Atitit 自然语言处理（NLP）的应用 与 搜索引擎

目录

[1.1. 搜索引擎并不是自然语言处理（NLP）的唯一应用。 2](#_Toc2513)

[1.2. NLP的应用 2](#_Toc32715)

[1.3. 社交网站信息流， 2](#_Toc28971)

[1.4. 语音助手，诸如苹果 Siri。 3](#_Toc26701)

[1.5. 垃圾邮件程序，比如 Google 的垃圾邮件过滤程序 3](#_Toc4499)

[1.6. 机器翻译 3](#_Toc22638)

[1.7. Ocr 3](#_Toc12233)

[1.8. automated chat bots, 3](#_Toc2596)

[1.9. article summarizers, 3](#_Toc27961)

[1.10. multi-lingual translation 3](#_Toc29499)

[1.11. opinion identification from data 3](#_Toc616)

[1.12. 文本相似（BM25） 3](#_Toc9334)

[2. Nlp常见功能 3](#_Toc30201)

[2.1. 词频统计 3](#_Toc10012)

[2.2. 分词 4](#_Toc31899)

[2.3. 同义词 反义词 4](#_Toc29593)

[2.4. 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。 4](#_Toc20891)

[2.5. 从文本中理解和提取重要信息 4](#_Toc12878)

[2.6. ，其主要任务包括语音识别和生成 4](#_Toc5373)

[2.7. 、文本分析、 4](#_Toc16332)

[2.8. 情感分析、机器翻译 4](#_Toc25401)

[2.9. 词云、情感分析和LDA主题建模。 4](#_Toc30739)

[2.10. 文本分类（Naive Bayes） 4](#_Toc15383)

[2.11. 转换成拼音（Trie树实现的最大匹配） 4](#_Toc10372)

[2.12. 繁体转简体（Trie树实现的最大匹配） 4](#_Toc14659)

[2.13. 提取文本关键词（TextRank算法） 5](#_Toc18087)

[2.14. 提取文本摘要（TextRank算法） 5](#_Toc23172)

[2.15. Tokenization（分割成句子） 5](#_Toc14848)

[2.16. text classification, 5](#_Toc7365)

[2.17. entity detection, 5](#_Toc14710)

[2.18. machine translation, 5](#_Toc31314)

[2.19. question answering, 5](#_Toc291)

[2.20. concept identification. 5](#_Toc9709)

[3. 现在有许多开源的自然语言处理（NLP）库。比如： 5](#_Toc2956)

[4. Ref 6](#_Toc8108)

## 搜索引擎并不是自然语言处理（NLP）的唯一应用。

有很多公司热衷收集所有这些数据，以便更好地了解他们的用户和用户对产品的热情，并对他们的产品或者服务进行合适的调整。

这些海量数据可以揭示很多现象，打个比方说，巴西人对产品 A 感到满意，而美国人却对产品 B 更感兴趣。通过NLP，这类的信息可以即时获得（即实时结果）。例如，搜索引擎正是一种 NLP，可以在正确的时间给合适的人提供适当的结果。

但是搜索引擎并不是自然语言处理（NLP）的唯一应用。还有更好更加精彩的应

## **NLP的应用**

以下都是自然语言处理（NLP）的一些成功应用：

* 搜索引擎，比如谷歌，雅虎等等。谷歌等搜索引擎会通过NLP了解到你是一个科技发烧友，所以它会返回科技相关的结果。

## 社交网站信息流，

* 比如 Facebook 的信息流。新闻馈送算法通过自然语言处理了解到你的兴趣，并向你展示相关的广告以及消息，而不是一些无关的信息。

## 语音助手，诸如苹果 Siri。

## 垃圾邮件程序，比如 Google 的垃圾邮件过滤程序

* ，这不仅仅是通常会用到的普通的[垃圾邮件过滤](https://likegeeks.com/linux-mail-server/" \l "Secure-Mail-Boxes-From-Spam-UsingSpamAssassin" \t "https://blog.csdn.net/hzp666/article/details/_blank)，现在，垃圾邮件过滤器会对电子邮件的内容进行分析，看看该邮件是否是垃圾邮件。

## 机器翻译

## Ocr

## automated chat bots,

## article summarizers,

## multi-lingual translation

## opinion identification from data

## 文本相似（BM25）

# Nlp常见功能

## **词频统计**

## 分词

正如所料，Mr. 是一个词，也确实被 NLTK 当做一个词。NLTK使用 nltk.tokenize.punkt module 中的 PunktSentenceTokenizer 进行文本分词。这个标记器经过了良好的训练，可以对多种语言进行分词

## 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。

## 从文本中理解和提取重要信息

## ，其主要任务包括语音识别和生成

## 、文本分析、

## 情感分析、机器翻译

## 词云、情感分析和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）

## 提取文本关键词（TextRank算法）

## 提取文本摘要（TextRank算法）

## Tokenization（分割成句子）

NLP 的任务包括

## text classification,

## entity detection,

## machine translation,

## question answering,

## concept identification.

# 现在有许多开源的自然语言处理（NLP）库。比如：

* Natural language toolkit (NLTK)
* Apache OpenNLP

NLTK 也很容易入门，实际上，它将是你用到的最简单的自然语言处理（NLP）库

另外， [jieba](https://github.com/fxsjy/jieba" \t "https://my.oschina.net/letiantian/blog/_blank)分词提供的基于TextRank的关键词提取工具。 [snownlp](https://github.com/isnowfy/snownlp" \t "https://my.oschina.net/letiantian/blog/_blank)也实现了关键词提取和摘要生成。

# Ref

(9+条消息)Python NLTK 自然语言处理入门与例程 - hzp666的博客 - CSDN博客.html