Document Manage System with Auto Classification, Tagging

Date: 04.12.2019

Course Name: Cognitive Computing and Deep Neural Networks (INFO7374)

Professor Name: SRI KRISHNAMURTHY

Team members: Wenjun Song, Haimin Zhang, Yanjun Liu

1. Overview

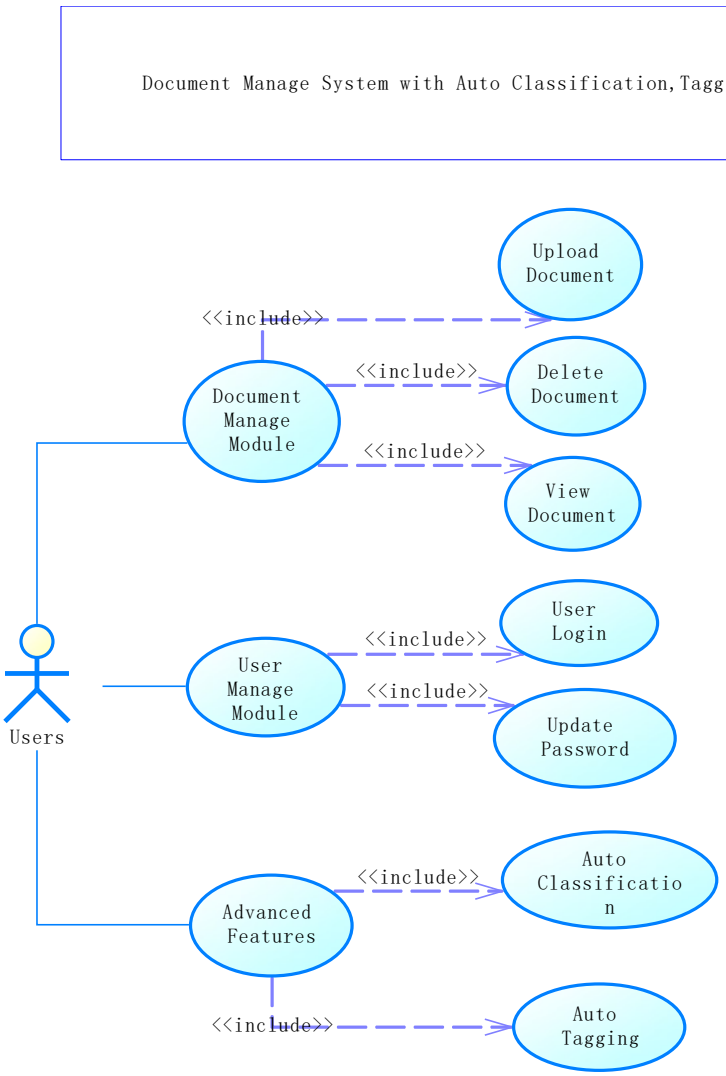
Nowadays, a lot of documents are created every day. By managing these large documents, people have to spend a lot of time on organizing and acting on document every day. Document Classification and Tagging reduces the burden of manual decision making that is done by employees by accurately and automatically organizing information. Document Classification and tagging helps organize unstructured content by analyzing the full text of documents and applying rules that automate classification decisions.

2. Goals

In order to solve problems above and enhance the efficiency of employee, we make these following goals.

1. To offer good algorithms to make good classification and tagging on documents.
2. To offer a simple web application to let people upload documents and manage documents manually
3. To advanced features to help people organize documents with auto classification and tagging.

3. Use Cases



3.1 Document Manage Module

Upload document: Users could upload file into system

Delete document: Users could delete file from system

View document: Users could view document online

3.2 User Manage

User Login: Users could login into CRMS with user name and password

Update Password: Users could update their password

3.3 Advanced Features

Auto Classification: Users could click classify button and documents would be separated into different types of folders.

Auto Tagging: Users could click tagging button and documents would be tagged with related key words.

4. Data

4.1 The 20 Newsgroups data set

The 20 Newsgroups data set is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. The 20 newsgroups collection has become a popular data set for experiments in text applications of machine learning techniques, such as text classification and text clustering. The data is organized into 20 different newsgroups, each corresponding to a different topic. Here is a list of the 20 newsgroups, partitioned (more or less) according to subject matter

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

4.2 AG's corpus of news articles

AG is a collection of more than 1 million news articles. News articles have been gathered from more than 2000 news sources by ComeToMyHead in more than 1 year of activity. The dataset has four categories: World, Business, Sports, SciTech.

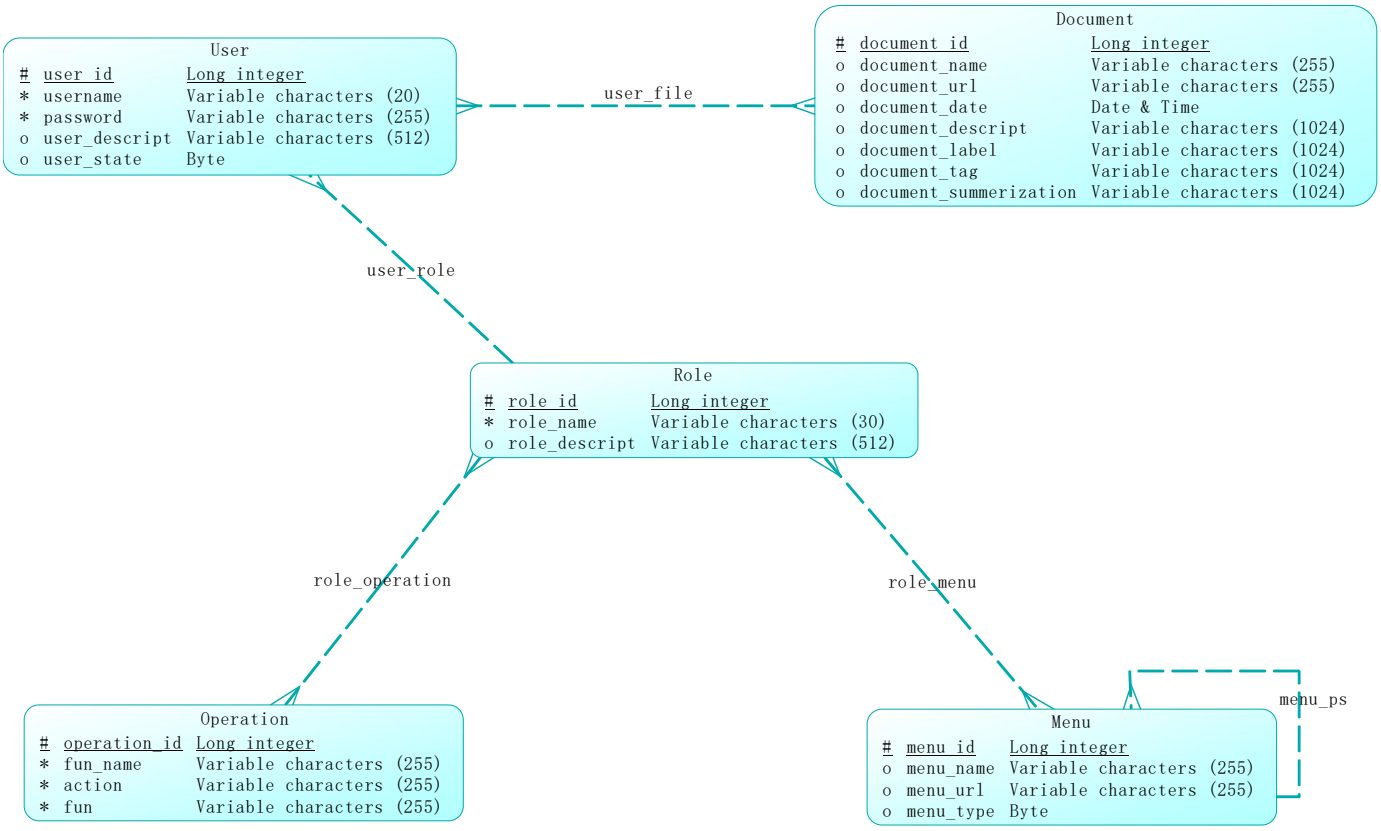
4.3 Reuters-21578 Text Categorization Collection

This is a collection of documents that appeared on Reuters newswire in 1987. The documents were assembled and indexed with categories.

5. Process Outline

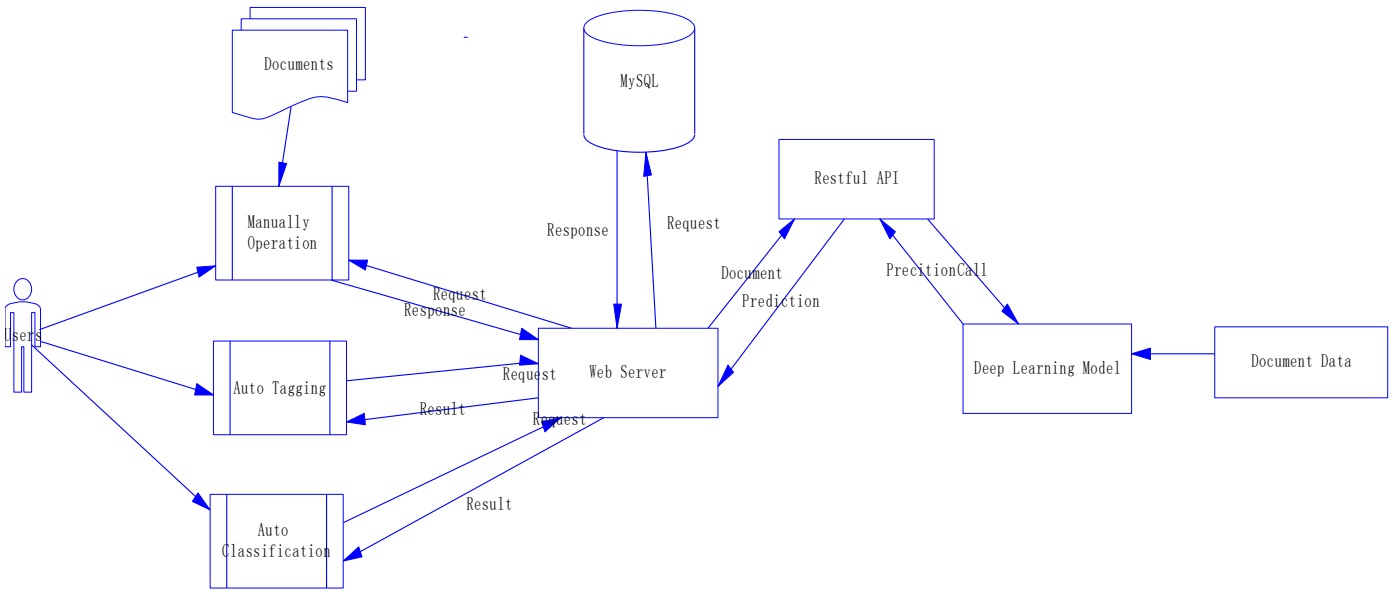
1. Data Preprocessing
2. Collect related papers and make a research on auto classification and tagging
3. Study of Supervised approaches and select the best model for prediction
4. Design of a pipeline and system to implement this approach and discussion on the system's capabilities
5. Build a web application to manually and automatically manage documents

6. Database Design

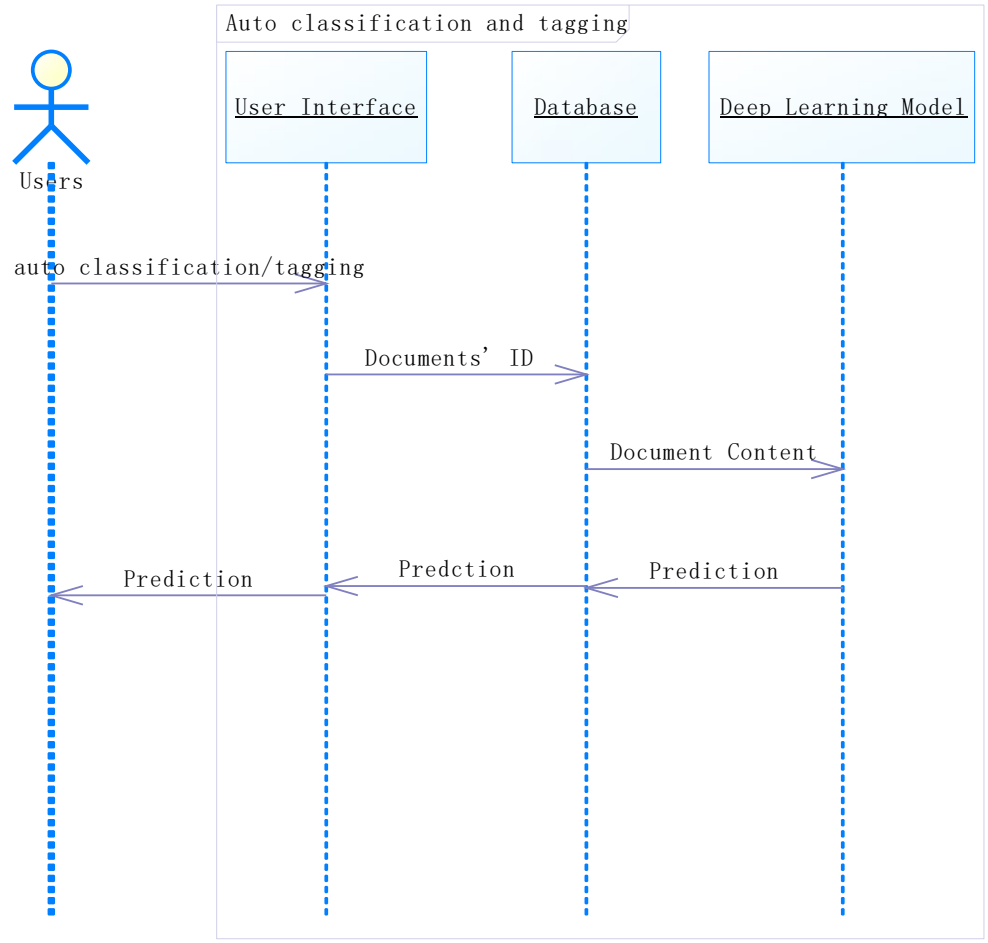


7. System Design

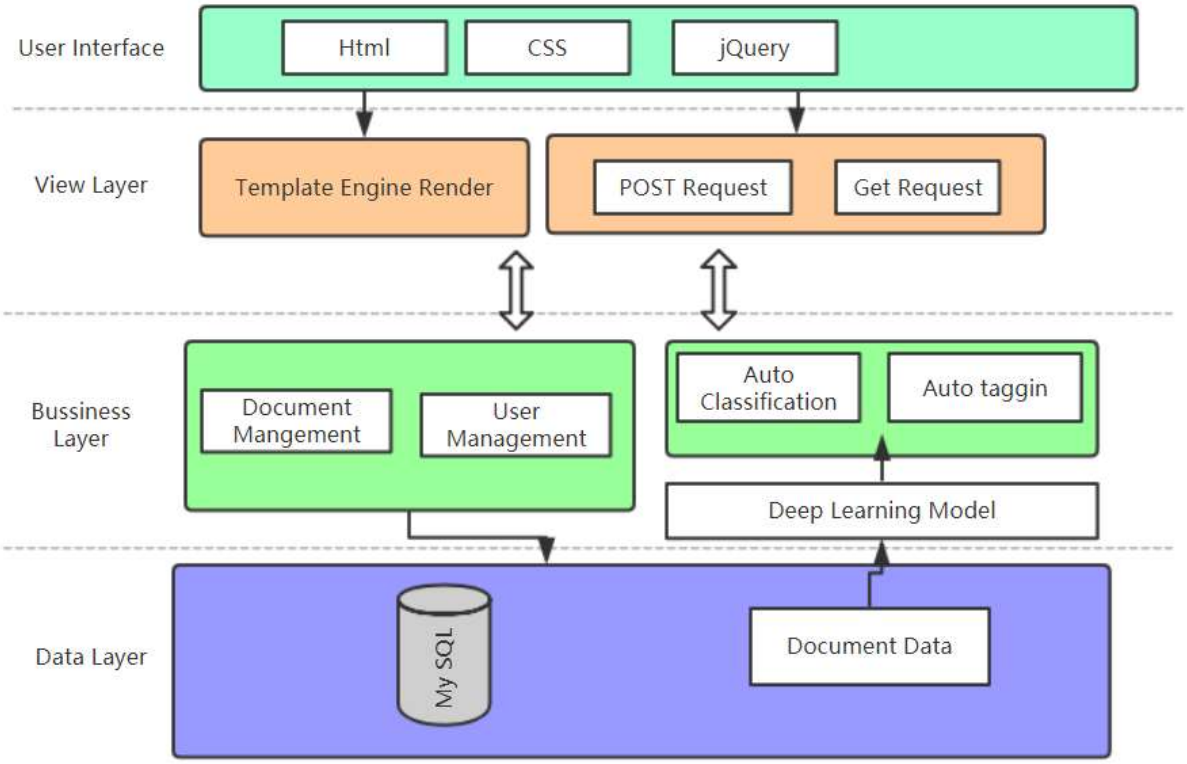
7.1 Workflow



7.2 Sequence Diagram



7.3 System Architecture



8. Milestones

Timeframe	Delivery
Day 1-2 (4.13~4.14)	Data Preprocessing and Exploratory Data Analysis, Read research paper
Day 3-11(4.15~4.24)	A. Model Building, Training, Selection(Wenjun Song Yanjun Liu) B. Deployment of models on cloud and build web application(Haimin Zhang)
Day 12-13 (4.24~4.25)	System integration and documentation

9. Reference and Sources

Datasets:

http://qwone.com/~jason/20Newsgroups/

http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html