示例性能对比图


3
目标和经费

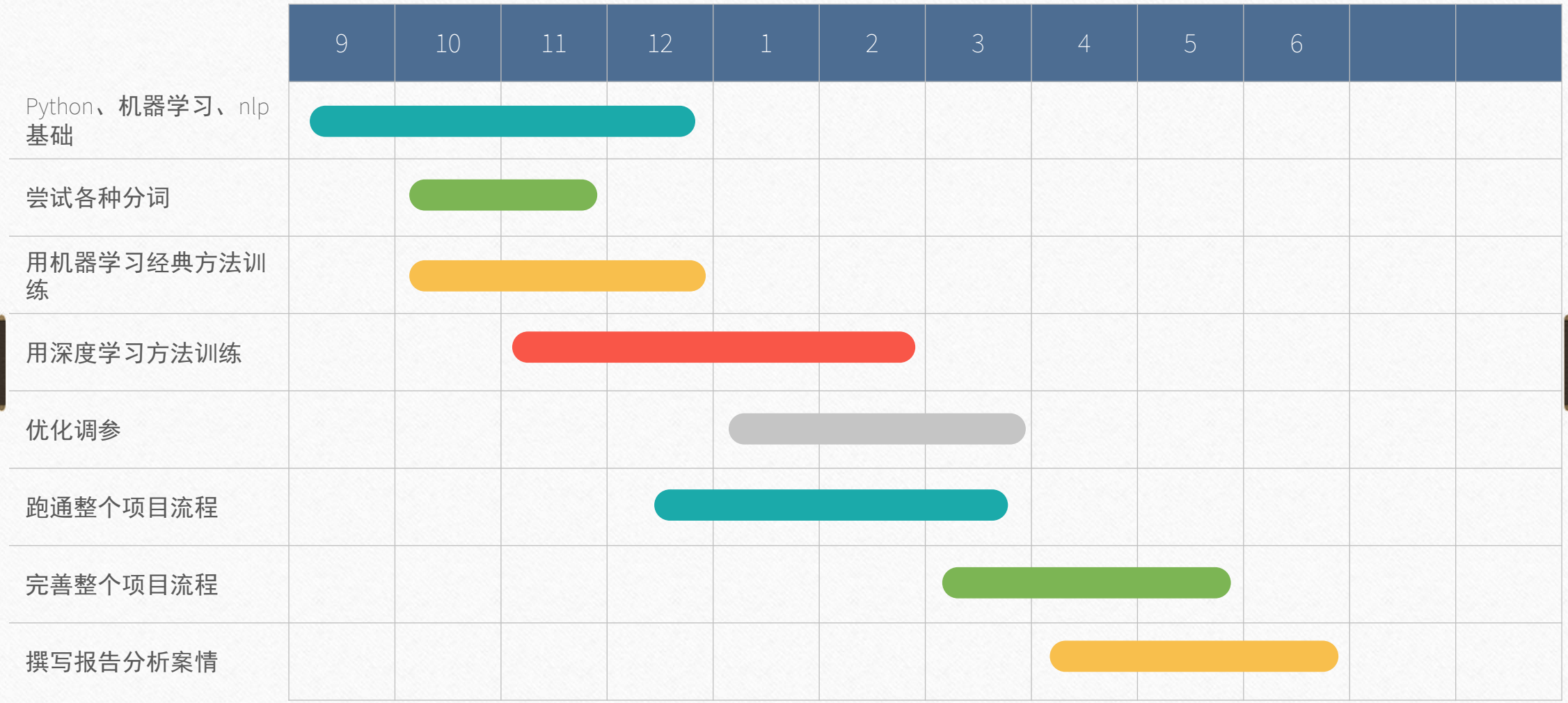
中期目标

- 完成前序知识的学习和准备
- 良好地完成数据预处理
- 初步简单实现项目完整的流程

结题目标

- 继续深入学习项目相关内容
- 优化模型，提高准确率
- 完善项目的各个流程，撰写好一篇报告

第三部分 计划表


9
10
11
12
1
2
3
4
5
6
Python、机器学习、nlp基础
尝试各种分词
用机器学习经典方法训练
用深度学习方法训练
优化调参
跑通整个项目流程
完善整个项目流程
撰写报告分析案情

预算	价格
书籍费	200
资料打印	50
服务器租赁	400

总计：650

2022

谢谢聆听！