Team Name

Team TechTorch

Team Members

Shantanu Digambar Dahitule

Deshansh Panchbhai

Moudipa Jana

Problem Statement:

Autonomous-Tagging-Of-Stack-Overflow-Questions

Skills Required:

Python,Machine Learning,Flask Integration,Numpy,Pandas, IBM Watson

Project Description:

Information sharing platforms have become very popular platform for question and answer sessions. Examples include Quora, StackOverflow, Reddit and OpenEDX. While the quantity of information available on these websites has increased many folds but there is no efficient way to classify data as such that is automated. Most such websites ask users to tag their queries which is not an intuitive way to ask questions. As users might not tag the problem properly which further leads to ambiguity in data. It would be useful to automate the process of tagging as the means to classify information in an efficient manner. A system that supports sovereign tagging can improve the user experience by clustering information into discrete common topics. The other benefit is that the user can be recommended queries related to his own problem which could help him find the answer in an efficient and effective manner. This project outlines a method for question and answer platforms that automatically allocates tags for a given query.

Purpose

1) It will improve the search results . User can easily search the topic they are interested in and system will return the relevant tag associated with the post.
2) For both users and moderators, the system saves time. It can take a lot of time to manually tag each question or piece of material, especially on platforms with plenty of submissions.
3) User engagement and participation are increased as a result of the use of tags to help users find questions and subjects that are relevant to their knowledge or interests.
4) Tagging correctly can lessen the possibility of repeated questions being asked. When tags appropriately describe material, users have a higher chance of finding already-existing solutions

Solution:

The solution contains three parts:
1) Machine learning model
2) IBM Cloud deployment
3) User Interface

Machine learning Model:

1. Automated Tagging Approach:
While manual tagging has been the norm for categorizing content on platforms like Stack Overflow, your solution introduces automation to the tagging process. This represents a significant shift in how tags are assigned, making the process faster, more efficient, and less error-prone.

2. Machine Learning for Tagging:
Leveraging machine learning algorithms like Logistic Regression and Linear SVC to predict tags based on question text and contextual information is a novel approach. It takes advantage of the algorithms' ability to learn patterns and relationships from large datasets, leading to accurate tag assignments.

3. One-vs-All Strategy:
Treating each tag as a separate binary classification task (one-vs-all strategy) is an innovative approach to handling the multi-class tagging problem. This technique allows the system to focus on learning the specific characteristics of each tag, leading to more accurate

predictions.

4. Scalability and Real-time Tagging:
The automated tagging system is designed to scale seamlessly as the platform grows. The ability to tag new questions in real-time as they are posted is a unique feature that ensures questions are categorized accurately and promptly.

5. User Experience Enhancement:
The solution directly contributes to improving the user experience on Stack Overflow. By accurately tagging questions, developers can find relevant information more easily, accelerating problem-solving and fostering a positive community atmosphere.

6. Consistency and Organization:
Achieving consistent and standardized tagging across the platform is a novel aspect. Automated tagging ensures that tags are applied uniformly, enhancing the organization of content and making it easier for users to discover related topics.

7. Combination of Algorithms:
The combination of Logistic Regression and Linear SVC provides a versatile approach to solving the tagging challenge. Each algorithm has its strengths, and utilizing both can enhance the overall tagging accuracy.

8. Adaptation to Platform Evolution:
As the platform's content and user demeanour evolve, the machine learning models can be retrained to adapt to changing patterns and trends. This adaptability ensures that the tagging system remains effective over time.
You've chosen an efficient and well-rounded technology stack for your autonomous tagging system:

IBM WATSON CLOUD & Tech Stack

1. IBM Watson Cloud Infrastructure
The solution contains the API of machine learning model which is deployed on the IBM Watson Cloud for the remote access this API is more faster responsive than the local server.
2. Flask:
Lightweight Python web framework for building APIs.
3. Scikit-learn:
Machine learning library for training and testing models.
4. Pandas and NumPy:
Libraries for preprocessing and handling data.
5. NLTK:

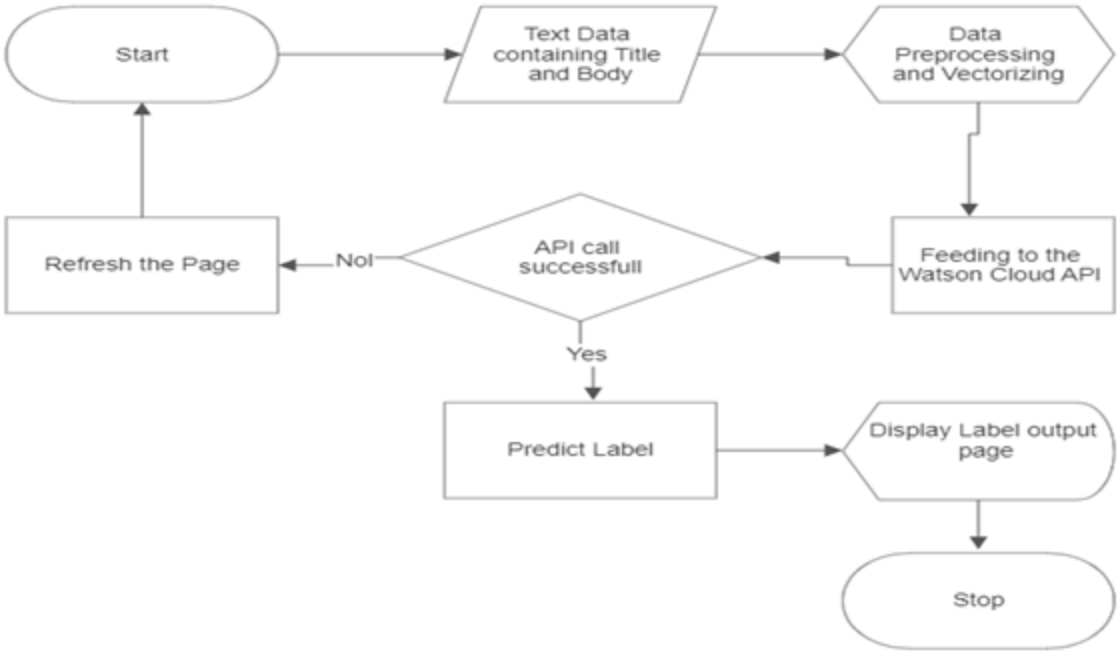
To preprocess the text data and Vectorize.

This stack covers everything from API development, machine learning model training, data preprocessing, and prediction. It will enable you to implement the autonomous tagging system for Stack Overflow questions efficiently.

User Interface

For this project we decided to keep user interface as similar as stackOverflow to keep the things conversant and to show the real world use case of our API in stackOverflow. So we can feel the users prospective for our solution.

FLOWCHART



ADVANTAGES AND DIS-ADVANTAGES

Advantages

a. **It will improves searchability and content structure, allowing users to locate pertinent information more quickly.**

b. **It will streamline the content categorization, saving users and moderators valuable time.**

c. **By leading users to customized material, the system's capacity to accurately propose tags decreases duplicate content and increases user engagement.**

d. **Data analysis supported by the system's effective tagging procedure enables administrators to spot patterns and important subjects.**

e. **This solution fosters a dynamic and user-friendly environment by optimizing user experience, content discovery, and platform management.**

f. **We have done the evaluation of the model based on F1- score , so based on the F1-score we got , we can say that our model is not overfitting so we can say that our model is more robust compared to other any models.**

Disadvantages

a. **If the training data is biased, incomplete, or unrepresentative of the full range of Stack Overflow content, the model's accuracy and effectiveness may be compromised.**

b. **Technical content often contains terms with multiple meanings or terms that can be used in different contexts. NLP models might struggle to accurately disambiguate such terms and assign the correct tags.**

c. **There is a lot of technical jargon and domain-specific vocabulary used in Stack Overflow conversations. Such specialized content may be difficult for NLP models with general language corpus training to comprehend and appropriately tag.**

d. **If the training data used to build the model is biased towards certain topics or perspectives, the model might reproduce and amplify these biases in its tagging suggestions.**

APPLICATIONS

a. **An autonomous tagging system can assist organize and index documents in enterprises and organizations, making it simpler for staff members to access the information they require.**

b. **Automated tagging can help social media networks locate relevant material for users based on their interests and increase the discoverability of content.**

c. **Automated tagging can be used by online learning systems to categorize and suggest instructional materials, tasks, and discussion threads based on their subject matter and applicability.**

d. **An autonomous tagging system can help platforms that house user-generated material, such as blogs, forums, and knowledge bases, organize and categorize content for simpler search and navigation.**