past_tense_detect

查找中文和英文的句子中的过去式个数

目标

- 找出英文文本过去式的单词
- 找出中文文本表示过去的词句
- 进行统计学分析（提取特征）
- 获取英语母语者的数据集

实现方法

方法如下：

英文文本过去式单词分析

使用Python的nltk库进行词性分析，按照宾夕法尼亚大学tag词性对照表来进行分类，可以获得标注出单词的词性，包括时态，**经测试，可以很好的查找出过去式的单词**

中文文本词性分析

采用中文分词库jieba来分词，可以获得每个词的词性。**由于中文没有语法一说，以“了”来作为过去式的标志。**

统计学分析

后续特征提取即可。

英语母语者数据集

现在有一些开源的数据集是这方面的，主要是用于作文自动批改和机器阅读理解的测试集和训练集，或许有帮助。

使用方法

中英文分词有细小的差别。

中文分词

安装*结巴*中文分词模块，将代码中中文替换成想查找的中文语句。

英文分词

安装*NLTK*模块，将代码中的英文句子换成想查找的英文语句即可。

参考文献

- https://github.com/fxsjy/jieba
- https://github.com/nltk/nltk_data

备注：

Author: Li Yunzhe

Contact: liyunzhe@whu.edu.cn