母语识别

云南大学信息学院
计算机工程综合实践
Tomasulo Joseph John
2020 年 07 月 08 日

任务

定义：母语识别是确定文本作者的母语的任务

主要作用：创建针对特定语言背景学习者的教材

本项目具体目标：开发一个可以区分以汉语为母语的人写的论文与以英语为母语的人写的论文

背景

Blanchard, et al. (2013)

- Tetrault, et al. (2013)
- Malmasi, et al. (2017)

数据获取

- **arXiv 是一个科学文献预印本的数据库**
  - 1991 创立
  - 150 万论文
  - s3cmd （$27）
  - 1200/2807 tarball
    - 几百个 LaTeX 文件
- **arxiv_archiv**
  - arXiv 的元数据数据集（作者，题目，摘要，领域 ...）

数据挖掘

问题： LaTeX 文件没有标准形式，怎么获取内容

- 找出内容开始、结束的 LaTeX 标签
  - /introduction 或 /end{abstract}

- **问题：没有作者母语的信息**
  - 用作者附属大学的国家的母语作为标签
  - 按国家划分大学名单
  - 用 Aho-Corsick 基于索引的字符串搜索算法

数据形式

来源	类别	内容
arxiv_archive (zenodo)	标题	Microscopic explanation for...
	筛选	Charged dilatonic black hole...
	arxiv_id	1812.11765
	作者（列表）	Yong Chen, Haitang Li...
	主要领域	hep-th
	创建日期	2018-12-31
arxiv.org (s3cmd)	内容	In the past decades, black...
	作者附属的大学	Zhejiang University of Tech.
	标签	Non-native

模型

- **向量化**
  - Count Vectorizer
  - TF-IDF Transformer

- **线性支持向量分类器**

sklearn!

模型结果

Num. samples: train 929, test 233

Training Time 9.59s

Linear SVC Score – 90.12%

Model size: 141MB

混淆矩阵	非母语作者	母语是英语
非母语	102	12
母语是英语	11	108

网页

- Python 的 Flask 模块
  - 上传 .txt 文件
  - 粘贴内容

- 突出显示特征并表示其权重

bootstrap！

结论

虽然提供的工具有很多改进的途径，基本的任务，就是区分母语和非母语作者的论文算成功

提醒：这个工具不是帮您找到语法的问题。大部分强烈表示非母语的作者的特征没有语法错误。反而这工具只能帮你看那些语句是非母语作者常用的，那些是母语是英语的作者常用的。所以你可以理解为一个不良书写风格检测器

Thanks for listening!