International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 03 | Mar 2019 www.irjet.net p-ISSN: 2395-0072

Question-Answer Text Mining using Machine Learning

Abhishek Kumar[1], Kunal Thakur[2], Pranav Nachnekar[3], Nayana Vaity[4]

[1,2,3]BE Student, Computer Engineering, Terna Engineering College, Nerul, Maharashtra, India
[4]Professor, Dept. of Computer Engineering, Terna Engineering College, Nerul, Maharashtra, India

---

Abstract - *Questions and Answers platform are an important part of today's academic world. They enable users to provide, search and search knowledge. However, most such platforms today have a major problem, that of duplication of knowledge. The aim of this project is to avoid this duplication by applying machine learning algorithms to the question, when it is being submitted and help the user find a pre-existing solution to the problem if it already exists on the platform. This helps the user navigate the platform more efficiently and reduces the amount of redundant questions on the platform. We intend to match questions based on the keywords used.*

Key Words: **Recursive Neural Network, Sequence to sequence model, MongoDB, Encoder, Decoder**

1.INTRODUCTION

The age of Internet has changed the way people express their views. It is now done through blog posts, online discussion forums, product review websites etc. People depend upon this user generated content to a great extent. When someone wants to buy a product, they will look up its reviews online before taking a decision. The amount of user generated content is too large for a normal user to analyze. so to automate this, various Machine Learning techniques are used. Symbolic techniques or Knowledge base approach and Machine learning techniques are the two main techniques used in sentiment analysis. Knowledge base approach requires a large database of predefined emotions and an efficient knowledge representation for identifying sentiments. Machine learning approach makes use of a training set to develop a machine learning model that classifies sentiments. Since a predefined database of entire emotions is not required for machine learning approach, it is rather simpler than Knowledge base approach.
Main aim of this project is to implement machine learning based sentiment analysis in a Q&A platform with the following features:
•Easy posting of question and answers
•Rating of questions and answers
•Ranking the questions and answers according to the rating
•Anonymous usage of the platform

2. DRAWBACKS OF EXISTING SYSTEM

The current & previous Question & Answers platforms are having some drawbacks which are as follows:

•Keywords extraction algorithms [1] like TF-IDF and RAKE (Rapid automatic keyword extraction) need large number of dataset to work properly.

•Multiple occurrences of similar type of questions which are having same meanings.

•To remove similar questions from database more manpower is required.

•Existing systems are not suggesting appropriate tags to the user, tags are used to categorize the questions.

3. SUGGESTED IMPROVEMENTS

•We are using Sequence to sequence model for keyword extraction which can extract keywords even from a sentence.

•No extra manpower is required for removing similar questions in our system.

•Appropriate tags are suggesting by analyzing the question by our system

4. METHODOLOGY AND IMPLEMENTATION

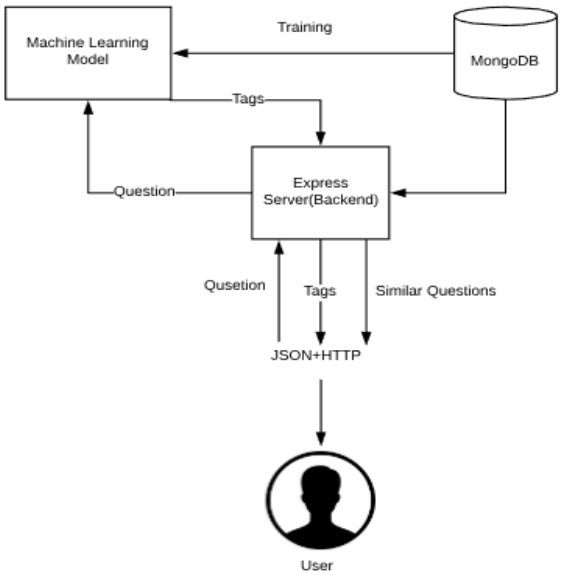

Fig – 1: Architecture of system

4.1 User Registration:

A user can register on the service to post question and answers. Users can also view the question and answers without registering on the service. However, they will not be able to post new questions or answer any previously asked question.

To register a user must provide a username, a password and an email address. The system will verify the user's email and grant them access to the service.

To authenticate user we are using JSON web tokens [2]. The user will first authorize themselves with their password, after which the system will assign them JSON web token. The user will have to include this token in every request to authorize themselves.

To prevent the misuse of tokens, all tokens will expire after 1 year of generation.

4.2 Express Server (Backend):

The Express Server will be used as an intermediate between machine learning model and user. It will handle different tasks such as:

- Authentication
- Registration
- Validation

The end points of this server will adhere to REST (Representational State Transfer) [3] standards. It will allow the user to perform CRUD (Create, Read, Update, Delete) on various resources in the system such as questions and answers.

4.3 Database:

We are using MongoDB for Database operations. We are going to be using NoSQL structure because of the un-uniformity of the data we will be using.

4.4 Machine Learning Model:

We are using RNN [4] in our system. In RNN, the information cycles through a loop. When it makes a decision, it takes into consideration the current input and also what it has learned from the inputs it received previously.

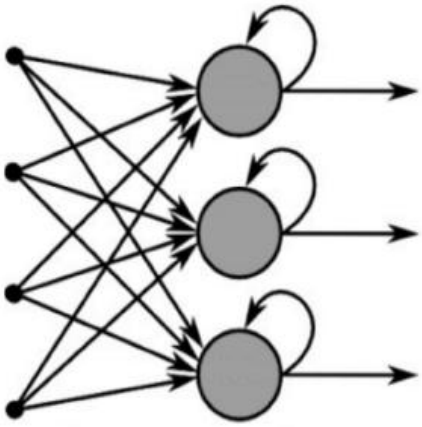

Fig – 2: Recursive Neural Network

For keyword extraction we are using Sequence to Sequence model [5] which comes under RNN.

A Sequence to Sequence network is a model consisting of two separate RNNs called the encoder and decoder. The encoder reads an input sequence one item at a time, and outputs a vector at each step. The final output of the encoder is kept as the context vector. The decoder uses this context vector to produce a sequence of outputs one step at a time.

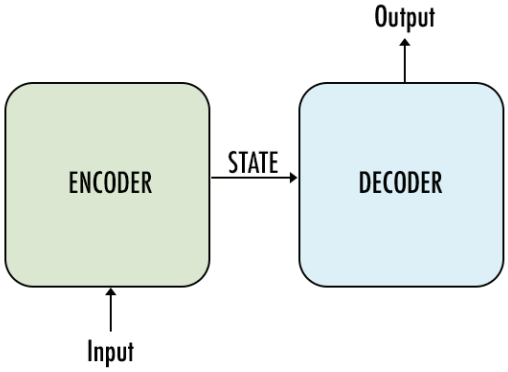

Fig-3: Sequence to Sequence model

4.5 Dataflow:

For posting a question user will enter title of the question and question description, after that question will be sent to backend server, then server will combine title and description of the question and feed it to the machine learning model. Model will provide tags to the user. An object containing question and tags will get save in database. If user is not satisfied from the tags then he can edit that tags and the tags will be updated in the database.

5. OVERALL DESCRIPTION

5.1 System Perspective:

Question and Answer platform is a online platform that enables easier question and answer interaction between the community. It is developed to allow for experts and beginners in a topic to share their knowledge.

It is a web based application that uses a REST API to provide data to the frontend web and mobile applications.

5.2 System Functions:

- Post a new question
- Post answer to a question
- Create, view and modify user profiles
- Find existing questions which are similar
- Search for questions
- Select top answer

5.3 System Features:

- Removing similar type of question

  i. Description:
  This feature aims to remove the redundancy in the platform and help make it easier for users to find answers.

  ii. Priority:
  High

  iii. Response Sequences:
  When a user attempts to ask a new question the question is sent to the backend and relevant tags to that question are returned, system also find questions with similar tags and displays them to the user.

- Achievements

  i. Description:
  To encourage users to answer questions, users are awarded achievements and XP for answering question and other actions.

  ii. Priority:
  Medium

  iii. Response Sequences:
  For example when the user answers a question for the first time they may be awarded with “First Answer” achievement while if they manage to get a certain number of upvotes on their answer they may be awarded with “Good Answer” achievement

6. CONCLUSION

Questions and Answers platform are an important part of today’s academic world. They enable users to provide, search and search knowledge. However, most such platforms today have a major problem, that of duplication of knowledge. Here, we tried to propose a Question and Answer platform based upon on machine learning techniques and minimize the duplication of question in the database.

We have used sequence to sequence model that comes under RNN in machine learning for providing tags to the user. Using that tags, user can categorize his question and check whether the question is already present in database or not. If similar type of question is present in our system then it will suggest all of the similar questions and if not, then user can post his question.

In the current system, they are not checking whether a newly asked question is previously asked or not and nothing about the tags as well.

As part of the future work, we can do reverse also i.e. we can form a question by having keywords only, implement common features from other similar platforms such as personalize question feeds. We can generalize the model to support other platforms for example making thumbnails on videos based on the title.

REFERENCES

[1] Slobodan Beliga University of Rijeka, Department of Informatics Radmile Matejčić 2, 51 000 Rijeka, Croatia
[2] JSON Web Token (JWT) Profile for OAuth 2.0 Client Authentication and Authorization Grants, M. Jones, B. Campbell, C. Mortimore, May 2015
[3] Hypertext Transfer Protocol (HTTP/1.1): Semantics and Content, R. Fielding, J. Reschke
[4] A Critical Review of Recurrent Neural Networks for Sequence Learning by Zachary C. Lipton, John Berkowitz
[5] Sequence to Sequence Learning with Neural Networks by Ilya Sutskever, Oriol Vinyals, Quoc V. Le