Atitit nlp 常见功能 总结

目录

[1. NLP的底层任务由易到难大致可以分为词法分析、句法分析和语义分析 2](#_Toc20473)

[1.1. 分词是词法分析（还包括词性标注和命名实体识别）中最基本的任务 2](#_Toc8155)

[2. NLP的最基本任务和知识，比如tagging, 各种parsing，coreference, semantic role labeling等等等等 2](#_Toc28902)

[2.1. 中文分词（Character-Based Generative Model） 2](#_Toc8931)

[2.2. NER识别其中的命名实体，最后使用关系识别搜索相近实体间的可能的关系。 2](#_Toc10237)

[2.3. 关系抽取 2](#_Toc26670)

[2.4. 情感分析 3](#_Toc5903)

[2.5. 文本分类（Naive Bayes） 3](#_Toc25152)

[2.6. 提取文本关键词（TextRank算法） 3](#_Toc18080)

[2.7. 提取文本摘要（TextRank算法） 3](#_Toc8025)

[2.8. 观点抽取,需要将文本当中有价值的信息提取出来 3](#_Toc15507)

[2.9. 词云 词频计算 3](#_Toc1002)

[3. 其他人物 3](#_Toc20803)

[3.1. 词性标注（TnT 3-gram 隐马） 3](#_Toc15816)

[3.2. tf，idf 3](#_Toc32376)

[3.3. 同义词 反义词 3](#_Toc24026)

[3.4. NLTK词干提取单词词干提取就是从单词中去除词缀并返回词根。 3](#_Toc32463)

[3.5. LDA主题建模。 4](#_Toc13396)

[3.6. 文本分类（Naive Bayes） 4](#_Toc2442)

[3.7. 转换成拼音（Trie树实现的最大匹配） 4](#_Toc11204)

[3.8. 繁体转简体（Trie树实现的最大匹配） 4](#_Toc5621)

[4. 艾提拉常用的功能 4](#_Toc371)

[4.1. 中文分词（Character-Based Generative Model） 4](#_Toc10934)

[4.2. NER识别其中的命名实体，最后使用关系识别搜索相近实体间的可能的关系。 4](#_Toc28140)

[4.3. 情感分析 歌词分析 4](#_Toc8379)

[4.4. 文本分类（Naive Bayes） 歌词分类 5](#_Toc9899)

[4.5. 提取文本关键词（TextRank算法） 5](#_Toc16052)

[4.6. 提取文本摘要（TextRank算法） 5](#_Toc10811)

[4.7. 观点抽取,需要将文本当中有价值的信息提取出来 5](#_Toc31028)

[4.8. 词云 词频计算 5](#_Toc4685)

[4.9. 搜索索引（Lucene/Solr/ES 5](#_Toc17641)

# NLP的底层任务由易到难大致可以分为词法分析、句法分析和语义分析

NLP的底层任务由易到难大致可以分为词法分析、句法分析和语义分析。

## 分词是词法分析（还包括词性标注和命名实体识别）中最基本的任务

可以说既简单又复杂。说简单是因为分词的算法研究已经很成熟了，大部分的准确率都可以达到95%以上

# NLP的最基本任务和知识，比如tagging, 各种parsing，coreference, semantic role labeling等等等等

## 中文分词（[Character-Based Generative Model](http://aclweb.org/anthology//Y/Y09/Y09-2047.pdf" \t "http://www.ppkanshu.com/index.php/post/_blank)）

## NER识别其中的命名实体，最后使用关系识别搜索相近实体间的可能的关系。

## **关系抽取**

## 情感分析

* （现在训练数据主要是买卖东西时的评价，所以对其他的一些可能效果不是很好，待解决）

## 文本分类（Naive Bayes）

## 提取文本关键词（[TextRank](http://acl.ldc.upenn.edu/acl2004/emnlp/pdf/Mihalcea.pdf" \t "http://www.ppkanshu.com/index.php/post/_blank)算法）

## 提取文本摘要（[TextRank](http://acl.ldc.upenn.edu/acl2004/emnlp/pdf/Mihalcea.pdf" \t "http://www.ppkanshu.com/index.php/post/_blank)算法）

## 观点抽取,需要将文本当中有价值的信息提取出来

## 词云 词频计算

* 当时自然语言处理（NLP）的核心是分词（IK分词、HanLP）、搜索索引（Lucene/Solr/ES）、文本分类、主题提取等，当时这些的主流技术栈都是Java系列，且由于C++难用、Python效率低的那个问题，这些组件至今为止还是以Java系为主流，只不过成熟了（虽然普及，但是不火了）

# 其他人物

## 词性标注（[TnT](http://aclweb.org/anthology//A/A00/A00-1031.pdf" \t "http://www.ppkanshu.com/index.php/post/_blank) 3-gram 隐马）

## tf，idf

## 同义词 反义词

## **NLTK词干提取**单词词干提取就是从单词中去除词缀并返回词根。

（比方说 working 的词干是 work。）搜索引擎在索引页面的时候使用这种技术，所以很多人通过同一个单词的不同形式进行搜索，返回的都是相同的，有关这个词干的页面。

词干提取的算法有很多，但最常用的算法是 Porter 提取算法。NLTK 有一个 PorterStemmer 类，使用的就是 Porter 提取算法。

## LDA主题建模。

## 文本分类（Naive Bayes）

## 转换成拼音（Trie树实现的最大匹配）

## 繁体转简体（Trie树实现的最大匹配）

# 艾提拉常用的功能

## 中文分词（[Character-Based Generative Model](http://aclweb.org/anthology//Y/Y09/Y09-2047.pdf" \t "http://www.ppkanshu.com/index.php/post/_blank)）

## NER识别其中的命名实体，最后使用关系识别搜索相近实体间的可能的关系。

## 情感分析 歌词分析

* （现在训练数据主要是买卖东西时的评价，所以对其他的一些可能效果不是很好，待解决）

## 文本分类（Naive Bayes） 歌词分类

## 提取文本关键词（[TextRank](http://acl.ldc.upenn.edu/acl2004/emnlp/pdf/Mihalcea.pdf" \t "http://www.ppkanshu.com/index.php/post/_blank)算法）

## 提取文本摘要（[TextRank](http://acl.ldc.upenn.edu/acl2004/emnlp/pdf/Mihalcea.pdf" \t "http://www.ppkanshu.com/index.php/post/_blank)算法）

## 观点抽取,需要将文本当中有价值的信息提取出来

## 词云 词频计算

## 搜索索引（Lucene/Solr/ES