**杜毅\_个人简历\_NLP工程师\_北京**

**基本信息**

|  |  |  |  |
| --- | --- | --- | --- |
| **姓名** | 杜毅 | **学历** | 本科 |
| **性别** | 男 | **毕业院校** | 石家庄学院 |
| **出生年月** | 1993年2月16日 | **邮箱** | 1079240024@qq.com |
| **联系电话** | 18310057309 | **现居地址** | 北京市 |

**求职意向**

|  |  |  |
| --- | --- | --- |
| **目标职位** | **意向城市** | **工作性质** |
| nlp算法工程师 | 北京 | 全职 |

**专业技能**

|  |
| --- |
| 1. 熟练掌握深度学习模型，如bp，CNN， RNN，生成对抗网络，自编码网络等； 2. 熟练掌握词嵌入模型，如N-gram，CBOW，Negative Sampling，word2vec； 3. 熟悉中英文分词、文本过滤、词性标注等常用文本挖掘方法； 4. 熟练掌握相似度计算、命名体识别、情感分析等技术； 5. 熟练掌握Tensorflow技术进行开发，实现自己设计的网络； 6. 熟练掌握sklearn、Numpy、Pandas等常用框架。 7. 熟悉机器学习模型，如RF, XGBoost, SVM等； 8. 了解网络爬虫技术，如requests, beautiful动态和静态爬取方法等。 9. 了解Linux系统及MySQL数据库常用命令。 10. 了解neo4j图数据库常用命令。 11. 了解熟悉前端后端开发。 |

**工作经历**

|  |  |  |
| --- | --- | --- |
| **时间** | **公司** | **职位** |
| 2019.04—2020.06 | Flipboard(中国) | NLP算法工程师 |
| 2017.08—2019.02 | 摩比神奇信息技术 | 算法工程师 |
| 2015.09—2017.08 | 北京农科院 | Android开发工程师 |

**项目经验**

|  |  |
| --- | --- |
| **项目名称** | XXXX行业智能新闻数据分类 |
| **开发工具** | Python， TensorFlow |
| **项目描述** | 红板报新闻需要对新闻文章数据进行归类和分类操作，以便根据用户的点击进行推荐文章的操作。 |
| **责任描述** | 1. 数据预处理：数据为设施信息及所属类别，处理缺失值，将字、计数和标签id化，出现次数排名前10000的字pickle持久化，文本pad为固定长度，生成器每次随机取64条数据。 2. 模型构建：RNN模型为2层128维的GRU，在最后一个时刻的输出后接fc、dropout层并使用Relu激活，Adam的学习率为1e-3，训练完成后测试用时30s，F1为92.76%。CNN模型使用256个尺寸为5的1d卷积核，全局最大池化GMP后步骤同上，测试F1为96.49%，用时仅为RNN的一半。 3. 模型优化：将模型替换为BERT，创建ClfProcessor类，self-attention机制首先通过全连接生成QKV layer，随机mask以充分学习上下文信息，根据公式计算attention，合并heads结果并经过3层线性映射，最终F1为98.67%。 |

|  |  |
| --- | --- |
| **项目名称** | XXXX场景智能聊天机器人 |
| **开发工具** | Python， TensorFlow |
| **项目描述** | 通过智能问答的方式替代前期工作人员的一些简单问答，由于机器人速度快而且时刻在线，可以提升用户体验，同时可以达到节省时间、节省人力资源的效果。 |
| **责任描述** | 1. 对用户的输入问题进行意图识别操作。采用textcnn和lstm对问题进行分类。 2. 使用jieba对所有问答对进行分词操作，并采用bucket分桶机制，根据长度对问答对进行分组 3. 基于含attention机制和embdding的seq2seq模型进行训练，并加入dropout机制使模型更具鲁棒性，其中encoder与decoder均为双层LSTM单元 4. 对模型进行优化，分别加入 dropout，使用不同的学习率，不同的梯度更新函数进行优化。 |

|  |  |
| --- | --- |
| **项目名称** | XXXX场景中文自然语言处理 |
| **开发工具** | Python， TensorFlow |
| **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。 |
| **责任描述** | **数据预处理：**   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%； |

|  |  |
| --- | --- |
| **项目名称** | XXXX场景下神经网络实现机器翻译 |
| **开发工具** | Python， TensorFlow |
| **项目描述** | 该项目旨在解决机器翻译的问题。神经网络机器翻译(Neural Machine Translation, NMT)是最近几年提出来的一种机器翻译方法。相比于传统的统计机器翻译（SMT）而言，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，这在翻译、对话和文字概括方面能够获得非常好的表现。 |
| **责任描述** | 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的index，利用one-hot方法将index变成NumPy array； 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，然后由encoder的最后一个隐藏层hidden vector作为decoder的输入；  （2） decoder：最初的输入为encoder的最后一个隐藏层，输出为目标序列词X，然后把该隐藏层以及它的输出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。  5、最后得到的loss为1.408； |

|  |  |
| --- | --- |
| **项目名称** | XX操作手册命名实体识别 |
| **开发工具** | Python， TensorFlow |
| **项目描述** | 为了更好地判断用户查询与候选内容的相关性，对手册中的厂站、平台、要素、专有名词等进行命名实体识别。项目分别试验了BLSTM/IDCNN-CRF、Lattice LSTM-CRF和BERT-NER模型，BERT可使F1值大幅提高。模块上线后，使下游任务问答系统的准确率提高约12%。 |
| **责任描述** | 1. 数据预处理：清洗数据，将标签转换为IOBES，构造字/标签id字典，按batch对文本长度排序后进行padding。 2. 模型构建：embedding为Xavier初始化字和切词信息的拼接，dropout后输入BLSTM或IDCNN层，膨胀卷积块之间shortcut连接。使用Adam优化CRF损失，验证集F1分别为77.83%、85.94%。测试时使用Viterbi解码标注结果，使用Flask部署模型至线上。 3. 模型改进：   （1）Lattice LSTM-CRF：加载Word2vec预训练词向量并归一化，不在词表中的词使用Xavier初始化，模型主干双向LSTM融合词序Trie信息，使用带动量的SGD优化NLLLoss，学习率按epoch衰减，F1为92.56%。  （2）BERT：下载BERT-Base Chinese，创建数据集处理类NerProcessor，将数据存入TFRecord，输入函数input\_fn返回batch数据。模型为12个768的隐层、12 heads的Transformer，将位置信息加入词向量，加载当前及预训练的参数，将最后一层序列输出池化、dropout得到loss后使用TPUEstimator训练3个epochs，测试F1为96.79%。 |

**教育经历**

|  |  |  |  |
| --- | --- | --- | --- |
| **时间** | **学校** | **专业** | **学位** |
| 2012.09—2016.07 | 石家庄学院 | 软件工程 | 学士学位 |

**自我评价**

|  |
| --- |
| 1. 工作适应能力较强，能很快地接收新事物并发现其中规律，总能想出一些新点子来提高工作效率； 2. 善于思考，喜欢钻研，做事认真严谨，具有良好的创新能力，喜欢有创造性的工作； 3. 有较强的事业心和责任感，沟通能力好，具备较好的团队意识，敢于迎接挑战。 |