**文本分析**

1. 文本分析定义

一．文本分析的定义

文本分析是指对文本的表示及其特征项的选取；文本分析是文本挖掘、信息检索的一个基本问题，它把从文本中抽取出的特征词进行量化来表示文本信息。文本（text），与 讯息（message）的意义大致相同，指的是由一定的符号或符码组成的信息结构体，这种结构体可采用不同的表现形态，如语言的、文字的、影像的等等。文本是由特定的人制作的，文本的语义不可避免地会反映人的特定立场、观点、价值和利益。因此，由文本内容分析，可以推断文本提供者的意图和目的。

二．文本分析的基础理论知识

1．语义分析与语法分析

从技术层面来看，文本分析任务主要分为语义分析和语法分析。有些文本分析任务只有其一，有些文本分析任务则兼具二者的特征。 语义分析和语法分析具有截然不同的研究框架。语义分析的主要目的是弄清楚段文本内容大体上讲述了什么主题:语法分析的主要目的是要弄清楚文本中各个语言要素之间的组成结构关系。

语义分析可以有不同的应用层次，即在分析粒度上的差异。对于一段文本内容，语义分析既可以很细致，也可以很粗糙。例如，对于“上海已经-一个月没有下雨了。”这句话进行粗糙的语义分析，可以知道这句话讲的是和天气有关的事情。但是如果进行细致的语义分析，不仅可以知道和天气有关系，还能知道和“上海”这个地方有关，并且关注的是“下雨”的问题，以及持续时间长达“一个月”等 很详细的内容。前者只需要观察到“下雨”这个词，就能判断是有关天气的主题， 后者则需要了解整个句子的各个组成部分， 以及各个部分在句子组成中扮演的角色。因此，如果进行粗糙的语义分析只需要与语义分析相关的技术即可，而如果进行细致的语义分析，就要配合使用用语法分析的技术手段。如果进行粗糙的语义分析只需要与语义分析相关的技术即可,而如果进行细致的语义分析,就要配合使用用语法分析的技术手段。

语义分析的典型应用情境有如下几种：

1. 判断两个词汇(或两篇文档)的语义相似性。
2. 将文本要素(词汇,句子、文章)进行结构转化。
3. 提取文章中的关键文本信息。

④提取文章中的主题信息。

......

语法分析的典型应用情境有如下几种：

1. 判断句子中词汇的词性。
2. 判断句子的语法结构,构建语法树。
3. 判断句子中代表特定角色的词汇或短语。

④基于语料库自动构建知识网络(本体或关系数据库)。

......

2.文本的结构化分析

文本类型数据是非结构化的数据，无论是管理类应用还是内容类应用，都只能对结构化的数据进行处理。因此，在文本挖掘中非常重要的技术环节就是数据的结构化过程。基于结构化的数据形式，可以进行统计建模分析，并在计算机中进行自动化处理。文本分析任务需要对文本中基本要素进行结构化处理，因此，需要先对文本数据中的基本要要素进行定义：

图2.2.1 文本分析基本要素框架

* 1. 字（字符）：字（字符）是文本内容的最小组成单元。中文的文本分析任务通常是字，英文的文本分析任务通常是字符。字（字符）通常难以独立构成语义，因此大多数文本分析任务不椅子为最基本的分析单元。
  2. 词汇（词组）：词汇是由字组成的，是具有特定语言含义的最小单元，因此大多数文本分析任务将词汇作为最基本的分析单元。
  3. 句子（短文本）：句子是由词或词组进一步组成的，具有一定的语法结构，通常在对文本进行语法分析时需要将句子作为主要的研究对象。
  4. 文档：文档是由句子组成的，对文本进行分析大多是针对文档进行分析。
  5. 语料库：语料库是由很多文档组成的。对语料库进行分析，有利于更好地了解词汇、句子及文章的内容含义。

3.文本的标准化分析

与文本挖掘技术关系十分密切的另外一个概念就是文本标准化。语言是非常灵活的一种信息形式,这导致其表现形式通常是杂乱无章的,这为文本内容分析带来诸多不便与困扰。因此,在进行文本分析之前,一般需要对其进行标准化处理,增加文本类型数据的一致性、整洁性。对文本进行标准化处理,可以降低后续文本建模、挖掘、分析等任务的难度,并提升结果的准确度。文本的标准化相当于数据挖掘技术的数据预处理阶段,一般在文本结构化任务之前。对文本进行标准化处理的具体方法很多,其依赖于所分析的文本集合的特征。很多标准化方法是基于丰富的数据处理经验进行设计与构建的,当前常用的文本信息标准化方法有以下几种：

1. 对文本集合的语言进行统一，去除其他语种文章。
2. 去除无效符号、特殊符号，以及网页标记。
3. 简体字与繁体字的统一(中文文本)。
4. 去除文本显示格式信息(如空格、回行等)。
5. 对文本中的错别字自动修复。
6. 构建同义词或近义词表，减少文本特征。
7. 过滤低频(词汇出现次数很低)的文本特征。

三．文本分析的主要技术

1.统计中文分词技术

1.1词法分析问题