

UNDERGRADUATE PROJECT PROPOSAL

|  |  |
| --- | --- |
| **Project Title:** | **Job matching and salary forecasting based on python using k-means** |
| **Surname:** | **Hu** |
| **First Name:** | **Jiao** |
| **Student Number:** | **201918010406** |
| **Supervisor Name:** | **Joojo Walker** |
| **Module Code:** | **CHC 6096** |
| **Module Name:** | **Project** |
| **Date Submitted:** |  |

Table of Contents

Contents

[1. Introduction 3](#_Toc118654488)

[1.1. Background 3](#_Toc118654489)

[1.2. Aim 4](#_Toc118654490)

[1.3. Objectives 5](#_Toc118654491)

[1.4. Product Overview 5](#_Toc118654492)

[1.4.1. Scope 5](#_Toc118654493)

[1.4.2. Audience 5](#_Toc118654494)

[2. Background Review 6](#_Toc118654495)

[2.1. Summary of existing approaches 6](#_Toc118654496)

[2.2. Brief summary of related literature 6](#_Toc118654497)

[3. Methodology 7](#_Toc118654498)

[3.1. Approach 7](#_Toc118654499)

[3.2. Technology 8](#_Toc118654500)

[3.3. Version management plan 8](#_Toc118654501)

[4. Project management 8](#_Toc118654502)

[4.1. Activities 8](#_Toc118654503)

[4.2. Schedule 9](#_Toc118654504)

[4.3. Data management plan 9](#_Toc118654505)

[4.4. Deliverables 10](#_Toc118654507)

[5. References 10](#_Toc118654508)

# 1. Introduction

## 1.1. Background

1. Background

Although there are many domestic recruitment websites and other platforms for publishing recruitment information, enterprises or employers are too scattered when publishing recruitment information, and the retrieval function of general recruitment websites is limited.

Therefore, there is an urgent need for a platform that can integrate as much recruitment information as possible, so that job seekers can search recruitment information through the platform, and obtain a large amount of recruitment information and statistical information in a short time, so that job seekers can find their favorite jobs faster.

1. Terms

Python clawer:

In order to replicate the act of browsing a website, the crawler actually makes use of a browser. The entire procedure is divided into three stages: opening a web page, data extraction, and data saving. There are tools available in Python for usage throughout these three periods (Medium,2022).

Data cleaning:

The practice of correcting or deleting inaccurate, damaged, improperly formatted, duplicate, or missing data from a dataset is known as data cleaning. There are several ways for data to be duplicated or incorrectly categorized when merging different data sources. Even though results and algorithms appear to be right, they are unreliable if the data is inaccurate (Tableau, 2022).

Information visualization:

A user can better comprehend data by displaying it visually and meaningfully through information visualization. (IxDF,2022).

collaborative filtering and k-means:

The collaborative filtering recommendation algorithm based on K-Means clustering can be applied to user based and project-based collaborative filtering recommendation algorithms. As one of the methods to reduce data sparsity and improve recommendation accuracy, a collaborative filtering recommendation process can achieve multiple K-Means clustering. A collaborative filtering recommendation process can accomplish multiple K-Means clustering as one of the ways to lessen data sparsity and increase suggestion accuracy (linge,2020).

1. Outline

The first section mainly introduces the background of this website and the related main contents, so that users who read this article can have a general understanding of this topic and my motivation to develop this application.

The second section have a comparison of features of related applications, to understand what advantages and disadvantages, to help figure out what will be original about my approach. There is also a summary of relevant literature.

The third section is the main method for developing：software development model，requirement gathering，specification method. Technology, version management plan.

The fourth section is project management which specifically discuss how to implement objective, schedule of the whole project and use GitHub for store documentations.

## 1.2. Aim

Computer-related job data platform built with Flask.

The crawler collects the job data from the boss application in the form of keywords and visually displays it on the web with tools for salary forecast and job matching.

## 1.3. Objectives

1. Home page data display: visual representation of the number of positions, industry average salary, industry minimum salary, industry maximum salary, statistics on education and salary, general education requirements, city average treatment tree, welfare treatment, popular positions, and other data.
2. Data overview: specific information on the classification of development positions and non-development positions is displayed.
3. Salary forecasting and job matching based on the user’s input information.
4. Test and evaluate

## 1.4. Product Overview

## 1.4.1. Scope

In order to allow job seekers to examine the relevant recruitment information by selecting their educational background and desired recruitment roles, I want to develop a website, crawl the information of many recruiting websites, and then integrate the information. The data will also be displayed so that job seekers may comprehend the demands of the organization rapidly (the visualized part includes salary, enterprise, company benefits and education background). Additionally, they have the option to anticipate the minimum salary and match the appropriate job in accordance with their own requirements.

## 1.4.2. Audience

It is convenient for job seekers to select and view the corresponding recruitment information by selecting their education background and desired job.

# 2. Background Review

## 2.1. Summary of existing approaches

At present, the recruitment visualization system based on web crawler is relatively rare at home and abroad. Search engine companies such as Baidu and Google provide this platform. However, most of them can only do fuzzy search on recruitment information, and cannot do statistics on recruitment information. In addition, because the recruitment information on recruitment websites cannot be reproduced and published without permission, there is no successful case of Python based recruitment visualization system.



Table 1 comparison of software features

## 2.2. Brief summary of related literature

K-means is most effortless and prevalent unsupervised learning calculation which understands the popular clustering issue. It is popular dividing technique in which objects are categorized as fitting in one of K bunches. The guideline thought is to characterize K centroids, one for each cluster. In each and each cluster there may show up centroid or a cluster operator (Institute of Engineering and Technology Devi Ahilya University,2017). Recommender systems empower individuals to share their inclinations, and after that they offer individuals the foremost curiously things. They are an intelligent method to deal with the issue of data overload. They can make valuable proposal for clients and make them come over their curiously things (Dakhel et al., 2011).

# 3. Methodology

## 3.1. Approach

1. Software development model: Waterfall
2. requirement gathering method: using Benchmarking and brainstorm.

The main purpose of this step is to understand the exact requirements of the customer. What are the advantages over other software and platforms, so that users can understand the purpose of designing this website.

1. Specification method

The software must be written with enough notations, explanations, flow diagrams, and schematics. In the future, it is handy to swiftly recollect the past.

To create a historic record, every software modification's specific should be documented in great detail. It is preferable to keep the changed version rather than overwrite the original one.

1. test and evaluation process:

Similar to the system development process, the testing process also needs to be carried out step by step, and each step is a logical continuation of the previous step. Large software systems usually consist of several subsystems, and each subsystem consists of a diversified small module.

large-scale software system testing basically consists of the following steps:

1. Module testing: often finds errors in detailed design and coding in the testing steps of modules.
2. system test: often finds errors in the software design during the test steps of the system, and may also find errors in the requirements specification.
3. The acceptance test: often finds errors in the system requirements specification in the test steps of acceptance.

## 3.2. Technology

system: Windows 10

chrome\_driver

Front-end： html css ajax

Back-end：flask requests Pandas Python (Pycharm)

Database：Mysql 8.0

## 3.3. Version management plan

Use Git repository to manage the product version.

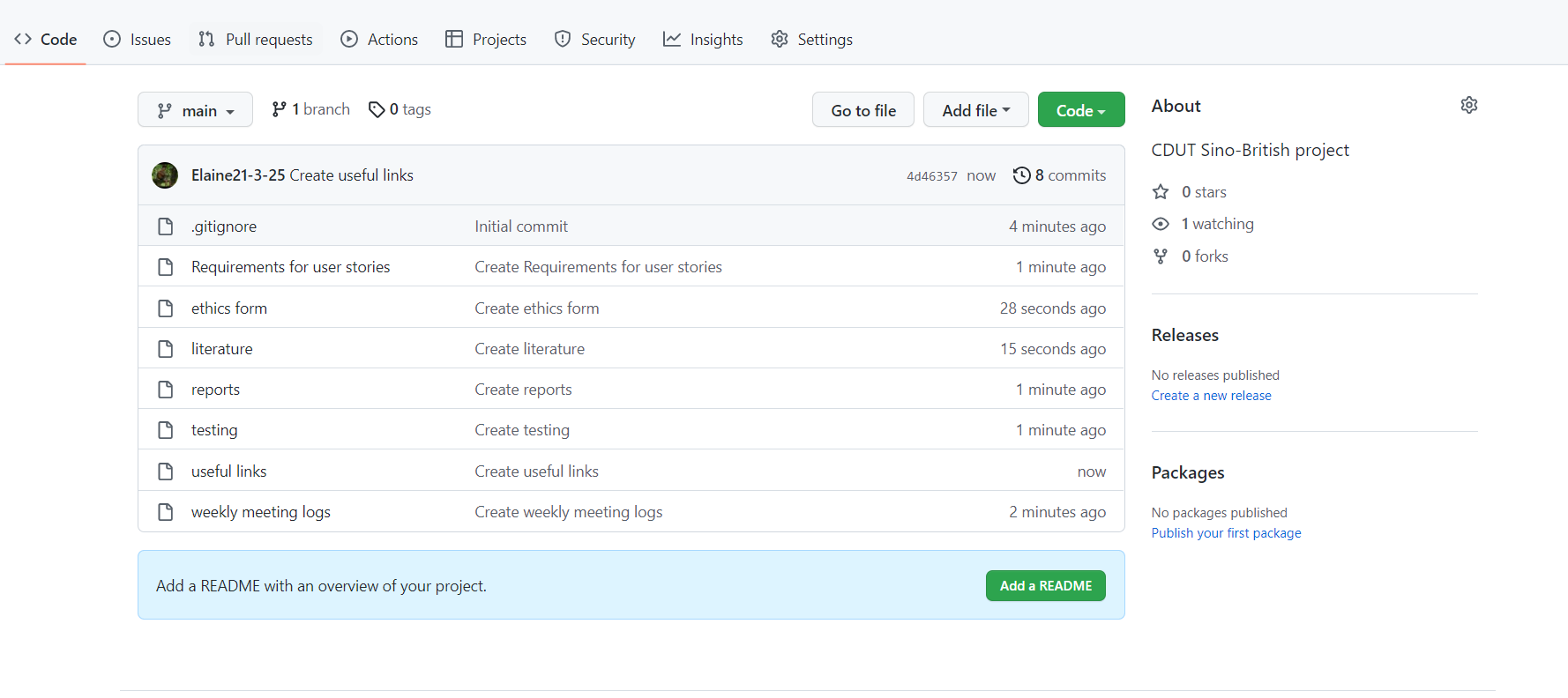


Figure 1 GitHub

# 4. Project management

## 4.1. Activities

1. Do literature research and practice with code for knowledge I need below.
2. Implement the front-end step by step, later will make the pages more detailed
3. Working on Python crawler and data cleaning.
4. Realize information visualization using word cloud and ecahrts.
5. Use K-means algorithm of collaborative filtering with machine learning for salary forecast and job matching.
6. Get information from Boss to test and evaluate.

## 4.2. Schedule

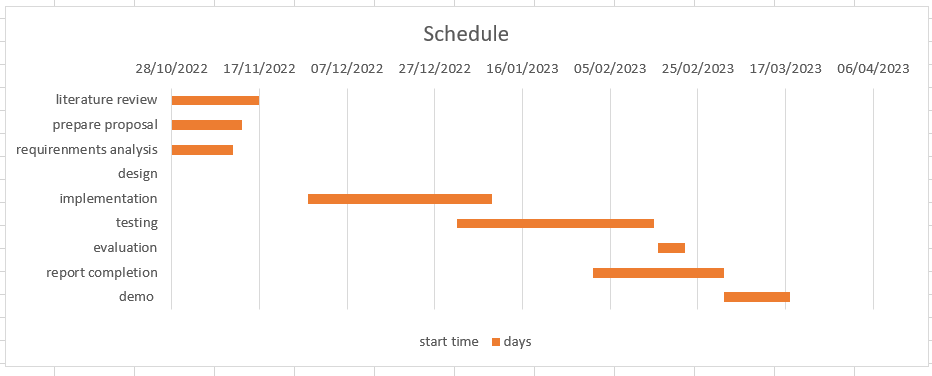


Figure Gantt Chart

## 4.3. Data management plan

Use Git repository to manage the data.

## 

Figure 3 GitHub

## 4.4. Deliverables

In the end, I should submit project proposal, progress report, final report, and project code.

# 5. References

G. M. Dakhel and M. Mahdavi, "A new collaborative filtering algorithm using K-means clustering and neighbors' voting," 2011 11th International Conference on Hybrid Intelligent Systems (HIS), 2011, pp. 179-184, doi: 10.1109/HIS.2011.6122101.

Institute of Engineering and Technology Devi Ahilya University,Indore, M.P, india (2017). Enhanced K-Means Clustering Algorithm Using Collaborative Filtering Approach Available at: <https://www.computerscijournal.org/vol10no2/enhanced-k-means-clustering-algorithm-using-collaborative-filtering-approach/> (Accessed: 4 November 2022)

IxDF (2022) What is Information Visualization. Available at: [www.interaction-design.org/literature/topics/information-visualization](http://www.interaction-design.org/literature/topics/information-visualization) (Accessed: 4 November 2022)

Linge (2020) Principle, process and code implementation of collaborative filtering recommendation algorithm based on K-Means clustering. Available at: <https://blog.csdn.net/u011291472/article/details/104348179> (Accessed: 4 November 2022)

Medium (2022) How to Write a Web Crawler in Python? Available at: <https://medium.com/data-analysis-enthusiast/how-to-write-a-web-crawler-in-python> (Accessed: 4 November 2022)

Tableau (2022) Guide To Data Cleaning: Definition, Benefits, Components, And How To Clean Your Data. Available at: <https://www.tableau.com/learn/articles/what-is-data-cleaning> (Accessed: 4 November 2022)