简　历

个人信息

姓　　名：白宁超　　性　　别：男

出生日期：1990 年 4 月　　政治面貌：中共党员

手　　机：15828515952　　电子邮件：1938036263@qq.com

个人博客：http://www.cnblogs.com/baiboy　职位意向：自然语言处理|数据挖掘

个人技能

1. 熟悉 NLP 体系知识，了解机器学习理论知识，具备数学和统计学基础。
2. 熟悉 Python、Java、Linux shell、C#、SQL 等编程语言，扎实的数据结构和编程能力。
3. 熟悉分类、聚类、序列标注、新词发现、NER、HMM、CRF、中文分词、信息抽取、模型评估等 NLP 算法。
4. 熟悉 StanfordNLP,HanLP、Apache OpenNLP、NLTK 等 NLP 开发工具，具有较强的分析能力和沟通能力。
5. 关注 NLP、机器学习相关论坛，个人博客中数个系列博文被机器学习研究会、CSDN、爱可可爱生活等转载。

工作经验

2014.12 －2015.11：成都淞幸科技有限公司（50-150）研发部，负责大数据平台下统计算法构件开发。

工作描述：统计算法构件模块是核格大数据平台下算法库的子模块，该平台是为服务西南地区大数据分析而自主研发的平台，其中业务包括宝洁公司等市场数据分析。本人与 10 多位在读研究生一起参与该平台下统计算法模块开发，该算法构件模块是用 Java 语言在 MyEclipse 平台下开发的，共计 12 个单元共计 70 余个子构件，单个构件是独立的且有单独的测试单元和功能描述。本人参与部分统计算法构件的编写并负责整个模块构件合并和文档汇总。

工作成果：通过算法适用环境、优缺点分析等为大数据平台提供基础构件支持。成功完成了统计图形化构件算法（条形图/折线图/直方图等）、统计度量构件算法(均值/中位数/众数等)、统计概率构件算法（随机过程/古典概率/条件概率/全概率/贝叶斯公式等）、统计分布构件算法（二项分布/正态分布/泊松分布等）、统计抽样构件算法、相关与回归构件算法等。

项目经验

2015.12 -至今：基于主动学习方法的中医古文献症状名识别方法研究

项目背景：本项目是导师申请的国家自然科学基金项目，中医知识由结构化和半结构化的中文自由文本构成，通过自然语言处理技术对症状名抽取，再生成条目清晰、简洁明了的关系数据，这对中医临床研究具有很高的价值。整个项目通过基于多种模型的实体识别，最终完成预期的效果。本人负责项目中基于 HMM 的症状名识别方法研究。

责任描述：本人在 MyEclipse 平台下以 1 万行伤寒杂病论数据为语料，首先对手工对语料 20%的数据进行序列标注，标注采用 BIO 方式，即 B 表示实体的边界开始 I 表示实体中间部分 O 表示实体结束。然后采用主动学习的方式对其进行序列标注，标注完成后通过训练数据的测试数据训练标注模型。最后，基于 HMM 模

型解决预测问题的知识即采用 Viterbi 算法建模实现，其中初始概率计算是个难点，需要机器学习算法对参数模型训练，本项目中简单采用词频技术解决的。然后就是转移概率和发射概率的求解了。每次采用 Viterbi 上次最大似然迭代，最后利用回溯法完成病症名识别工作。之后通过精确率、召回率、F 度量值作为指标对结果进行分析。

2015.10 - 2015.11 ：基于文本处理技术的研究生英语等级考试词汇表构建系统设计与实现

项目背景：在学习文本挖掘相关技术之后，临近研究生英语等级考试，结果发现研究生等级考试不同于高考中考等专门有机构整理出的词汇书。介于此抱着学以致用，自己动手丰衣足食的态度完成此项目。最终实现的效果是一批各种格式的英文文献，通过此程序自动完成词汇表构建，其中词频阈值可以自己指定，进一步可扩展为各专业领域术语知识库，也可以进一步实现在线答题系统。

责任描述：本人通过网络爬虫技术下载历年研究生等级考试真题，再在 MyEclipse 平台下利用 Apache Tika 工具对文本预处理为统一 txt 格式，然后进行文本数据清洗、数据预处理（停用词、去重等）等工作，将基本格式一致、质量较高的数据作为待处理语料。之后通过实现词频统计，词频排序等技术构建出英语等级考试词汇表。

2014.12 - 2015.06 ：基于朴素贝叶斯模型的宾馆入住评价分类的设计与实现

项目背景：本项目是对 5 万行连锁酒店和普通宾馆入住的评价为原始数据，最终将其分为不同满意度的四类数据。包含不同情感文本类别诸如满意、还好、一般、差劲。在 MyEclipse 平台下以 70%数据为训练集和 30%数据为测试集，训练出朴素贝叶斯分类器，并对分类器进行模型评估。

责任描述：本人负责整个项目的开发，首先对数据进行收集为统一格式的原始文本数据，采用 StandfordNLP 工具进行分词处理，之后进行特征提取，通过词语的统计，形成一个词典向量。包含了训练数据里的所有词语（停用词已去除），且每个词语代表词典向量中的一个元素。通过词频方法将原始数据数值化，然后 70%训练集 30%预测集来训练分类模型。最后结合平滑技术和模型指标进行评估，最终精确度为 74.18%，召回率为 90.92%，F 度量值 81.70。

教育经历

2014.09 - 至今	成都信息工程大学	软件工程	硕士（统招学硕）
2010.09 - 2014.06	南阳理工学院	软件工程	学士（统招）

自我评价

1. 曾获“优秀共产党员”（硕士：1/1）、“国家励志奖学金”（硕士：1/1）、“省级一等学业奖学金”（硕士：1/1）、“省级三等学业奖学金”（硕士：1/6）、“优秀本科毕业生”（本科：5%）、“国家励志奖学金”（本科：3%）等。
2. 具有强烈的责任心和对技术的热衷，主页博文（博客园：134篇），读研期间创立成信大技术爱好者社区。
3. 通过国家气象局（2016.02—2016.03）、四川省科技厅（2016.08—2016.09）工作经历，使自己更加稳健。
4. 价值取向积极主动，坚信稳定的团队是执行力的有力保障，座右铭：“事交我您放心，文交我不出错”。