MOOCS RECOMMENDER BASED ON LEARNING STYLES

ONLINE FORUM THREAD DISCUSSION ANALYSIS TO AID IN RECOMMENDING MOOCS

Software Requirement Specification
Project ID: 19-089

Hilmy S.B.M.

IT 16 0053 72

Bachelor of Science Special (Honors) in Information Technology
Specializing in Software Engineering

Department of Software Engineering

Sri Lanka Institute of Information Technology
Sri Lanka

Date of Submission: 2019-05-13

MOOCS RECOMMENDER BASED ON LEARNING STYLES

ONLINE FORUM THREAD DISCUSSION ANALYSIS TO AID IN RECOMMENDING MOOCS

Project ID: 19-089

Hilmy S.B.M.

Supervisor: Mr. Nuwan Kodagoda

Co-Supervisor: Ms. Kushnara Suriyawansa

Date of Submission: 2019-05-13

Declaration

I hereby declare that the project work entitled “MOOCs Recommender Based on Learning Styles” submitted to the Sri Lanka Institute of Information Technology, is a record of original work done by our group under the guidance of Mr. Nuwan Kodagoda (Supervisor) and Ms. Kushnara Suriyawansa (Co- Supervisor), and this project work is submitted in the fulfillment for the award of the Bachelor of Science (Special Honors) in Information technology Specialization in Software Engineering. The results embodied in this report have not been submitted to any other University or Institute for the award of any degree or diploma. The diagrams, research results and all other documented components were developed by us and we have cited clearly any references we have made.

Name	ID	Signature
S.B.M. Hilmy	IT16005372	

Supervisor: Mr. Nuwan Kodagoda Signature:

Co-supervisor: Ms. Kushnara Suriyawansa Signature:

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Purpose

The purpose of this document is to specify the exact user and system requirements of the Forum Analysis component of the research project. The contents of this document are primarily intended for the customers of the application but will also be of interest to the software engineers building or maintaining the software and other stakeholders of the project. The document describes the services needed by the customer from the system, the processes by which the system will be developed and the constraints under which it operates and developed. The statements regarding the services the system provides will be detailed under functional requirements and the constraints on the service or functions offered will be under non-functional requirements.

1.2 Scope

This document describes the details of the "Forum Analysis" module of the project. This software component will produce a score which is a reasonably accurate rating of the forum activity regarding a specific online MOOC. The application component will support the analysis of forums linked to MOOCs on the platforms Coursera, Edx and FutureLearn, the application will analyze forums within these platforms and external forum sites to deduce the rating. The external forum sites used to gather information will be Class Central, Reddit and Quora.

1.3 Definitions, Acronyms and Abbreviations

MOOC	Massive Open Online Course
SRS	Software Requirements Specification
RAM	Random Access Memory
CLI	Command Line Interface
GCP	Google Cloud Platform
SSD	Solid State Drive
NLP	Natural Language Processing
JSON	JavaScript Object Notation

1.4 Overview

This iteration of development will vastly improve MoocRec by implementing additional features and improving the effectiveness of existing ones. This component aims to improve the recommendation of MOOCs by matching the level of forum activity to the preferences of the learner and by improving the accuracy of the existing MOOC ratings.

Users who prefer learning through peer-to-peer interactions will be suggested MOOCs which have higher active forum users so that he/she will be able to participate in it. And, learner reviews will be analyzed to deduce another rating which can be used alongside the normal ratings to improve the accuracy of it.

This SRS (Software Requirements Specification) will cover the functional and non-functional requirements of the Forum Analysis module of the MoocRec system. Each of these have been described in detail in the three sections of this document.

The first section (Section 1.) will contain an introduction to this document and will go over all the areas which will be covered. The second section (Section 2.) will contain an overall description of the module of the system and the contents of the document.
The third section describes the specific requirements and is intended to be understood by all stakeholders of this project. It will cover all the requirements by describing each one in an in-depth manner.

2 Overall Descriptions

2.1 Product Perspective

2.1.1 System Interfaces

No special system interfaces are required. This module does not directly interact with the other modules in the application. This module (Forum Analysis Module) and the others will be connected through the database.

2.1.2 User Interfaces

All the sub-modules of this module (Forum Analysis) will have command line interfaces (CLI). Graphical User Interfaces (GUI) do not exist because the end-user of this product will not directly interact with these components.

2.1.3 Hardware Interfaces

No special hardware interfaces are required.

2.1.4 Software Interfaces

MongoDB will be used as the database component. Google Cloud Platform for computationally complex processes like NLP.

2.1.5 Communication Interfaces

The computer in which this application will be running on needs to have a download speed of more than 60Mb/s so that the web crawlers will be able to gather data at a reasonable speed.

2.1.6 Memory Constraints

At least 8GB of RAM (memory) will be needed to run the multiple web crawlers and the forum analyzing sub modules.

2.1.7 Operations

The operations of the Forum Analysis component will be carried out in the background without any user involvement once the processes are up and running.

2.1.8 Site Adaption Requirements

No site adaptation requirements are required as it does not directly involve the end-user.

2.2 Product Functions

- Gather online forum data from Coursera, FutureLearn and Edx
  Forum data from MOOCs which are on Coursera, FutureLearn and Edx will be gathered and stored in the database to be used later for processing. Several web crawling sub-modules will be created for this purpose which will gather data repeatedly based on a time interval (daily basis, weekly basis, monthly basis, etc.). Several web crawlers are needed since MOOC platforms are vastly different and have a completely different routing ang page structure. The collected data will be filtered based on certain conditions to maintain the integrity of the data. Once filtered they will be formatted so that they can be processed easily. Finally, the formatted information will be stored in the database.

- Review-Type Forum Analysis
  Data from review-type forums regarding a certain MOOC is used to come up with a rating for the MOOC. Review-Type forums include comments sections and reviews. Sentiments from the messages in comments and review forums is analyzed to obtain a numerical value which can then be used for further processing. This derived rating is used in conjunction with the learners' directly submitted ratings to get a more accurate overall rating. The increased accuracy is because reviews give a more accurate description of the learner's sentiments towards the MOOC. This is done by analyzing the sentiments in the message posts and reviews. This MOOC rating can then be used when recommending MOOCs to learners.

- Forum Meta-data Analysis
  The metadata contained within forum threads can be used to get a rating of how active it is. Metadata refers to data attributes like the date a certain thread was created and when the thread was last active. These types of data are used to achieve this functionality. The data which is available depends on the platform of the MOOC, as some platforms track a lot of information while others do not.

2.2.1 Overall Design

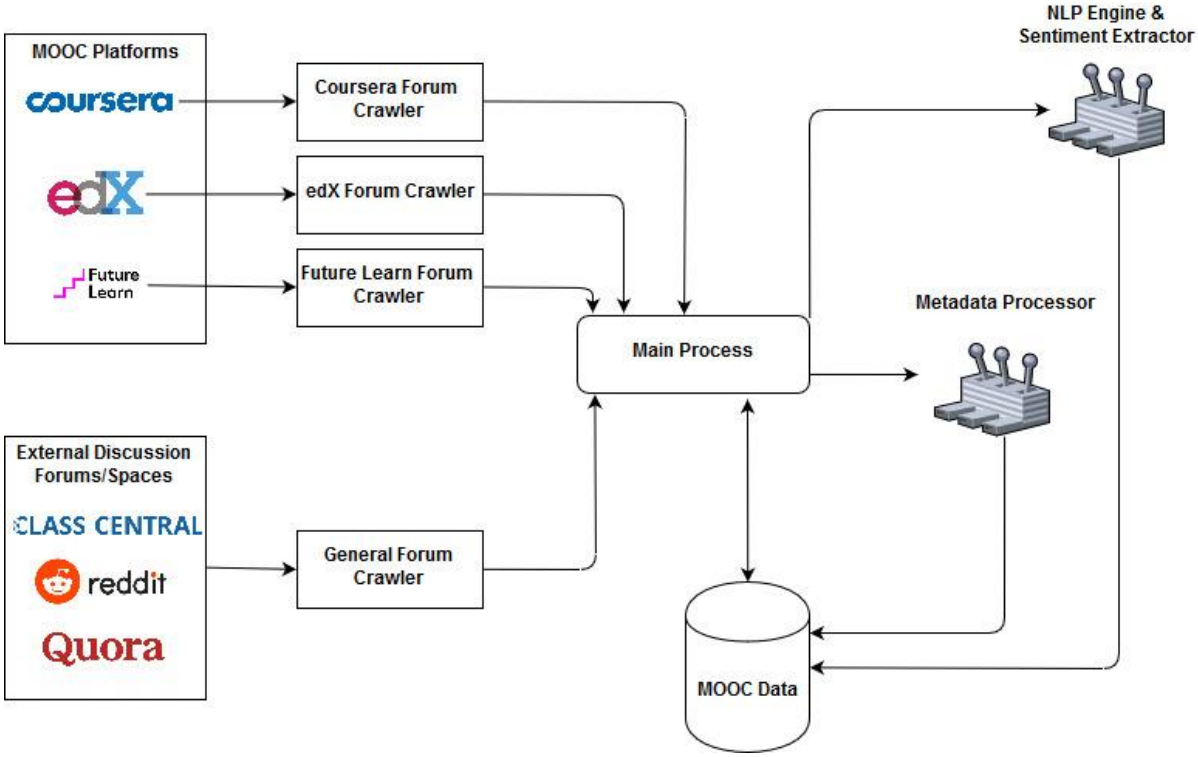

Figure 2.1: Overall Design

2.2.2 Use Case Scenarios

Table 1: Use Case Scenario 1

Use Case Name	Gather forum data regarding a MOOC
Pre-Condition	Forum data about the MOOC • does not currently exist in the database • which currently exists in the database is outdated
Post-Condition	Data about the MOOC • should exist within the database • should not be outdated
Actor	Web Crawler - Backend System
Main Success Scenarios	1. Visit the MOOC platform 2. Login using the user credentials 3. Access the specific MOOC 4. Go to the forums section 5. Traverse the forum threads and store the require data in the database
Extension	1a. Connection issues 3a. The MOOC no longer exists 5a. The connection with database has been severed

Table 2: Use Case Scenario 2

Use Case Name	Calculate Forum Activity Score using Metadata Analysis
Pre-Condition	Forum data about the MOOC • should exist within the database • should not be outdated Metadata sub score not should be present
Post-Condition	Metadata sub score should be present
Actor	Forum Processor - Backend System
Main Success Scenarios	1. Retrieve the attribute data from the database 2. Analyze forums using the attributes available 3. Save information in database
Extension	1a. Data does not exist in the database 3a. Connection with database has been severed

Table 3: Use Case Scenario 3

Use Case Name	Calculate Forum Rating using Sentiment Analysis and Existing Ratings
Pre-Condition	Data about the MOOC • should exist within the database • should not be outdated Sentiment sub score not should be present
Post-Condition	Sentiment sub score should be present
Actor	Forum Processor - Backend System
Main Success Scenarios	1. Retrieve forum thread posts from database 2. Process posts to get sentiment sub score 3. Save information in database
Extension	1a. Data does not exist in the database 3a. Connection with database has been severed

Table 4: Use Case Scenario 4

Use Case Name	Normalize sub scores and calculate final score (rating)
Pre-Condition	The following information must be available • Forum Metadata sub score • Forum Sentiment sub score
Post-Condition	Final forum activity score should be available
Actor	Forum Processor - Backend System
Main Success Scenarios	1. Get sub scores from the database 2. Normalize the sub scores 3. Calculate final score 4. Save final score in the database
Extension	1a. Data does not exist in the database 4a. Connection with database has been severed

2.3 User Characteristics

An ideal user of this back-end component of this system would be a Software Engineer who has experience in Python code, MongoDB and has worked with software which handle large amounts of data. The user also should have some experience with the Google Cloud Platform (GCP) as it is used for Natural Language Processing (NLP).

The ideal user should be capable of going through logs when errors occur and make the required rectifications to ensure the module functions as expected in the future operations.

2.4 Constraints

Forum Analysis is a backend component which involves computationally intensive tasks therefore the computer which is running the component has certain constraints which must be adhered to for optimal functionality.

- 8GB of System Memory (RAM)
- Python is the implementation language (Python 3.7.0)
- MongoDB is the database
- Linux, Windows 8 or higher

2.5 Assumptions and Dependencies

- The computer will have a stable internet connection 24/7
- The internet speed is at least as fast as 100Mbit
- At least 250GB of storage
- Google Chrome is installed
- Chrome Driver is present (Standalone Application)
- An active GCP (Google Cloud Platform) account

2.6 Apportioning of Requirements

2.6.1 Essentials Requirements

- Gathering forum related data from the MOOC platforms Coursera, Edx and FutureLearn
- Analyzing the sentiments of review-type forum posts and calculating the sentiment sub score
- Analyzing the metadata of forums and calculating the metadata sub score
- Normalize the sub scores to compensate for the different types of data used to calculate them
- Use the sentiment sub score along with the normal ratings of MOOCs to get a new improved rating
- Use the two scores when the learner is being recommended MOOCs to improve the accuracy of recommendation

2.6.2 Desirable Requirements

- Compress the data to improve the efficiency of storing forum related information in the database.
- Implement additional web crawlers to gather data from more external web sites like Reddit and Quora.
- Enable multiple web crawler instances to be able to work simultaneously which would significantly increase the scalability of this component.

- Enable the forum analyzer instances to be able to work simultaneously which would significantly increase the scalability of this component.

3 Specific Requirements

3.1 External Interface Requirements

3.1.1 User Interfaces

As this is a back-end component it will not be exposed to the end-user and so it does not have any user interfaces. This component will contain Command Line Interfaces to be used by technical operators who maintain or administrate the application.

3.1.2 Hardware Interfaces

- A network adapter which supports at least 60Mb/s of data transfer
- At least 250GB of storage – Hard disk or SSD
- At least 8GB of memory (RAM)

3.1.3 Software Interfaces

- Python 3.7
- MongoDB will be used as the database component.
- Google Cloud Platform for computationally complex processes like NLP

3.1.4 Communication Interfaces

The computer in which this application will be running on needs to have a download speed of at least 60Mb/s so that the web crawlers will be able to gather data at a reasonable speed.

3.2 Functions

3.2.1 Gathering Data from Forum Threads

Table 5: Gathering Data from Forum Threads

Description	The web crawlers will gather data from forum threads regarding a MOOC
Sequence of Operations	1. Initialize and start platform specific crawler (e.g. Coursera Crawler if MOOC exists on Coursera) 2. Initialize and start external forum site crawler if topic related data does not exist within the database Below operations are common to both web crawlers 1. Retrieve the list of threads present 2. Iteratively traverse the threads and collect forum posts data 3. Store the collected data in the database 4. Notify main process that the task is complete

Validity Checks	It should have been more than a day (24 hours) since data was last gathered from forum threads regards that specific MOOC.
Input	MOOC Details (URL, Title, ….)
Output	Forum Thread Data
Error Handling	If collecting data from a certain forum thread fails, the operation will keep retrying for a set amount of time. If the time threshold exceeds the forum thread is considered deleted (non-existent)

3.2.2 Calculate Forum Activity Score using Metadata Analysis

Table 6: Calculate Forum Activity Score using Metadata Analysis

Description	Analyzing the metadata of posts of all forum threads regarding a MOOC and calculating the forum activity score
Sequence of Operations	1. Analyze forum metadata of each forum thread to get a thread specific score 2. Get average of all the scores of all the threads regarding the MOOC which is currently in analysis 3. Normalize the data to compensate for the different data attributes available in each MOOC platform 4. Notify main process that the task is complete
Validity Checks	Check whether all the required data is present and exists in the proper format
Input	Forum Thread Data
Output	Forum Activity Score (Numerical Value)
Error Handling	If task fails it will be considered unprocessed and a log record will be produced so that the maintenance personnel are notified.

3.2.3 Calculate Forum Rating using Sentiment Analysis and Existing Ratings

Table 7: Calculate Forum Rating using Sentiment Analysis and Existing Ratings

Description	Analyzing sentiments of review type of forum thread and using existing ratings to come up with a more accurate rating
Sequence of Operations	1. Analyze sentiments of every post to get a score 2. Get average of all the post scores 3. Use existing rating along with the calculated sentiment score to get a new rating 4. Notify main process that the task is complete
Validity Checks	Check whether all the required data is present and exists in the proper format
Input	Forum Thread Data
Output	Forum Rating (Numerical Value)
Error Handling	If task fails it will be considered unprocessed and a log record will be produced so that the maintenance personnel are notified.

3.3 Performance Requirements

This component should be able to process data from

- Gathering data from a single forum thread related to a MOOC should not take more than 30 seconds
- Analyzing a single forum thread should not take more than 30 seconds
- Normalizing the sub scores should not take more than 10 seconds
- Usage of these new attributes should not add an overhead of more than a single second when recommending MOOCs to learners
- The web crawlers must collect data at least once a day

3.4 Logical Database Requirements

3.4.1 Data Format

The data format used to represent the forum information in the database is JSON. The database used for this purpose is MongoDB.

The logical database requirements include the retention of the following data elements. The below mentioned data structures and lists are not complete and are only meant to be a starting point for development.

Forum Threads

1. MOOC Platform

2. URL
3. Section
4. Type
5. Title
6. Author
7. Description
8. Created Date/Posted Date
9. Posts – Array of objects with the following attributes
    a. Author
    b. Post
    c. Date Posted
10. Is Archived?
11. Last Active Date
12. Is Closed?
13. Upvotes
14. Views

The data records/documents/objects will be uniquely identified using an ID so that it will be linked to the MOOC collection which is separate from this. Multiple records of the above-mentioned collection can be linked with a single MOOC record (Many-to-One relationship) as a single course can have multiple forum threads regarding it.

For the functionality of this component the MOOC collection will have the following attributes in addition to the ones which already exist and from the other components of the research project.

MOOCs

1. Date of last forum analysis
2. Activity Score
3. Rating

3.4.2 Data Accessibility

End users will not have direct access to this data, instead this data will be used to assist the learner's operations.

3.5 Design Constraints

3.5.1 Standards Compliance

The web crawlers should comply to the web crawling policies of the MOOC platforms and the external sites.

3.5.2 Data Constraints

All data must be represented in JSON format so that all modules of the system can interoperate.

3.6 Software System Attributes

3.6.1 Reliability

This component shall fail not more than once a week. This component shall be capable of resolving any data inconsistencies when resuming operations after a failure. Server admin intervention will be required to resume operations in case a failure occurs.

3.6.2 Security

The database used by this component will be secured by a password to prevent unauthorized access. User credentials of active accounts for the MOOC platforms will have to be provided so that the web crawler can use it to gather data which requires the user to be logged in to access.

3.6.3 Maintainability

Functions which are used to traverse the MOOC platforms in the platform specific web crawler might have to be changed when the corresponding MOOC platform page routing structure changes.

3.6.4 Scalability

This component will consist of several sub components which can operate independently of each other. The sub components will be capable of using additional hardware resources (Memory, Storage, etc.) if they are added. Therefore, this component will component will support both horizontal and vertical scaling.

3.6.5 Availability

The back-end system will be operating on a cloud platform which will ensure very high availability (at least 99.99%). The cloud platform will be having multiple instances of the system distributed globally which in turn ensures low latency and redundancy in case an instance fails.

3.7 System Mode

This module will be operational in an automated system mode where human intervention is not required. Several operations will be carried out on a set time period. The user will be required to restart the failed modules if by chance it happens.

4 References

[1] C. Kent, E. Laslo, and S. Rafaeli, “Interactivity in online discussions and learning outcomes,” *Comput. Educ.*, vol. 97, pp. 116–128, 2016.

5 Appendix