哈爾濱工業大學

自然语言处理

题　　目　作业二

专　　业　人工智能

学　　号　1190201303

班　　级　1903601

学　　生　王艺丹

指 导 教 师　杨沐昀

计算机科学与技术学院

2021 年 10 月

一、 请找到一个英文的 tokenization 工具，分析其代码中如何处理这些问题(歧义)?

NLTK 分词

1. word_tokenize
   利用英文自身的特点，利用空格和标点符号等进行分词
2. TweetTokenizer
   按空格进行分词，同时针对推文一些特性，去除@用户名，保留表情等一些特殊符号
3. MWETokenizer
   再分词基础上对已经先保留的一些短语，或者组合，进行重组
4. RegexpTokenizer
   使用正则表达式进行分词
5. StanfordTokenizer
   按空格进行分词，对于$4.28 之类的，将符号与数字分开

对于歧义，可以使用正则表达式（regular expression）进行识别和特殊处理。为了使后续处理能识别同个单词的不同变体，一般要对分词结果提取（stemming），即提取出单词的基本形式。比如 do、does、done 这 3 个词统一转化成为词干 do。提取词干可以利用规则处理，比如著名的 Porter Stemmer 就是采用一系列复杂的规则提取词干。可以利用停用词和惯用短语表达进行歧义的处理

二、 从最长匹配到最大频率分词，体现了什么工程实践中的普遍规律?

经验主义

2.1 站在工程技术高度，分词/tokenization 于 NLP 的意义是什么(提示：方法论角度，此题可在第 1 讲课后再提交)

从方法论的角度来看，NLP 主要是通过对语言进行建模来让及机器解决一些自然语言问题，而分词则是语言模型建立的必要前提，是自然语言与机器之间的桥梁。分词是其他信息处理的基础。

2.2 可否证明最长匹配分词的合理性?(要求超越直觉说明和个例说明的层次，要更客观可信;此题可在第 2 讲课后再提交)

个人认为可以看作是有向图的路径匹配，如果存在路径存在 a->b->c 的有向路，则 abc 一定是一个有意义的词，但 ab 与 bc 不一定是一个有意义的词。即最长匹配分词更能够保证分词的合理性，并且在保证分词结果有意义的前提下分词数量更少，全局最优。