东华大学

毕业设计（论文）任务书

课 题 名 称 ： 基于 SVM 的程序设计相关网页判别系统

学　　院 ： 计算机科学与技术

专　　业 ： 网络工程

姓　　名 ： 杜易初

学　　号 ： 111330125

指 导 教 师 ： 朱国进

二零一五 年 三 月 七 日

一、毕业设计（论文）的目的与要求：

1．培养学生综合运用所学基础课、技术基础和专业课的知识，分析和解决工程技术问题的工作能力。

2．培养学生的创新能力和团队精神，树立良好的学术思想和工作作风。

3.培养学生查阅文献和收集资料的能力。

4.参与社会生产和学校科研和实验室建设工作，为现代化建设作出贡献。

5.机器学习是一种有效的处理数据的方法，而基于统计的机器学习在机器学习领域中有着不俗的表现，期望通过该课题使学生接触到较高级的技术并提升自己的数学素养。

二．毕业设计（论文）的内容：

该系统要从头到尾实现一个网页分类判别系统。按照信息处理顺序，依次要完成数据收集、数据处理、交付使用三个阶段。

数据收集要完成的内容为网页获取、网页解析、添加标注、数据存入数据库。

数据处理模块又可以分为预处理模块和训练模块。前者完成数据库中读取数据、中文分词、特征空间统计、特征向量计算等所有前期预处理功能；后者通过对训练集的训练得到训练模型，交付给最后的数据判别模块使用。

判别模块提供用户交互与程序交互两种使用模式。它们背后的实现流程是一样的，依次是网页获取、网页解析、中文分词、计算特征向量、送往 libsvm 工具结合已有的训练成果进行预测。

三、毕业设计（论文）课题应完成的工作：

3.1　利用Jsoup在互联网中摘取指定url的页面并剔除html标签，得到正文。
3.2　利用IKAnalyzer与Lucene对中文文本进行分词，得到满意的分词后的词组列表。
3.3　从训练样本集中统计出出现过词组的全集，并通过卡方检验进行特征抽取。
3.4　根据卡方检验得到的特征空间计算文章的特征向量并进行归一化。
3.5　将训练集标签与训练集特征矩阵送往SVM进行训练，并对SVM进行参数调优。
3.6　得到训练好的SVM模型，它可以对新的未知样本进行预测，达到目的。
3.7　运用 k-折交叉检验对机器学习的效果进行评判。

四．毕业设计（论文）进程的安排：

序 号	设计（论文）各阶段名称	日　期	备　注
1	摘取指定 url 的页面并得到正文	2014.12.1-2014.12.30	
2	对文本进行分词	2015.1.1-2015.1.30	
3	统计训练样本集，抽取特征空间	2015.2.1-2015.2.28	
4	计算每篇文档的特征向量并归一化	2015.3.1-2015.3.30	
5	理解 SVM 的数学原理和关键参数 gamma	2015.4.1-2015.4.30	
6	完成整个系统，进行答辩	2015.5.1-2015.5.29	

五．应收集的资料及主要参考文献：

[1]同济大学数学系。工程数学线性代数。北京：高等教育出版社，2012。
[2]（土耳其）Ethem Alpaydin 著，范明，昝红英，牛长勇等译。机器学习导论。北京：机械工业出版社，2014。
[3]（加）Simon Haykin 著，申富饶等译。神经网络与机器学习。北京：机械工业出版社，2014。
[4]Microsoft Developer Network 资源库。
http://msdn.microsoft.com/library。
[5]刘浩，韩晶著。Matlab R2012a 完全自学一本通。北京：电子工业出版社，2014.
[6]陈明 等著。MATLAB 神经网络原理与实例精解。北京：清华大学出版社，2014.
[7]Fabrizio Sebastiani.Machine learning in automated text categorization[J].ACM Computing Surveys,2002,334(1):1-47.
[8]Arbib,M.A.,The Handbook of Brain Theory and Neural Networks,2d ed.,Cambridge,MA:MIT Press.2003.
[9]Boyan,J.A.,"Technical update:Least-squares temporal difference learning,"Machine Leaning,vol。49，pp.1-15.2002.
[10]埃史尔 著，陈昊鹏 译。Java 编程思想(第 4 版)。北京：机械工业出版社，2007。

六、任务执行日期：

自 2014 年 11 月 10 日 起，至 2015 年 5 月 29 日 止。

学　　生（签字）_____

指导教师（签字）_____

系 主 任（签字）_____