《大数据分析与处理》第3次作业

姓名： 李燕琴 学号： 20195633 日期： 2022-04-21

1. 简述文本大数据分析与传统数据分析的区别和联系，以及文本大数据分析存在的主要挑战。（20分）

**(1)区别和联系：**传统数据挖掘大部分研究主要针对结构化数据，如关系的、事务的和数据仓库数据。 然而现实中大部分数据存储在文本数据库中，如新闻文章、研究 论文、书籍、WEB页面等，且存放在文本数据库中的数据是半结构化数据，文档中可能包含结构化字段，如标题、作者、出版社、出版日期等，也包含大量非结构化数据，如摘要和内容等。

**(2)挑战：**①无确定形式并且缺乏机器可理解的语义和表示形式；②针对文本数据，需要设计新的挖掘方法。

2. 什么是单词的局域性表示（Local Representation）和分布式表示（Distributed R epresentation），它们各自的优缺点是什么？各有哪些代表性方法？（20分）

**(1)单词的局域性表示：  
①含义：**在将单词表示为向量时，每个单词使用向量中独有且相邻的维度。  
**②优点：**简单，容易表示；所有单词都是相互独立的，在其向量空间中所有的词向 量都是正交的，因此其具有很强的判别能力。  
**③缺点**：单词之间相互独立，不能体现出不同单词之间的语义关联和关系远近；且如果单词较多，需要较高的维度进行表示，面临维度灾难问题，其表示和计算远超出普通计算机的计算能力。  
**④代表方法**：独热表示

**(2)单词的分布式表示：  
①含义：**将单词映射到特征空间中，每个单词由刻画它的多个特征来高效表示；在形式上使用稠密实数向量（向量多于一个维度非0，通常为低维向量）来表示单词； **②优点：**分布式表示可以编码不同单词之间的语义关联。可用于相似度评价和单词类比任务。 **③缺点**：不同单词之间的语义关联难以表示；预训练模型需要基于大量的语料库和强大的计算机算力。  
**④代表方法**：  
横向组合关系

隐性语义索引(Latent Semantic Indexing， LSI)

概率隐性语义索引(Probabilistic Latent Semantic Indexing， PLSI)

隐性狄利克雷分布(Latent Dirichlet Allocation，LDA)

纵向聚合关系

神经网络概率语言模型(Neural Prob. Language Model, NPLM)

排序学习模型(C&W)

上下文预测模型(Word2Vec)

全局上下文模型(GloVe)

3. 简述潜在狄利克雷分配（Latent Dirichlet Allocation, LDA）模型的基本原理（20分）

**基本思想：**基于贝叶斯学习的话题模型，假设每个文本由话题的一个多项分布表示，每个话题由单词的一个多项分布表示，进而进行文本集合的概率生成。

**基本步骤：**首先随机生成一个文本的话题分布；之后在该文本的每个位置，依据该文本的话题分布随机生成一个话题；然后在该位置依据该话题的单词分布随机生成一个单词，直至文本的最后一个位置，生成整个文本；重复以上过程生成所有文本。

4. 简述TF-IDF模型的基本原理（20分）

TF是指词频，即一个具体的词在该文章中出现的频率；

IDF是指逆文档频率，可由总文档数目除以包含该词语之文档的数目，再将得到的商取对数得到；

如果某个词或短语在一篇文章中出现的频率TF高，并且在其他文章中很少出现即IDF也高，则认为该词或者短语具有很好的类别区分能力，基于统计学的知识，适合用于分类。

5. 简述文本表达、文本匹配和文本生成的评价方法，并比较它们的联系和区别。（20分）

（1）文本表示

单词表示可用相似度评价（对即衡量单词之间的相似度）、单词类比任务、将单词表示作为机器学习任务的变量特征衡量其解决实际问题的性能。

**更看重文本表示的表征能力和执行相关学习任务的能力；文本表示将人类的语言符号表示成计算机能够处理的向量，是理解文本大数据的基础。**

（2）文本匹配

分类准确率、精确率P@k，召回率R@k、MAP（平均精度均值指标）、MRR（平均倒数排名）、nDVG（归一化折扣累计收益）等方法

**基于文本匹配特定任务设定的评价指标；文本匹配将人们的信息需求与文本内容相关联，是运用文本大数据的基础。**

（3）文本生成

一是内在评价方法，即内在评价衡量系统的性能，它通常与文本质量有关。如，所生成文本的正确性和可读性都是内在的评价，包括人的主观评价、基于语料库的仿人工评价；评价指标有n-gram重合度，句子距离，内容重合度。

二是外在评价方法，外在评价是评估系统是否真正实现了任务的目标。

**文本生成旨在模拟人类语言交流，是人工智能的高级目标，是产生文本大数据的基础。**