**浙江大学城市学院第七届大学生服务外包创新应用竞赛**

基于jieba分词的

主观题评分程序

小组成员： 徐余浩(组长)

王文海，胡奕舟，潘启涵

目录

[一. 前言 3](#_Toc28337686)

[**1.1** 编写目的 3](#_Toc28337687)

[**1.2 公司**背景 3](#_Toc28337688)

[**1.3** 业务背景 3](#_Toc28337689)

[二. 创意描述 3](#_Toc28337690)

[**2.1** 业务说明 3](#_Toc28337691)

[**2.2** 主要参加人员 4](#_Toc28337692)

[**2.3 产品** 4](#_Toc28337693)

[2.3.1 程序 4](#_Toc28337694)

[2.3.2 文件 6](#_Toc28337695)

[三. 功能简介 6](#_Toc28337696)

[**3.1** 大致功能 6](#_Toc28337697)

[**3.2** 技术要求与指标 6](#_Toc28337698)

[四. 特色描述 7](#_Toc28337699)

[**4.1** 小组特色 7](#_Toc28337700)

[**4.2** 算法分析与介绍 7](#_Toc28337701)

[五. 开发工具与技术 9](#_Toc28337702)

[**5.1**开发工具 9](#_Toc28337703)

[**5.2** 技术要求与指标 9](#_Toc28337704)

[**5.3** 主要阶段工作产品 9](#_Toc28337705)

[六. 应用环境 10](#_Toc28337706)

[**6.1**应用环境分析 10](#_Toc28337707)

[七. 结语 10](#_Toc28337708)

[**7.1**总结 10](#_Toc28337709)

# 一. 前言

## **1.1** 编写目的

编写此可行性研究报告的目的，是为了研究项目的问题是否值得去解决，首先需要进一步分析和澄清问题定义。在问题定义阶段初步确定的规模和目标，如果是正确的就进一步加以肯定，如果有错误就应该及时改正，如果对目标系统有任何约束和限制，也必须把它们清楚地列举出来。

AI 时代下，取景翻译，智能分词已不是新鲜的技术，初高中英语作文的自动化批改，也在一些线上学习系统中有所应用，但对于长篇大段复杂中文语义分 析，仍未有较好的解决方案推出，对于大量应用主观题的公考、职业教育、资质 类考试，主观题部分目前依旧依靠老师人工阅卷评分。

## **1.2 公司**背景

华云数据集团连续四年被评为中国云计算、大数据独角兽。深圳市中诺思科 技股份有限公司是以物联网、大数据为技术核心，提供信息化建设为主的高新技 术企业，多年来双方在行业产品开发和研究有众多合作，尤其在物流、大数据相 关专业的教育类产品已经比较成功，但随着教育越来越鼓励学生对知识的综合运 用，而非单纯的知识水平考试，客观题的考察占比越来越低，方案设计等主观题 的考察占比越来越高。传统的考评软件，仅支持客观题自动评分，主观题仍需要 老师人工打分。如何将老师从批改试卷、作业这类重复性工作中完全解放出来， 特编写此命题进行共同探讨。

## **1.3** 业务背景

随着职业教育越来越鼓励学生对知识的综合运用，而非单纯的知识水平考试， 客观题的考察占比越来越低，方案设计等主观题的考察占比越来越高。传统的考 评软件，仅支持客观题自动评分，主观题仍需要老师人工打分。在教学智能化日 益提升的今天，函需一款软件将老师从批改试卷、作业这类重复性工作中完全解 放出来。

# 二. 创意描述

## **2.1** 业务说明

随着职业教育越来越鼓励学生对知识的综合运用，而非单纯的知识水平考试， 客观题的考察占比越来越低，方案设计等主观题的考察占比越来越高。传统的考 评软件，仅支持客观题自动评分，主观题仍需要老师人工打分。在教学智能化日 益提升的今天，函需一款软件将老师从批改试卷、作业这类重复性工作中完全解 放出来。

与开放性题目不同，主观题虽然不限用户的答题内容，但仍需要求用户能回 答到几个关键得分点。而由于中文的语义多样性，对于关键得分点的表述，仍可 能存在多种文字描述，而即便出现关键得分点，如果语义完全错误，仍不能认为 表述得分，所以需要对用户的回答语义进行正确识别，并对关键词进行匹配。学 生可以针对提供的部分测试样例做训练，本实验提供的测试样例均来自教师对于 学生的真实评分。

如以下示例： 问题描述：物流系统由什么组成？

答案描述：物流系统是由物流作业系统和支持物流信息流动的物流信息系统 两部分组成。

答案关键字：物流作业系统、物流信息系统

回答方式： 物流作业系统和物流信息系统 满分 物流作业和物流信息 满分 不是物流作业和物流信息 0 分 作业和信息 50% 物流作业 50% 物流管理、物流采集、物流统计 0 分

## **2.2** 主要参加人员

组长：徐余浩

王文海、胡奕舟，潘启涵

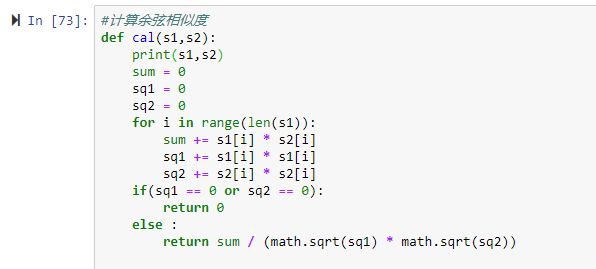
## **2.3 产品**

### 2.3.1 程序

**文本数据的读入。**



**相似度计算函数：**



**主程序**



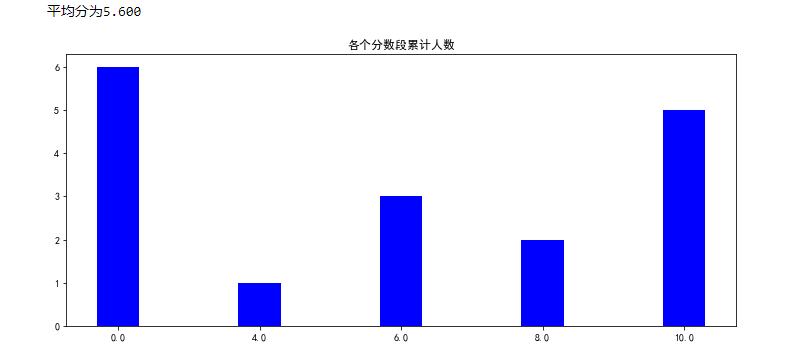
**计算结果：**



**Excle中数据修改与存储：**



**数据分析：**



### 2.3.2 文件

（1）项目概要介绍；

（2）项目简介 PPT；

（3）项目详细方案；

（4）项目演示视频；

（5）团队自愿提交的其他补充材料。

# 三. 功能简介

## **3.1** 大致功能

与开放性题目不同，主观题虽然不限用户的答题内容，但仍需要求用户能回 答到几个关键得分点。而由于中文的语义多样性，对于关键得分点的表述，仍可 能存在多种文字描述，而即便出现关键得分点，如果语义完全错误，仍不能认为 表述得分，所以需要对用户的回答语义进行正确识别，并对关键词进行匹配。学 生可以针对提供的部分测试样例做训练，本实验提供的测试样例均来自教师对于 学生的真实评分。

如以下示例： 问题描述：物流系统由什么组成？

答案描述：物流系统是由物流作业系统和支持物流信息流动的物流信息系统 两部分组成。

答案关键字：物流作业系统、物流信息系统

回答方式： 物流作业系统和物流信息系统 满分 物流作业和物流信息 满分 不是物流作业和物流信息 0 分 作业和信息 50% 物流作业 50% 物流管理、物流采集、物流统计 0 分

## **3.2** 技术要求与指标

（1）能够阐明自己设计的系统的评分机制；

（2）考虑算法的效率，包括时间和空间性能；

（3）针对测试样例（部分提供给学生），统计参赛人员设计的算法得到的 学生各题分数与教师提供分数之间的绝对差值，所有学生的各题分数绝对差值最 小者为优。测试为开放性测试，不仅限于提供的样例数据。

# 四. 特色描述

## **4.1** 小组特色

Git:小组文件上传和下载的地方，担任中转站

Jieba分词工具：介绍结巴中文分词涉及到的算法包括：

(1) 基于Trie树结构实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图（DAG)；

(2) 采用了动态规划查找最大概率路径, 找出基于词频的最大切分组合；

(3) 对于未登录词，采用了基于汉字成词能力的HMM模型，使用了Viterbi算法。

分词模式

结巴中文分词支持的三种分词模式包括：

(1) 精确模式：试图将句子最精确地切开，适合文本分析；

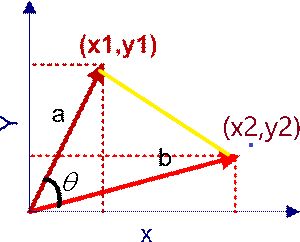
(2) 全模式：把句子中所有的可以成词的词语都扫描出来, 速度非常快，但是不能解决歧义问题；

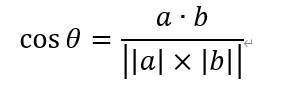
(3) 搜索引擎模式：在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。

## **4.2** 算法分析与介绍

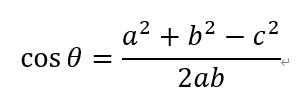
核心算法：余弦相似度

余弦相似度用向量空间中两个向量夹角的余弦值作为衡量两个个体间差异的大小。余弦值越接近1，就表明夹角越接近0度，也就是两个向量越相似，这就叫"余弦相似性"。

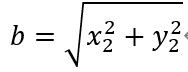
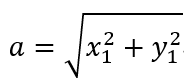




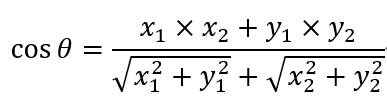
根据三边长度求余弦值：



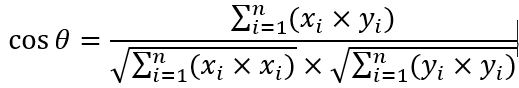
可知三边长度为：



代入计算可得：



扩展至N维空间向量中：



具体算法实现过程为：

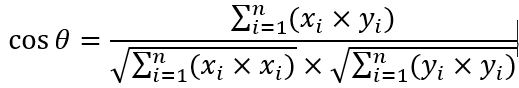
1、分词

2、列出所有词，并按顺序进行标号

3、按出现顺序对两个想要比较的字符串中的每个分词进行标号匹配

4、对每个分词进行独热编码，统计出现次数

5、余弦相似度计算得出结果



复杂度：



# 五. 开发工具与技术

## **5.1**开发工具

Jupyter Notebook：

Jupyter Notebook是一个交互式笔记本，支持运行 40 多种编程语言。

Jupyter Notebook 的本质是一个 Web 应用程序，便于创建和共享文学化程序文档，支持实时代码，数学方程，可视化和 markdown。 用途包括：数据清理和转换，数值模拟，统计建模，机器学习等等 。

Jieba分词：

分词模式：jieba中文分词支持的三种分词模式包括：

(1) 精确模式：试图将句子最精确地切开，适合文本分析；

(2) 全模式：把句子中所有的可以成词的词语都扫描出来, 速度非常快（使用中）；

(3) 搜索引擎模式：在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。

Pandas：

pandas 是基于NumPy 的一种工具，该工具是为了解决数据分析任务而创建的。Pandas 纳入了大量库和一些标准的数据模型，提供了高效地操作大型数据集所需的工具。pandas提供了大量能使我们快速便捷地处理数据的函数和方法。在本次项目中，pandas主要用于excle数据的预处理，包括读写操作。

## **5.2** 技术要求与指标

（1）能够阐明自己设计的系统的评分机制；

（2）考虑算法的效率，包括时间和空间性能；

（3）针对测试样例（部分提供给学生），统计参赛人员设计的算法得到的 学生各题分数与教师提供分数之间的绝对差值，所有学生的各题分数绝对差值最 小者为优。测试为开放性测试，不仅限于提供的样例数据。

## **5.3** 主要阶段工作产品

第一阶段：概要介绍文档

第二阶段：算法分工

第三阶段：项目详细方案

第四阶段：功能以及总体功能的结合

第五阶段：PPT制作

# 六. 应用环境

## **6.1**应用环境分析

我国一直贯彻实施科教兴国战略,强调了科学技术和教育的重要地位,将计算机先进技术应用到教育行业是顺应时代潮流的一项举措。越来越多的学科教育将信息化引入了日常教学,对传统教学方式进行了改革,在线考试系统也得到了广泛使用,计算机基本可以完成组卷、发卷、阅卷、分析等全部活动。但是,在线考试系统中的自动阅卷部分仍有技术局限,对于主观题一般还需要人工辅助评分,也就是说目前在线考试系统的美中不足就是针对简答题、论述题等主观题完成自动评分。主要工作如下:首先,完成了对文本进行分句、分词、语句匹配等预处理操作,并给出了详尽的核心算法步骤和合理的分词方案选择。其次,完成了基于多特征融合和单项贴近度的句子相似度算法改进,将句表层特征融合之后作为关键词以外的因素参与到基于单项贴近度计算句子相似度的算法中。并应用到简答题自动评分中,根据相应的成绩计算公式设计完成评分。再次,完成了词语的知网语义相似度算法改进,尤其是在义原相似度计算部分。给出了词语的词林语义相似度计算方案。本文使用基于动态加权策略的知网和词林信息融合的词语语义相似度算法改进方案。并应用到论述题的自动评分中,根据相应的成绩计算公式设计完成评分。最后,将短文本主观题自动评分模块应用到在线考试系统中,实现阅卷模块的简答题和论述题自动评分,完善了自动阅卷模块的部分功能。经过实验和数据分析,验证了本文设计的文本相似度改进算法的合理性以及自动评分方法在短文本主观题自动评分应用中的有效性,基本满足在线考试系统中短文本主观题自动评分模块的需求。

# 七. 结语

## **7.1**总结

一、项目基本情况：

在这几个星期，项目的所有组员团结合作，每一个文档，PPT，算法的分析和解决，都有每一个小组成员日夜奋斗的身影。最后项目的算法，概要文档，详细文档，项目介绍PPT都已顺利完成，小组中每一个成员都功不可没。这次参赛对我们小组的每一个成员来说意义都十分重大，小组中有些是经常参加各种比赛的，经验丰富的成员，也有一些是第一次接触团队合作完成项目的成员。对于经验丰富的成员来说，这一次项目无疑是一个增加阅历的机会。而对于第一参加这种团队项目的成员来说，这次项目的意义更为重大。首先，他们体会到了小组合作完成项目的每一个流程，对一个项目需要干什么，自己要如何融入到一个项目中都有了一定的了解，更重要的是，这一次阅历对于他们来说更为重要，这也会是他们团队合作的最重要的一步。

二、建设中的工作情况（最好给每一个小标题都起一个煽情的名字）

首先，我们结合了四个人的理解，选择了一个最适合每一个人的选题，虽然这是最早的一步，但是考虑到小组人员的能力情况，我们给了足够多的时间让每一个人都能理解每一个项目的考察方向，从而做出最优的选择。

第二步，我们开始着手概要简介，我们在进行分工后，每个人对自己的任务都进行了制作并出色完成任务。

第三步，我们开始着手算法的实现，这一步是我们小组重点讨论的部分，在明确了业务要求之后，我们更具每一个成员的能力，做了大概的分工。在实现过程中，有成员遇到了代码方面的问题，只要有成员提出了相应的问题，组员中的所有人东欧会对他代码中的问题进行解答。这一个过程在整个环节中虽然占了最多的时间，但我认为，这也是做能体现出一个团队的的各步骤，成员们都互帮互助，一起谈论最佳的解决方案。最后，虽然有超时，但还是完成了代码的编写。

第四步，我们开始着手项目详细方案，我们在进行分工后，每个人对自己的任务都进行了制作并出色完成任务。

第五步，我们开始着手概要简介PPT，我们在进行分工后，每个人对自己的任务都进行了制作并出色完成任务。

四、存在的不足：在整个项目的基础上，每一个成员都完成了任务，但是在细节方面，对于我们所列出的项目计划甘特图来看，有一些成员并没有按时完成进度，导致整体的项目都向后推移。主要原因是由小组成员是第一次接触项目，对项目的流程并没有很强的意识。对于这个问题，我想在经历了这一次的项目后，所有成员都应该深刻的印象，并在以后的项目中引以为戒。

五、几点体会：在这几个星期，项目的所有组员团结合作，每一个文档，PPT，算法的分析和解决，都有每一个小组成员日夜奋斗的身影。最后项目的算法，概要文档，详细文档，项目介绍PPT都已顺利完成，小组中每一个成员都功不可没。这次参赛对我们小组的每一个成员来说意义都十分重大，小组中有些是经常参加各种比赛的，经验丰富的成员，也有一些是第一次接触团队合作完成项目的成员。对于经验丰富的成员来说，这一次项目无疑是一个增加阅历的机会。而对于第一参加这种团队项目的成员来说，这次项目的意义更为重大。首先，他们体会到了小组合作完成项目的每一个流程，对一个项目需要干什么，自己要如何融入到一个项目中都有了一定的了解，更重要的是，这一次阅历对于他们来说更为重要，这也会是他们团队合作的最重要的一步。

三、项目改进计划：

在项目功能上，在一些否定词，情感词以及双关词的判断上仍存在一定的缺陷，在接下来的工作中我们会通过一系列机器学习与NLP文本处理的方法去改进和弥补这些缺陷。

数据分析也是未来会进行重点研究的内容，在数据集逐渐完整的情况下，我们会针对一些数据特征进行数据可视化以及数据的分析，最后在web端呈现给用户。