任务介绍

文本分类是为一个句子或文档分配合适类别的任务。此次任务是根据新闻的的内容来分类新闻的主题。采用的数据集是 AG's News Topic Classification Dataset 。

数据集

- 类别： Word, Sports, Business, Sci/Tech
- 训练集： train_texts.txt每行对应一则新闻文本，train_labels.txt每行则是对应新闻的主题； 每个类别30000个样本，共计120000个样本。
- 测试集： test_texts.txt每行对应一则新闻文本，test_labels.txt每行则是对应新闻的主题； 每个类别1900个样本，共计7600个样本。

评价指标

测试集上的错误率

$$err = \frac{wrong\_nums}{all\_nums} * 100\%$$

建议方法

统计学习方法

使用Bag-of-words and its TFIDF 或 Bag-of-ngrams and its TFIDF 提取特征，再使用 Multinomial Logistic Regression 或 Naive Bayesian Classifier 学习这些特征。

深度学习方法（了解）

参阅NLP-progress上列出的方法。

提交内容

- 实现代码
- 实验报告，应至少包含以下内容：
  - 阐述问题
  - 介绍方法
  - 汇报结果
  - 思考改进