智能问诊

摘要：随着时间的推进，人们越来越感受到整个国家的电子行业的兴起，其中电商，云技术，还有前几年还未出现的智能手机，现在也是满大街都是，最让人瞩目的应该新起之秀，人工智能。首先，人工智能是数学与哲学的结合体，这么说可能不太懂，人工智能研究的的人类的思维方式和对自然语言的处理，而数学就是将自然语言以数学形式在机器上运行，如同高级语言向二进制的转变，使用数学将自然语言变成机器能懂的形式，比如最基本的词向量，这个关键词很重要，后面章节会经常涉及这个。本次核心部分，也就是深度学习用的Google的Word2vec无监督深度学习算法，这个算法相比其他的深度学习算法实现上要简单，训练的时候有更高的效率，对于结果的分析准确度也很高。其选择的是高维转低维用的是PCA主成分分析技术，其能够降低训练成本，不会产生维数爆炸，而且提炼了词与词之间的关系。输出矩阵用的是softmax函数，词与词不会产生极端的没有联系的情况。总体来说，训练出来的model完全符合预期的model。

关键词：深度学习；Word2vec无监督深度学习；词向量；Python；

1. 绪论
2. 引言

随着人工智能的推广，开始走入了人们的日常生活中，比如，火热的小爱同学和天猫精灵，也因此以前的只有很少一部分的研究人才到现在有了大量的技术和技术人员，又因此人工智能飞速成长，做出了更好的产品，这一套的良性循环，无不说明人工智能正在发展并崛起，而且以后会影响着人们生活的方方面面。

而智能问诊也是人工智能激流中的一员，他可以像医生一样对病人的病症进行分析，而得出病人得的是什么病，而且它的学习不需要像医生那样需要十几年甚至更多年的学习，词汇量也是远超医生的词汇量，不会产生人类的一段时间不接触某样东西，对这个东西的记忆就会淡忘的情况。如果智能问诊做到对每个细节都把握好，比如可以对人体进行扫描，提取体液，进行方方面面的分析，而不是以简单的病人的口诉病症去分析，那么其准确度绝对完美。本次奈何自己能力不够，人体扫描技术目前市场也没有，但坚信，此技术以后只要不断完善，绝对能够给人类带来巨大的贡献。

1. 智能问诊
2. 智能问诊概述

目前智能问诊主要核心部分对病人口诉病症进行分析然后返回反馈信息。其包括唤醒服务，能够只要开机状就可以不断的提供服务，只要喊一下它的名字，就可以完成唤醒，下句话只要说明自己的症状就OK，比如我怎么怎么样或者他怎么怎么样，然后等待系统反馈，反馈完成，一次服务就算完成，进入休眠状态，等待唤醒。

目前其目的是完成急诊医生完成的工作，以后会比医生做的更好，能够不出现误诊情况，而且迅速准确，其可以很容易产生大量终端，可以解决很多病人已经在医院预约好却还一直处于等待看病的情况。

1. 智能问诊构建

开发语言是Python，语言及其灵活，开发库多。唤醒使用snowboy，可以选择公用的唤醒词，也可以自己声音录制唤醒词，区别是设置自己的别人唤醒不了，因为人与人音色不同。深度学习使用Word2vec，训练模型为cbown。文字转语音使用的是pyttsx，声音转byte使用的是Google的SpeechRecognition，byte转文字使用的是讯飞的一个HTTP接口。

1. 智能问诊系统操作

操作及其简单，目前设置的唤醒词是小老弟（个人觉得好玩用了这个词），只要喊声小老弟，系统就被唤醒，然后说出病症，系统就会分析并返回结果。

1. 核心技术及其原理讲解
2. Python

Python语言今年报道已经超过了Java，凡事皆有原因，Python的语法简练，语言灵活，尤其在自然语言方面，强大到无可替代的地步，其解释器不断的改进，性能也完全够用，而且如果真的要比性能，调个c++编写的demo，绝对不慢，所以不过多讨论性能。在开发速度上也是远超其他语言，不像Java 的臃肿，c++的晦涩复杂，比如知乎就是Python开发的。跨平台性，其是解释性语言，跨平台性毫无疑问。而且动态语言，数据类型在运行时确定，也就不用预先去定义类型，方便。所以此系统最终使用了Python作为开发语言。

1. Word2vec无监督深度学习

word2vec中Word词的意思，vec是vector的缩写，向量的意思，放在一起就是词向量，所以我在上面的关键词里写了词向量。虽然只是三个字，但是接下里内容之多不敢相信。

1. 原理简介
2. 数据源构建

比如我们自然语言说一句话“我周末就可以回家啦”，我们现在对这段话分析，我这里用jieba 进行分词 分出来 ['我', '周末', '就', '可以', '回家', '啦'] 得到这样一组词，那我们如果想用Word2vec训练出来的model，至少需要选个词作为目标词汇，我们这里选择周末这个词，那怎么样提炼语境呢，需要控制俩个参数，第一个是skip\_window，如果设置为2，那么此算法会选择以周末为中心选择左边俩个词和右边俩个词，这个是确定窗口大小，不过这个时候会有三组训练数据出现，（‘周末’，‘我’）， （‘周末’，‘就’），（‘周末’，‘可以’），有时候我们不需要全都要，于是还有个参数num\_skips用来控制选择多少个训练数据。

1. 概率model形成

这个是一次拿去数据，而算法会多次去拿数据，那我们就可以得到很多类似于上面的数据源，最终，神经网络基于这些数据源得到得到一个概率分布，这个概率分布就是我们的model，比如拿我们刚刚拿出的一组数据源（‘周末’，‘可以’），周末作为input 词，另一个作为output词，输入input词，整个model会去计算所有output词的可能性，因为上面的数据源，我们已经知道有个“可以”是output里的一员，现在假设数据源里只有“可以”这个output，那此时model会计算出output是“可以”还有其可能性为百分之百。好累啊，讲解一个东西。

1. 词向量
2. one-hot

one-hot，历史上伟大的一步，从上面我说的数据源，可以很容易看出，机器，他是处理不了那些自然语言的，于是one-hot出现了，它的作用是把所有分好的词标好号，还是上面那句话，里面总共有六个词，于是得到一个一组编号，那我们还是拿“周末”来用， 假设周末编号为2，则one-hot表示就是[0,1,0,0,0,0]，按规律来排，“我”的表示就是[1,0,0,0,0,0]，其他我就不多写了，目前解决了第一个问题，计算机如何理解自然语言。

1. pca

主成分分析（pca），上一步拿到的词向量是完全不行的，因为词库可能十几万词汇甚至更多，首先会产生维数爆炸问题，其次，拿到的词向量之间毫无关系，那每个都去分析也是孤立的。于是这里就用到了降维，降维是对高维特征数据处理然后得到低维特征数据。