

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **王晨（****男）** |  | 年龄: | 21 |
|  | 电话: | 18742024733 |
|  | 意向岗位：算法 |  | 邮箱: | wc18742024733@163.com |

意向城市：杭州、深圳 | 求职类型：校招 | 期望薪资：5k-10k

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **教育经历**   |  |  |  |  | | --- | --- | --- | --- | | **2017.9-2021.6** | **大连理工大学** | **本科** | **数字媒体技术** |   英语一90英语二91分布式系统97软件工程92日语90信号处理94数据库系统90计算机图形学90  GPA 3.56/5.0 专业排名45/210 六级606/710 日语N2 143/180    **相关技能**  熟悉常用的机器学习算法，熟悉不同模型算法原理及特性，能够灵活运用模型进行数据分析；  了解 python 语言及其各种数据分析包，如numpy，pandas；  了解SQL语言；  **奖项证书**  大学英语六级、大学英语四级、日语二级证书、亚太数模三等奖、大连理工大学院优秀团干、精神文明奖学金  **个人总结**  **个人总结**  本人责任心强，拥有较好的学习能力；对数据挖掘与数据分析有很大兴趣；喜欢面对挑战，探索创新； 担任两年团支书； 多次参与社区挂职活动、曾参与暑期支教活动、疫情期间帮扶前线医生儿女志愿家教活动。  **项目部分**  项目一：神经网络实现机器翻译  项目描述： 该项目旨在解决机器翻译的问题。该项目基于encoder-decoder方式的NMT能够训练一张能够从一个序列映射到另一个序列的神经网络，输出的可以是一个变长的序列，在翻译、对话和文字概括方面能够获得非常好的表现。encoder把源语言序列进行编码，并提取源语言中信息，通过decoder再把这种信息转换到另一种语言即目标语言中来，从而完成对语言的翻译。  主要职责： 1. 分别取出中英文训练集，在句子的头部和尾部分别加上<BOS>、<EOS>，<EOS>是句子终止符，创建一个index对应文字的字典表； 2. 将训练集的所有文字变成其对应index，利用one-hot方法将index变成NumPy的array类型用于训练模型； 3. 训练 encoder-decoder 神经网络： （1） encoder：每一时间步输入一个单词直到输入终结符为止，由encoder最后一个隐藏层为decoder输入； （2） decoder：将encoder最后一个隐藏层以及其输出X作为下一时间步的输入来生成目标序列中第二个词Y，这样依次进行直到< EOS>，进行线性变化后得到概率值； 4. 计算loss，如果loss大于criterion就使用Adam optimizer更新模型，用gradient descent来降低loss；  最后得到的loss为1.408，得到了较好的模型；  项目二：中文自然语言处理  项目描述：  该项目旨在对中文进行处理，应用jieba库来进行分词，然后分别采用TF-IDF、TextRank、LDA、朴素贝叶斯、FastText、Tensor Flow、LSTM、GRU进行建模分析预测，对比出各个使用场景和正确率。  主要职责：   1. 使用pandas载入数据进行数据处理，去掉na值，将数据转换成list格式； 2. 将list中一个个单词以空格分隔成一个字符串格式； 3. 使用TF-IDF抽取关键词以及使用TextRank关键词提取；使用jieba.analyse.extract\_tags； 4. 将处理后的数据使用gensim库中的corpora、models、similarities模块进行构建模型； 5. 使用gensim.models.ldamodel.LdaModel()分类主题，可以得到每个主题类型的前二十个关键词   朴素贝叶斯处理问题：   1. 针对5个类别先载入数据，去掉na值，并每类取20000个； 2. 预处理：先载入停用词，并且遍历分词进行过滤操作，如果词条长度小于1或者属于停用词则去除； 3. 生成训练集以单词和类别的形式将所有类别的数据都放到一个list中； 4. 将list打乱顺序（shuffle）使得训练集更加可靠，使用zip分出单词和类型，将数据集分割成训练集和评估集，取出训练集中出现频率最大的4000个词转化成向量； 5. MultinomialNB（朴素贝叶斯）进行建模，用score计算准确率，得到准确率为83.5%； 6. 通过StratifiedKFold进行对评估集进行评估，得到准确率为88.2%；   项目三：视觉聊天机器人  项目描述：  该项目旨在实现一个可以对图片和问题进行分析并返回结果的视觉聊天机器人。使用了word2vec和LSTM进行训练，提高了准确率，降低了loss。  主要职责：  1. 预处理数据：先选取最热的1000个问答：构建字典表，answer和出现的次数，按照出现次数进行排序，将1000个最热的问答和图片提取出来，作为train\_data，将答案编号sklearn的preprocessing 来赋值label(labelencoder,fit)用.pkl存储下来，将answer转化成label：输入answer和encoder，构成1000维的数组，由01构成；  3. 利用scipy载入VGG features提取图片特征向量，再和图片一一对应，得到数字化表达格式：用来取得任何一个input图片的数字化表达形式的，也就是一个matrix，再加一层flatten得到的4096维数组，得到CNN features  4. 特征处理：用scipy自带的英文模型，把问句中所有的英文转化为vector，平均化整个句子，也就是NLP中的word2vec，载入scipy英语库，图片维度大小4096维，单词的维度（每个）300维  5. 模型采用MLP多层神经网络完成建模，使用softmax得到概率值，最高概率的结果就是答案；  6. 保存：model.to\_json变成json模式，以hdf5形式保存权重文件供项目使用。  7. 训练之前使用chunklist方法将原始的数据分成一组组的形式为了之后的batch training做准备  9．训练模型并调整一些超参数，进行VGG完成模型训练 |  |