

InfoRetrieval

3120102087 庞罕天

pipeline

1. 预处理数据
 1. 分词（使用spacy，去除非字符词项，还原为词根）
 2. 构建倒排索引
 3. 预先计算tf, idf（实际是将wf, idf以及norm计算保存）
2. 检索
 1. 将检索文本分词，同样使用spacy，并且还原词根
 2. 筛选词项，如果词项是字符，并且在词典中，并且idf大于0.4（speed up），则保留，否则去除，如果没有词项保留，则不设置阈值限制
 3. 对每个词项查找同义词（nlTK wordnet），将每个同义词的集合对应的文档合取，作为该词项最终的文档集合
 4. 对每个词项对应的文档集进行合并，同时要统计每个文档出现的次数（用于后续的speed up）
 5. 按文档出现的次数group成多个文档集，然后按照出现次数从大到小对每个文档集进行6, 7操作
 6. 对每个文档和检索文本计算相似度并排序
 7. 按相似度从大到小添加到最终结果中
 8. 如果文档数量足够，则不继续操作下一个文档集，否则取下一个文档集重复6, 7操作

依赖

- spacy
- nlTK
- beautifulsoup4

使用

1. 出现>Please input a query:"后输入检索短语
2. 出现>Please input K:"后输入显示条目数量
3. 回车显示下一条检索结果，输入q再回车则跳出显示