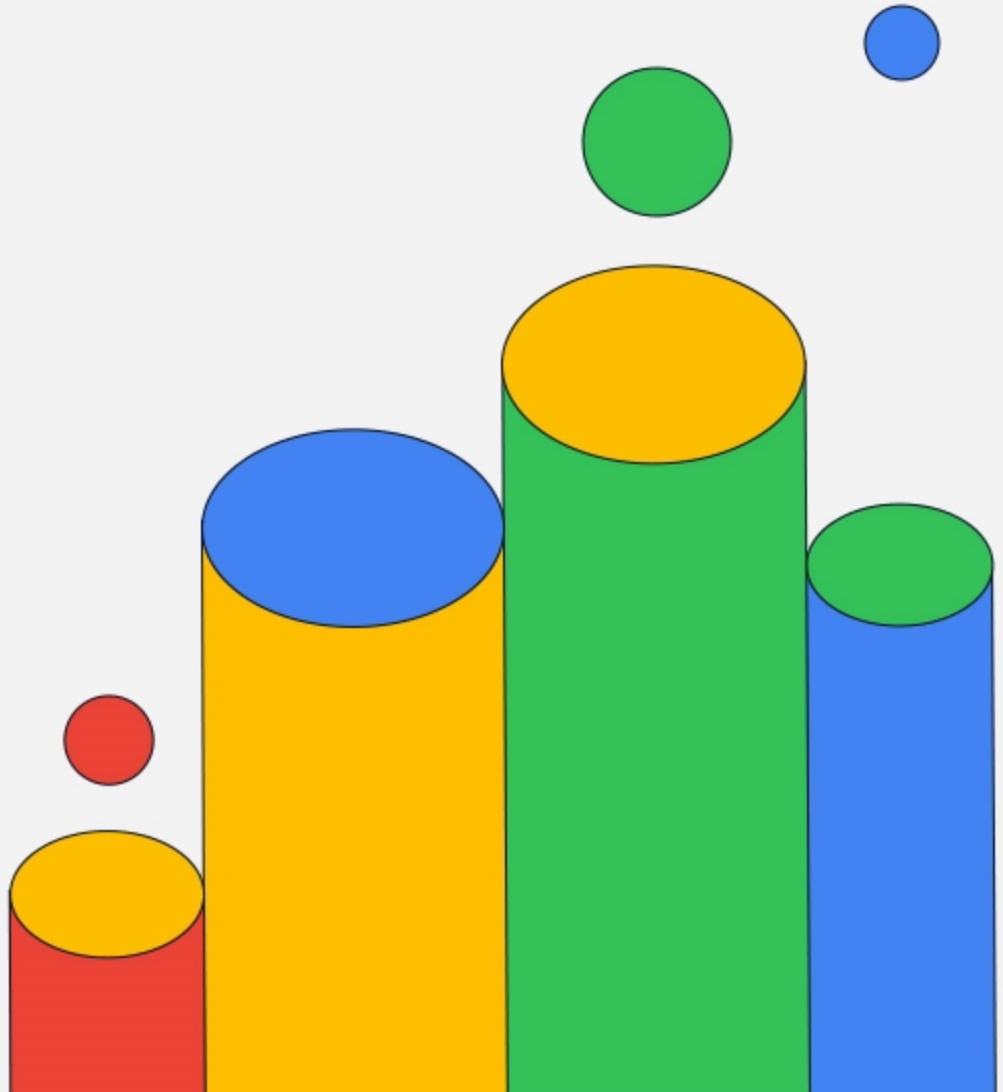
2023

REPORTING

基于统计方法从文本中抽取分词词典

汇报人：

2024-02-03



2023

目录

CATALOGUE

PART 01

基于统计方法的文本处理

统计方法简介

统计方法是一种基于数据驱动的文本处理技术，通过对大量文本数据的统计分析，挖掘出文本中的语言规律和模式。

相比于基于规则的方法，统计方法具有更好的适应性和泛化能力，能够处理更加复杂和多样的文本数据。

在自然语言处理领域，统计方法被广泛应用于分词、词性标注、命名实体识别、句法分析等任务，取得了显著的效果。

文本处理流程

文本预处理

对原始文本进行清洗、去噪、标准化等处理，得到规范化的文本数据。

特征提取

从文本中提取出对后续任务有用的特征，如词频、词性、上下文信息等。

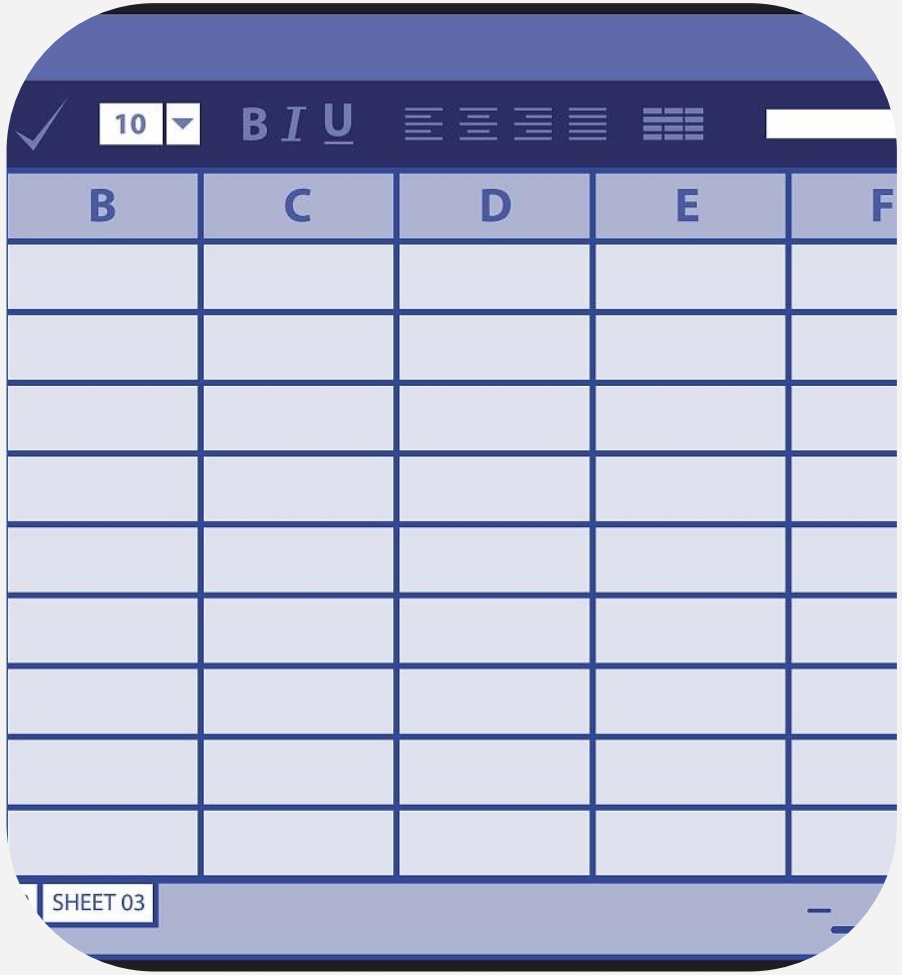
模型训练

利用机器学习算法，基于已标注的训练数据，训练出能够自动处理文本数据的模型。

模型评估与优化

对训练好的模型进行评估，根据评估结果对模型进行优化和调整，提高模型的性能和准确率。

分词词典在文本处理中作用



分词词典是文本分词的重要基础，包含了大量词语及其对应的分词结果，为文本分词提供了可靠的依据。

在基于统计方法的文本处理中，分词词典的质量直接影响到后续任务的准确性和效果。

通过不断优化和完善分词词典，可以提高文本处理的效率和准确性，为自然语言处理应用提供更好的支持。

PART 02

数据准备与预处理

文本数据来源及特点

来源广泛

包括新闻、社交媒体、学术论文等各类文本数据。

语言多样性

可能涉及多种语言和方言，需要考虑语言特性。

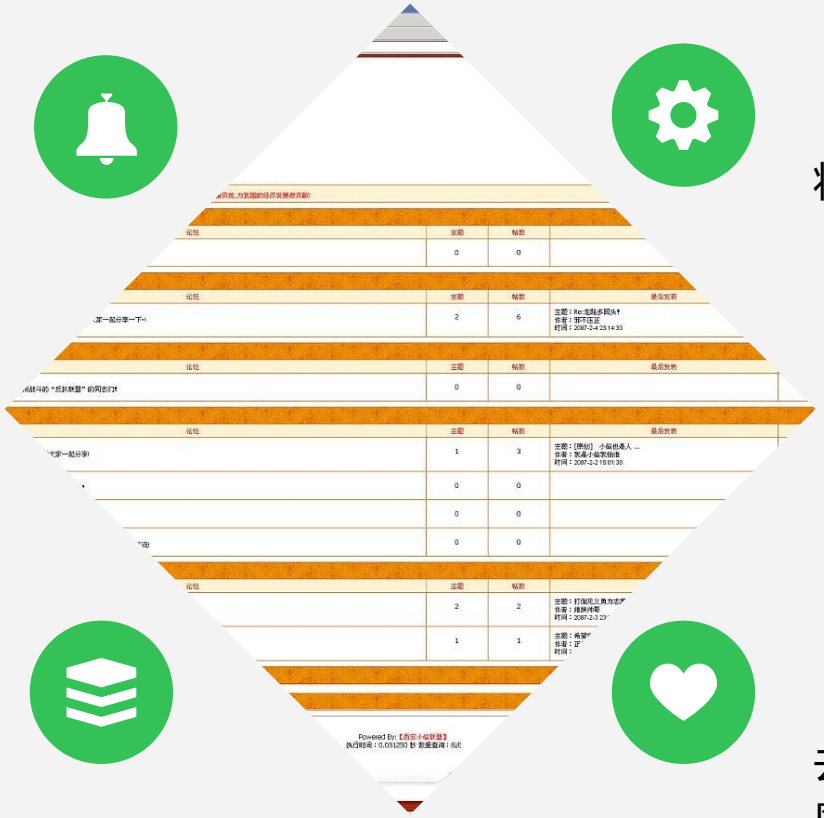
噪声与不规范

文本中可能包含拼写错误、语法错误、缩写等噪声和不规范用语。

数据清洗与整理

去除无关字符

如HTML标签、特殊符号等，保留纯文本内容。



文本分句与分词

将长文本分成句子或单词，便于后续处理。

去除停用词

如“的”、“了”等常用词，减少计算量和干扰。

处理重复与相似文本

去除重复文本或高度相似的文本，提高数据质量。

文本预处理方法

词干提取与词形还原

将单词还原为其基本形式，减少词汇多样性。

特征选择与降维

通过TF-IDF、Word2Vec等方法提取文本特征，降低维度。

N-gram模型

将连续出现的n个词作为一个整体进行处理，捕捉局部信息。

文本向量化

将文本转换为向量形式，便于计算和分析。

PART 03

统计分词算法介绍

基于字典匹配算法

最大匹配法（MM法）

根据词典，将待切分文本从左至右按最大词长进行匹配和切分，适合长词较多的文本。

最小匹配法（Min-Match）

与最大匹配法相反，按最小词长进行匹配和切分，适合短词较多的文本。

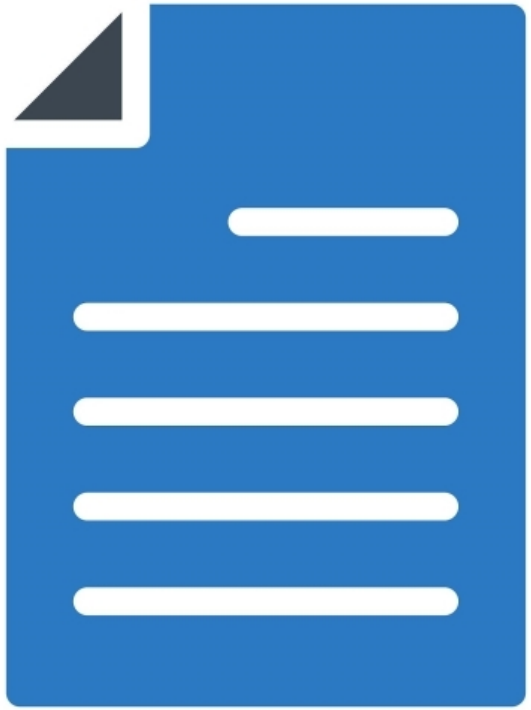
逐词匹配法

将词典中的词逐一与待切分文本进行匹配，找到则切分，适合词典较小且文本较简单的场景。

双向最大匹配法（Bi-MM法）

同时从左至右和从右至左进行最大匹配，比较两种切分结果的词数，选择词数较少的一种，适合歧义消解。

基于统计学习算法



隐马尔可夫模型（HMM）

将分词问题转化为序列标注问题，通过训练得到模型参数，进而对待切分文本进行切分，适合处理复杂文本和未登录词。

条件随机场（CRF）

与HMM类似，也是一种序列标注算法，但考虑了全局最优解，适合处理标注之间存在依赖关系的场景。

深度学习模型

如长短期记忆网络（LSTM）、卷积神经网络（CNN）等，通过训练大量语料学习文本特征，进而实现分词，适合处理大规模语料和复杂文本。