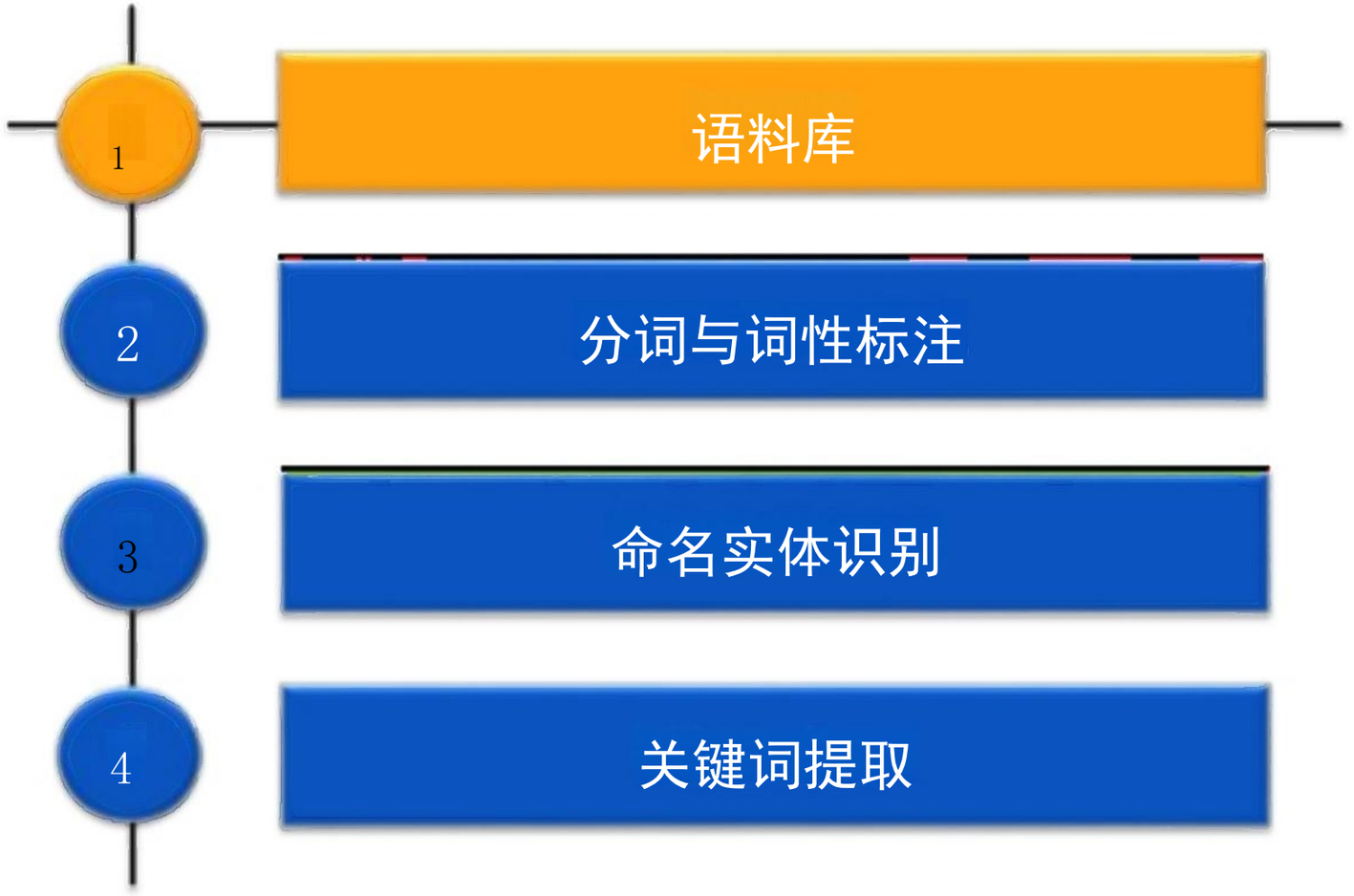
文本基础处理

2023/9/6

目录



语料库概述

语料库是为某一个或多个应用而专门收集的、有一定结构的、有代表性的、可以被计算机程序检索的、具有一定规模的语料集合。

语料库概述

1. 语料库简介

>语料库的实质是经过科学取样和加工的大规模电子文本库。语料库具备以下3个显著的特征。

- 语料库中存放的是真实出现过的语言材料。
- 语料库是以计算机为载体，承载语言知识的基础资源。
- 语料库是对真实语料进行加工、分析和处理的资源。

>语料库不仅仅是原始语料的集合，而且是有结构的并且标注了语法、语义、语音、语用等语言信息的语料集合。

>任何一个信息处理系统都离不开数据和知识库的支持，这点对于使用NLP技术的系统自然也不例外。在NLP 的实际项目中，通常要使用大量的语言数据或者语料。语料作为最基本的资源，尽管在不同的NLP 系统中所起到的作用不同，但是却在不同层面上共同构成了各种NLP 方法赖以实现的基础。

语料库概述

2. 语料库的用途

>语料库的产生起始于语言研究，，后来随着语料库功能的增强，它的用途变得越来越广，以下将从4个方面阐述语料库的几大用途。

(1) 用于语言研究

- 语料库为语言学的研究提供了丰富真实的语言材料，在句法分析、词法分析、语言理论及语言史研究中都起到了强大的作用。如今，人们对语料库内的语料进行了更深层次的加工处理，为语义学、语用学研究、会话分析、言语变体、语音科学及心理学研究等方面提供了大量支持。

语料库概述

(2) 用于编纂工具参考书籍

- 一些对语言教学有重要影响的词典和语法书均是在语料库的基础上编写的。例如，《朗曼当代英语词典》第3版的编写利用了3个大型的语料库，分别是上亿词的BNC语料库、3000万词的朗曼兰开斯特语料库和朗曼学习者语料库。该词典中最常用词及频率、成语、搭配和例句等都是根据这三大语料库统计出来的。

(3) 用于语言教学

- 在语言教学中，语料库可以帮助减少课堂上学习的语言与实际使用的语言之间的差距，发现过去被忽略的语言规律，能够更准确地理解一些词语在实际交际中的意义和用法，发现学习者使用语言时的一些问题。此外，语料库还可以用于语言测试、分析语言错误等用途。

语料库概述

（4）用于NLP

- 语料库按照一定的要求加工处理后可以应用到NLP 的各个层面的研究中。
- 语料库在词层面上进行分词、 词性标注后，可以用于词法分析、拼写检查、全文检索、词频统计、名词短语的辩识和逐词机器翻译等。
- 语料库在句层面上进行句法标注、语义标注后，可以用于语法检查、词义排歧、名词短语辩识的改进、机器翻译等。
- 语料库在语篇层面上进行语用层的处理后，可以用于解决指代问题、时态分析、目的识别、文本摘要和文本生成等。

语料库概述

>语料库包含的语言词汇、语法结构、语义和语用信息为语言学研究和NLP 研究提供了大量的资料来源。

>j 语料库即是时代的产物 1 也是科技进步的成果，让处于大数据时代的人们得以拥有和享受语料库带来的便利。

语料库的产生，既丰富了语言研究中词汇的数量、语法的形态及语句的结构，又让学习和研究语言的方式产生了巨大的变化。各种随时代而兴起的技术也有了更为准确的语言研究基础。

语料库种类与原则

- 语料库的类型主要依据它的研究目的和用途进行划分。

>根据不同的划分标准，语料库可以分为多种类型。

例如，按照语种划分，语料库可以分为单语种语料库和多语种语料库；按照记载媒体不同划分，语料库可以分为单媒体语料库和多媒体语料库；按照地域区别划分，，语料库可以分为国家语料库和国际语料库等。

语料库种类与原则

1. 语料库种类

将语料库以语料库结构进行划分可分为平衡结构语料库与自然随机结构语料库，以语料库用途进行划分可分为通用语料库与专用语料库，以语料选取时间进行划分可分为共时语料库与历时语料库。

(1) 平衡结构语料库与自然随机结构语料库

>平衡结构语料库的着重点是语料的代表性和平衡性，需要预先设计语料库中语料的类型，定义好每种类型语料所占的比例并按这种比例去采集组成语料库。

>例如，历史上第一个机读语料库布朗语料库就是一个平衡语料库的典型代表，它的语料按三层分类，严格设计了每一类语料所占的比例。自然随机结构的语料库则是按照某个原则随机去收集组成语料，如《圣经》语料库、狄更斯著作语料库、英国著名作家作品库、北京大学开发的《人民日报》语料库等。

语料库种类与原则

(2) 通用语料库与专用语料库

>所谓的通用语料库与专用语料库是从不同的用途角度上看问题得来的结果。

>通用语料库不做特殊限定， 而专用语料库的选材可以只限于某一领域，为了某种专门的目的而采集。

>只采集某一特定领域、特定地区、特定时间、特定类型的语料所构成的语料库即为专用语料库，如新闻语料库、科技语料库、中小学语料库、北京口语语料库等。通用领域与专用领域只是一个相对的概念。

语料库种类与原则

(3) 共时语料库与历时语料库

>共时语料库是为了对语言进行共时研究而建立的语料库，即无论所采集语料的时间段有多长，只要研究的是一个时间平面上的元素或元素的关系，则是共时研究。共时研究所建立的语料库就是共时语料库，如中文地区汉语共时语料库（ Linguistic Variation in Chinese Speech Communities,LiVac）， 采用共时性视窗模式，剖析来自中文地区有代表性的定量中文媒体语料，是一个典型的共时语料库。

所谓的历时语料库是为了对语言进行历时研究而建立的语料库，即研究一个历时切面中元素与元素关系的演化。例如，原国家语委建设的国家现代汉语语料库，收录的是1919年至今的现代汉语的代表性语料，是一个典型的历时语料库。根据历时语料库得到的统计结果是依据时间轴的等距离抽样得到的若干频次变化形成的走势图。

语料库种类与原则

2. 语料库的构建原则

>从事语言研究和机器翻译研究的学者逐渐认识到了语料库重要性，国内外很多研究机构都致力于各种语料库的建设。各种语料库的研究正朝着不断扩大库容量、深化加工和不断拓展新的领域等方向继续发展。建设或研究语料库的时候，　一般需要保证语料库具有以下4个特性。

（1）代表性

在一定的抽样框架范围内采集的样本语料尽可能多地反映无限的真实语言现象和特征。

（2）结构性

收集的语料必须是计算机可读的电子文本形式的语料集合。语料集合结构包括语料库中语料记录的代码、元数据项、数据类型、数据宽度、取值范围、完整性约束。

语料库种类与原则

(3) 平衡性

平衡性是指语料库中的语料要考虑不同内容或指标的平衡性，如学科、年代、文体、地域、使用者的年龄、性别、文化背景、阅历、语料的用途(公函、私信、广告)等指标。 一般建立语料库时，需要根据实际情况选取其中的一个或者几个重要的指标作为平衡因子。

(4) 规模性

大规模的语料库对于语言研究特别是对NLP 研究具有不可替代的作用，但随着语料库的增大，垃圾语料带来的统计垃圾问题也越来越严重。而且当语料库达到一定的规模后，语料库的功能不能随之增长。因此在使用时，应根据实际的需要决定语料库的规模。

NLTK库

1.NLTK简介

NLTK(Natural Language Toolkit)是一个用于构建处理自然语言数据的Python应用开源平台，也是基于Python编程语言实现的NLP 库。

NLTK 提供了超过50个素材库和词库资源的接口，涵盖了分词、词性标注、命名实体识别、句法分析等各项NLP 领域的功能。

NLTK支持NLP 和教学研究，它收集的大量公开数据集和文本处理库，可用于文本分类、符号化、提取词根、贴标签、解析及语义推理等。NLTK也是当前最为流行的自然语言编程与开发工具，在进行NLP研究和应用时，利用NLTK 中提供的函数可以大幅度地提高效率。

NLTK库

>NLTK 的部分模块及功能描述如表所示。

模块	功能	描述
nltk.corpus	获取语料库	语料库和词典的标准化切口
nltk.tokenize、nltk.stem	字符串处理	分词、分句和提取主干
nltk.tag	词性标注	HMM、n-gram、backoff
nltk.classify、nltk.cluster	分类、聚类	朴素贝叶斯、决策树、K-means
nltk.chunk	分块	正则表达式、命名实体、n-gram
nltk.metrics	指标评测	准确率、召回率和协议系数
nltk.probability	概率与评估	频率分布

NLTK库

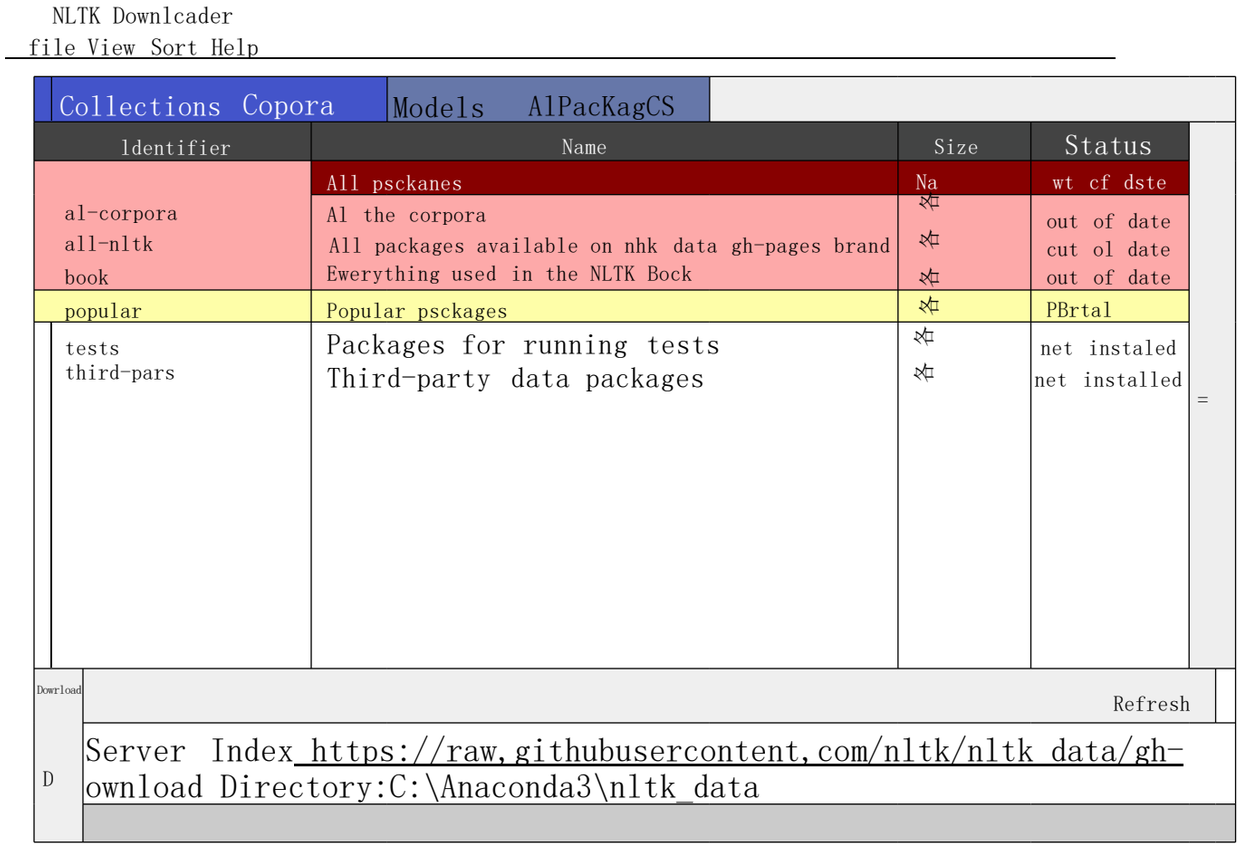
2. 安装步骤

>本书1.3小节已经介绍了Python开发环境的安装和环境变量的配置，以及如何在Anaconda Prompt里创建一个名为NLP的虚拟环境，在本节不再重复介绍。在成功安装Python开发环境和创建NLP虚拟环境的条件下，NLTK 的安装步骤如下。

- 进入NLP 虚拟环境。在Anaconda Prompt命令行激活NLP 虚拟环境。
- 安装NLTK库。在Anaconda Prompt的NLP虚拟环境里安装NLTK库。
- 检查是否存在NLTK 库。
- 下载NLTK 数据包。在成功安装NLTK 库后，打开Spyder,新建一个新文件，编写代码，下载NLTK 数据包。

NLTK库

> 下 载NLTK 数据包时，会显示可供下载的NLTK 数据包的对话框，如图所示。



NLTK库

>由图可看出，首先选择需要下载的包，如all、all-corpora、all-nltk、book、popular、tests、third-partty中的book，然后在“Download Directory”修改下载路径。下载路径可选择为Anaconda3的安装位置，将NLTK数据包放置于Anaconda3的下级目录，如“C:Anaconda3 \nltk_data”(注意需要先在“C:\Anaconda3”目录下新建一个名为“nltk_data”的文件夹)。然后单击“Download”按钮(下载需要一些时间，需耐心等待)。

>下载完数据包以后，还需要进行环境变量的配置，具体步骤为：右键单击计算机图标，在弹出的右键菜单中依次单击“属性”→“高级系统设置”→“高级”→“环境变量”，在“系统变量”里双击“Path”，在输入框中输入下载路径“C:Anaconda3\nltk_data”。

最后检查NLTK数据包是否安装成功，在成功安装NLTK数据包之后，界面会显示NLTK当中Book数据包的示例文本。

语料库的获取

>除了自行构建语料库之外，还有许多已经构建的好的语料库可以直接获取使用。NLTK 中就集成了多个文本语料库，除此之外还有许多网络的在线语料库被共享出来以供人们使用。

>NLTK 中有多个文本语料库，其中包含古腾堡项目(数字图书馆)电子文本档案的一小部分文本、网络和**聊天文本、即时消息聊天会话、路透社语料、就职演说语料、标注文本语料、其他语言语料等。**

>NLTK 中定义了许多基本语料库函数，如表所示。

函数	说明
fileids0	获取语料库中的文件
fileids([categories])	分类对应的语料库中的文件
categories0	语料库中的分类
categories((fileids))	文件对应的语料库中的分类

语料库的获取

>NLTK 中定义了许多基本语料库函数，如表所示。

函数	说明
raw0	语料库的原始内容
raw([fileids=[f1,f2,f3)]	指定文件的原始内容
raw(categories=[c1,c2])	指定分类的原始内容
words0	查找整个语料库中的词汇
words(fileids=[f1,f2,f3])	指定文件中的词汇
words(categories=[c1,c2])	指定分类中的词汇
sents0	指定分类中的句子

语料库的获取

>NLTK 中定义了许多基本语料库函数，如表所示。

函数	说明
sents(fileids=[f1,f2,f3])	指定文件中的句子
sents(categories=[c1,c2])	指定分类中的句子
abspath(fileid)	指定文件在磁盘上的位置
encoding(fileid)	文件编码
open(fileid)	打开指定语料库文件的文件流
root()	到本地安装的语料库根目录的路径
readme()	语料库中的README文件的内容

语料库的获取

>NLTK　包含网络文本、获取网络文本需要先加载NLTK，　然后调用fileids函数获取文本。对于文本可进行以下3个操作。

(1)查找某个文件，统计词数。

(2)索引文本。

(3)获取文本的标识符、词、句。

语料库的构建与应用

1. 构建作品集语料库

- 本节演练如何构建作品集语料库，并在构建完之后对该语料库进行简单的分析。

>下载目前比较火的影视作品集构建作品集料库，完成数据采集和预处理工作，获取保存的文件的列表。

- 构建完成语料库之后，可以利用NLTK 基本函数进行搜索相似词语、指定内容、搭配词语、查询文本词汇频数分布等相应操作。

语料库的构建与应用

2. 古装影视语料库分析

>通过下载的《琅琊榜》语料构建古装影视语料库，具体实现步骤如下。

(1) 读取本地语料。

(2) 查询词频。

(3) 查看《琅琊榜》部分文本。

(4) 统计高频词次数。

(5) 查询词频在指定区间内的词数量。

语料库的构建与应用

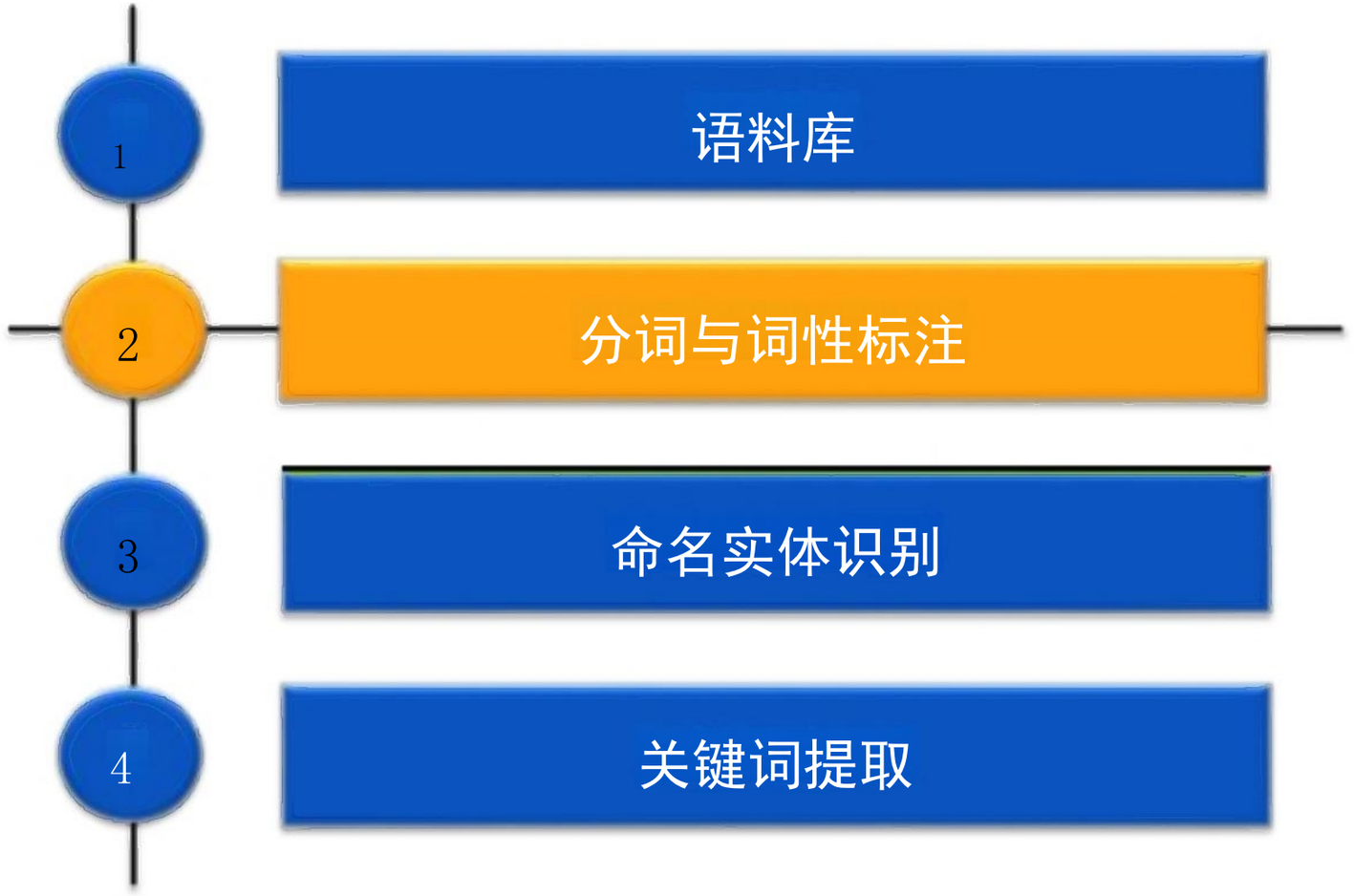
（6）使 用jieba进行分词。NLTK 虽自带了很多统计的功能，但是部分函数只能处理英文语料，对中文语料并不通用。为了使用这些NLTK中的函数，需要对中文进行预处理。首先对中文进行分词，然后将分词的文本封装成NLTK的“text”对象，最后再使用NLTK中的函数进行处理。分词的目的是为NLTK的“text”对象提供封装的语料，这里使用jieba包的Icut函数进行分词(jieba的使用将在第4章介绍)。

(7)查看指定单词上下文。

(8)搜索相似词语。

(9)绘制词汇离散图。

目录



中文分词简介

- 中文分词是指将汉字序列按照一定规范、逐个切分为词序列的过程。在英文中，单词之间以空格为自然分隔符，分词自然地以空格为单位切分，而中文分词则需要依靠一定技术和方法寻找类似英文中空格作用的分隔符。

- 基于规则的分词是中文分词最先使用的方法，常见的方法有正向最大匹配法、逆向最大匹配法等。

  随着统计方法的发展，又提出了一些基于统计的分词模型，常见的分词模型有n元语法模型、隐马尔可夫模型和条件随机场模型。

基于规则的分词

>基于规则或词典的分词方法是一种较为机械的分词方法，其基本思想是将待分词语句中的字符串和词典逐个匹配，找到匹配的字符串则切分，不匹配则减去边缘的某些字符，从头再次匹配，直至匹配完毕或者没有找到词典的字符串而结束。

- 其于规则分词主要有正向是大匹配法(Mavimum Match Methed MM 法)、逆向最大匹配法(Reverse Maximum Match Method,RMM法)和双向最大匹配法(Bi-direction Matching Method,BMM法)这3种方法。

基于规则的分词

1. 正向最大匹配法

- 假设有一个待分词中文文本和一个分词词典，词典中最长的字符串长度为。从左至右切分待分词文本的前个字符，然后查找是否有和词典一致的字符串。若匹配失败， 则删去该字符串的最后一个字符，仅留下前个字符，继续匹配这个字符串，以此类推。如果匹配成功，那么被切分下来的第二个文本成为新的待分词文本，重复以上操作直至匹配完毕。如果一个字符串全部匹配失败，那么逐次删去第一个字符，重复上述操作。

基于规则的分词

>例如，假设待分词文本为“西藏生态屏障建设”，词典为“{”西藏”，”西藏生态", "生态屏障", "建设"}"。由词典得到最长字符串的长度为4，具体分词步骤如下。

(1)切分待分词文本“电影全产业链体系”前4个字符，得到“西藏生态”，在词典中找到与之匹配的字符串，匹配成功。此时，将文本划分为“西藏生态”，“屏障建设”。

(2)将分词后的第二个文本“屏障建设”作为待分词文本。此时词典中找不到与之匹配的字符串，匹配不成功。

(3)删去“屏障建设”的最后一个字符，匹配失败，删去最后两个字符，得到“屏障”，匹配成功，此时，将文本划分为“西藏生态”，“屏障”，“建设”。

(4)将分词后的第三个文本“建设”作为待分词文本。此时词典中找到与之匹配的字符串，匹配成功。

>综上所述，用正向最大匹配法分词，得到的结果是“西藏生态”，“屏障”，“建设”。

基于规则的分词

2. 逆向最大匹配法

DN NAX N 心 + R 权 L 大 不 大 结 从 词 式 未 的 民 个 之 续 中 本 长 目 不 友 和 得 也 ， 列 的 立 焚 中 共

匹配失败，仅留下待分词文本的后个词，继续匹配这个字符串，以此类推。

如果匹配成功，则被切分下来的第一个文本序列成为新的待分词文本，重复以上操作直至匹配完毕。如果一个词序列全部匹配失败， 则逐次删去最后一个字符，重复上述操作。

基于规则的分词

>同样以待分词文本“西藏生态屏障建设”为例说明逆向最大匹配法，具体分词步骤如下。

(1)切分待分词文本“西藏生态屏障建设”后4个字符，得到“屏障建设”，在词典中找不到与之匹配的字符串，匹配不成功。

(2)删去“屏障建设”的第一个字符得到“障建设”，匹配失败，删去“障建设”的第一个字符得到“建设”，匹配成功，将文本划分为“西藏生态屏障”和“建设”。

(3)将分词后的第一个文本“西藏生态屏障”作为待分词文本，与词典匹配不成功，一次删去第一个字符，直到得到“生态屏障”，匹配成功，将文本划分为“西藏”，“生态屏障”和“建设”。

(4)将分词后的第一个文本“西藏”作为待分词文本，与词典匹配成功。

> 综上所述，用逆向最大匹配法分词，得到的结果是“西藏” “生态屏障” “建设”。

基于规则的分词

3. 双向最大匹配法

>双向最大匹配法基本思想是将MM 法和RMM 法的结果进行对比，选取两种方法中切分次数较少的作为切分结果。用正向最大匹配法和逆向最大匹配法对“西藏生态屏障建设”进行分词，结果分别为“西藏生态” “屏障” “建设”和“西藏” “生态屏障” “建设”。选取切分次数最少的结果为“西藏生态” “屏障” “建 设”。

- 研究表明，利用正向最大匹配法和逆向最大匹配法匹配，中文分词大约90%的词句完全重合且正确，有9%左右的句子得到的结果不一样，但其中有一个是正确的。剩下不到1%的句子使用两种方法进行切分都是错误的。因而，双向最大匹配法在中文分词领域中得以广泛运用。

基于统计的分词

基于规则的中文分词常常会遇到歧义问题和未登录词问题。中文歧义问题主要包括交集型切分歧义和组合型切分歧义两大类。交集型切分歧义是指一个字串中间的某个字或词，不管切分到哪一边都能独立成词，如“打折扣”一词，　“打折”和“折扣”可以是两个独立的词语。组合型切分歧义是指一个字串中每个字单独切开或者不切开都能成词，如“将来”一词，可以单独成词，也可以切分为单个字。

未登录词也称为生词，即词典中没有出现的词。未登录词可以分为四大类，第一类是日常生活出现的普通新词汇，尤其是网络热门词语，这类词语更新换代快，且不一定符合现代汉语的语法规定；第二类是专有名词，主要指人名、地名和组织机构名，它还包括时间和数字表达等；第三类是研究领域的专业名词，如化学试剂的名称等；第四类是其他专用名词，如近期新上映的电影、新出版的文学作品等。遇到未登录词时，分词技术往往束手无策。

基于统计的分词

基于统计的分词方法有效解决了中文分词遇到歧义问题和未登录词问题。基于统计的分词方法的基本思想是中文语句中相连的字出现的次数越多，作为词单独使用的次数也越多，语句拆分的可靠性越高，分词的准确率越高。基于统计的分词方法通常需要两个步骤：建立统计语言模型；运用模型划分语句，计算被划分语句的概率，选取最大概率的划分方式进行分词。

常见的基于统计的分词方法包括n元语法模型和隐马尔可夫模型。

基于统计的分词

1. n元语法模型

（1）概念

>n　元语法指文本中连续出现的n个语词。 n元语法模型是基于(n-1)阶马尔可夫链的一种概率语言模型，通过n个语词出现的概率来推断语句的结构。这一模型被广泛应用于概率论、通信理论、计算语言学(如基于统计的自然语言处理)、计算生物学(如序列分析)、数据压缩等领域。 n-gram是一种基于统计语言模型的算法。它的基本思想是将文本里面的内容按照字节进行大小为n的滑动窗口操作，形成了长度是的字节片段序列。

>每一个字节片段称为gram, 对所有gram的出现频度进行统计，并且按照事先设定好的阈值进行过滤，形成关键gram列表，也就是这个文本的向量特征空间，列表中的每一种gram就是一个特征向量维度。

基于统计的分词

(2) 类型

> 当n分别为1、2、3时，又分别称为一元语法(unigram)、 二元语法(bigram) 与三元语法(trigram)。

> 一 元模型(unigram model):把句子分成一个一个的汉字。

> 二元模型(bigram model):把句子从头到尾每两个字组成一个词语。

> 三元语法模型(Trigram model):把句子从头到尾每三个字组成一个词语。

基于统计的分词

(3)中文分词与n元语法模型

>假设语句序列为={小孩，喜欢，在家，观看，动画片)，估计这一语句的概率。以二元语法模型为例，需要**检索语料库中每一个词以及和相邻词同时出现的概率。假设语料库中总词数7542，单词出现的次数如图**所示。