

浙江工业大学



文本分析与挖掘

上机实验

计算机科学与技术学院

实验一、文本预处理和基本表示

Part 1. 文本预处理

一、实验目的

1. 熟悉英文常用预处理方法;
2. 熟悉中文分词;
3. 了解预处理的重要性;

二、实验内容

1. 英文预处理（可参考电子书 chapter3。）

- a. 编写英文预处理函数 `EngPreprocess()`, 对输入的一个英文段落实现以下功能: 分词、词干提取、词性还原、去停用词。以上功能可以直接调用 `nltk`、`spacy` 相关方法实现;
- b. 对比 `text.split()` 与分词结果;
- c. 对词干提取、词性还原、去停用词分别进行测试（用多个不同样例），观察结果并讨论准确性。
- d. （选作）对以上函数功能进行适当完善，比如去 HTML 标签、大小写转换、拼写错误纠正等，并进行功能测试。

2. 中文分词

- a. 编写函数 `ChTokenize()`, 基于 `jieba` 实现中文分词。
- b. 对不同参数设置进行测试，给出对应结果并讨论每个模式的差别。

3. 分句

- a. 编写函数 `Doc2Sent()`, 实现对英文、中文文档进行分句。可以调用相关函数或自己实现。
- b. 对以上功能用中、英文分别进行测试并讨论结果。