基于Spark的场景应用调研

刘新明，蒋昱风，刘钰，李瑞

一、自然语言理解及关键词提取

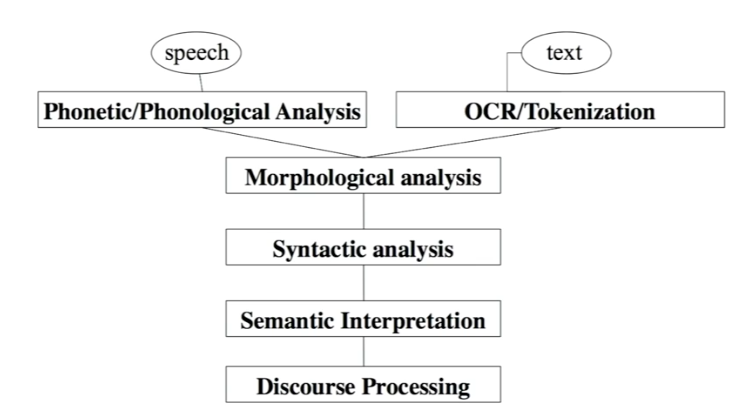
概要

通过使用自然语言处理与深度学习，使得机器具有初步理解自然语言的能力。进一步，机器在理解一段文字的 基础上提取出这一段文字的关键词等有效信息。例如，针对一篇新闻稿，机器在理解这篇新闻稿的基础上，从 中提取出这篇新闻稿的关键词，包括新闻中的时间、地点、人物、事件发展以及情感色彩等。这一实现可以使 人们快速了解一篇文章的主要内容。

这一实现可以应用于其他地方。例如，通过比较不同文章中提取出的关键词，搜索相似的文章新闻稿等，或者 对不同文章进行分类。又比如，通过深度学习算法，对某一类型的文章进行学习并实现对于文章的仿写。

自然语言处理

自然语言处理是计算机科学、人工智能和语言学的交叉学科，主要用于让计算机处理或者理解自然语言，从而 在此基础上做一些其他事情，除了上文提到的，还有网络购物、预约系统、问答系统等。自然语言处理设计的 主要难题是对于语言意思的完全理解及表征（representation)。

人类语言是一种用于传达谈话者或者写作者想要表达的信息的信号系统，具有离散性、符号性、模糊性等 特点。而且一段文章中的标点、语气词、修辞手法（比喻、反语、拟人等）都会影响到一段文章的理解。人类 语言具有的这些特点决定了自然语言处理的复杂性。如下图为自然语言处理的层次。  


深度学习

深度学习是属于机器学习的一个领域。大多数机器学习算法需要人们手动设计表示方法并输入特征才能起作用，这就需要不断地优化各个特征的权重以做出最好的预测。

而深度学习尝试自动地学习好的特征或者表征方法,主要通过神经网络实现。对比传统机器学习，深度学习这一特点的好处显而易见。传统的机器学习需要花费很长的时间去设计特征并优化，而深度学习提供了一种灵活的通用的学习框架来表征语言。并且深度学习可以实现无监督（通过原始数据）的学习和监督式（比如指定积极或消极标签）的学习。

自然语言处理+深度学习

可以通过使用表征学习和深度学习中的方法来解决处理自然语言处理中的问题。

在自然语言中，一个单词由各种词素构成。而在深度学习中，可以用一个多维向量来表示一个词素或者单词。这个多维向量的位置方向等都有特定的意义，并且具有相似意义的单词在多维空间中具有详尽的位置。在神经网络中，表示词素的向量构成表示表示单词的向量，表示单词的向量构成表示词组的向量，表示词组的向量又构成了表示句子的向量。通过这种方式，建立起了一张用来表征自然语言的神经网络。并在表征自然语言的神经网络的基础上，进行对于自然语言中词性、语法、语义的分析，从而提取出可能对人们有用的关键信息。

## 二、基于Spark平台的个性化音乐推荐系统

在互联网时代，各类的音乐网站提供了成千上万的歌曲，满足了人们对音乐的需求，为人们带来了极大的方便。但是，数字音乐的迅速发展造成了音乐信息的过载，面对海量的歌曲人们难以快速找到符合自己兴趣的音乐。音乐推荐系统的出现，在大量用户听歌行为和音乐数据上进行分析与计算，能够为用户推荐可能喜欢的音乐，帮助用户快速的发现或者找到自己想要的歌曲。这种推荐服务能够为用户提供良好的使用体验。

音乐网站每天产生大量的用户行为数据，推荐系统的计算往往要耗费大量的时间。传统的单机处理方式难以满足海量数据的计算效率，以及时效性要求，分布式平台则成为大数据时代下推荐系统发展的新机遇, 大数据存储技术和大数据处理技术能够有效地解决不断增长的数据集存储需求和传统推荐系统扩展性差、计算耗费大等问题。

近年来，很多学者针对传统的推荐算法的缺陷做出了改进或者提出了一些新的推荐算法。随着大数据、深度学习的火热兴起，基于内容的音乐推荐算法有了新的发展，A. Ven Den Oord等人[1]提出了一种新的基于内容的深度音乐推荐方法，通过深度学习方法，利用卷积神经网络训练回归模型，直接对音频的内容信号进行分析来预测收听喜好，免去了专家标注的麻烦。基于协同过滤算法的推荐则在Netflix推荐系统大赛和KDD Cup 2009的音乐预测推荐中得到了高度的关注和发展，Y. Koren等人[2]提出了基于矩阵分解模型的协同过滤算法和加入时间动态因素的改进技术，这种考虑隐式反馈信息的模型缓解了数据稀疏性问题，加入时间因素能够把握用户兴趣发生的变化和音乐流行度的改变，提高了推荐的准确性和稳定性。但是现有的主流算法也都有各自的不足，比如主流的协同过滤算法数据稀疏、海量多元数据得不到有效利用、冷启动和可扩展性差的问题;传统的推荐算法没有考虑到音乐行为中特有的连续性行为和环境的影响，造成了推荐效果不理想。

分析现有推荐系统存在种种的问题，基于Spark平台提出并设计一个个性化音乐推荐系统。

对现有的推荐算法进行深入的研究，对协同过滤算法进行改进及基于大数据分析平台的并行化实现； 同时，基于音乐收听的特点，实现一种音乐列表的推荐算法。基于Spark平台构建分布式日志采集模块和分布式日志数据传输模块,用于采集和传输用户隐式音乐行为日志，当算法提交到Spark集群后，分析集群任务的划分及执行情况。构建用户、歌曲、音乐行为之间的模型，从实际需求出发，实现对用户长期喜好预测的音乐推荐和当前会话期内需求偏好的连续列表推荐功能。并将离线推荐结果与实时推荐结果通过统一介质融合的方案,实现对用户隐式行为进行实时推荐反馈的功能。

而针对这个推荐系统，实现“猜你喜欢”的推荐功能和“私人音乐电台”功能。对应到设计的音乐推荐算法，训练模型时通过HDFS导入的数据集中应包括用户属性数据等等；产生推荐结果时导入模型以及音乐数据，为用户推荐Top-N歌曲。系统可以指定产生的推荐音乐数目N，以用户和时间标识不同的推荐结果，存储于推荐结果数据库。“私人音乐电台”功能，是在推荐歌曲的基础上输入模型并确定用户当前正在收听的种子歌曲后，系统可以推荐最符合当前情境和连续性习惯的下一首歌曲，并形成一个连续不断的音乐播放列表，同样以用户和时间标识不同的推荐结果并存储。

[1] Oord A V D, Dieleman S, Schrauwen B.Deep content-based music recommendation[J]. Advances in Neural Information Processing Systems, 2013, 26:2643-2651.

[2] Koren Y. Collaborative filtering with temporal dynamics[C]. ACM, 2009:447-456.

三、手写汉字识别

参考

<https://zhuanlan.zhihu.com/p/24698483?refer=burness-DL>

背景

手写汉字识别技术已取得了较大的进步，然而无约束的手写汉字识别仍然是汉字识别领域有待解决的问题之一。其中影响识别率的一个重要因素是手写汉字中存在大量的相似字，如“夫”，“失”和“天”以及“看”，“着”和“春”等，书写随意性引起的不规则变形更增加了识别的困难。

因此，研究相似汉字识别，是提升无约束手写汉字识别率的关键，该问题的研究最近几年引起了广泛关注，例如有学者在梯度特征提取的基础上采用两级手写汉字识别架构来改善系统识别性能。第一级采用简单快速的分类器进行粗分类，分类结果需要以很高的概率测试样本的正确类别，即获得相似字表；第二级分类器采用一个速度较慢但性能更好的分类器进行精细分类，即从候选相似字集中挑出正确的标注。

训练集

数据集来自于[中科院自动化研究所](http://www.nlpr.ia.ac.cn/CN/folder/folder8.shtml)。下载地址：<http://www.nlpr.ia.ac.cn/databases/download/feature_data/HWDB1.1trn_gnt.zip>

解压后是一些gnt文件，然后可将所有文件都转化为对应label目录下的所有png的图片，共3755个常用汉字，每个汉字约240张图片。如下图：





选择该题目的优点

* 很容易找到的训练集
* 有一定应用价值

选择该题目的缺点

* 最终模型好坏只与最后的算法及参数有关，“分布式”的概念只用于最初的训练模型阶段，可能与主题不符
* 相关的研究有很多了，不容易出彩

可以创新的地方

一些相似汉字即使是人类自己也很难说得清到底是哪个字，但是如果上下文中就很容易辨认出来。因此我们可以尝试在单个汉字识别的基础上进行句子甚至段落的识别，这样准确性可能更高，但同时也会更有挑战性（可能会涉及到NLP）

# 四、疾病预测

随着大数据技术的发展，其应用领域也越来越广泛，尤其是在医疗领域的发展极为迅速。其中疾病预测也是其发展的一个重要应用方向。

2008年谷歌推出了一款名为“谷歌流感趋势”（Google Flu Trends）的产品。工程师们假定：一旦人们患上流感，就可能会在搜索引擎上输入特定的检索词条以获得与流感相关的信息。通过汇总和分析这些检索词条，就能预测流感将在何时何地爆发。2009年，这款产品在甲型H1N1流感爆发几周前成功预测了其在全美范围的传播。不过在之后预测流感爆发趋势时却几乎是美国疾病控制与预防中心统计数据的两倍，存在了很大的偏差。

在国内百度也有一个相关项目叫百度预测，地址如下：<http://trends.baidu.com/open>。也是通过各地的搜索数据来推测疾病的发生情况。比如说最近某个地区有很多人搜索感冒的相关词汇，就可能预示某地发生了流感。政府和相应的机构，比如医院和药店，了解到这些，就可以做成相应的准备。

不过这个预测还是很有局限性的，一个是搜索关键词和疾病爆发即使有相关性，也没有什么因果性，一个人搜索感冒，也不一定就是病了，也可能就是随便搜索玩玩。再次，这样判断出的疾病趋势有一个滞后性，对某些传染病来说可能并没有多大意义。最后，这种预测有很大的偏差，因为公众并不了解疾病和症状的对应情况。比如搜索感冒，但很可能得了非典，关键词和疾病的对应关系无法建立。

仅靠在线数据进行疾病识别和预测是存在风险的，如果要进一步提高准确率的化，就不能仅仅依靠“搜索相关词汇”等单一在线数据来源，还需要从医院的信息管理系统(HIS)及各类医疗卫生信息化平台获取更多的患者信息，通过多样生物标记来确定疾病的发生，提高实时性和准确性。