**IBM Hack Challenge 2023**

Tag Suggest:Automated System for Predicting Tags in StackOverflow

Team name : tackleAI

Project ID : SPS\_PRO\_3991

Project Title : Autonomous-Tagging-Of-Stack-Overflow-Questions

Team members

Shriyash Gulhane: shriswimmer@gmail.com

Hajna Saik: hajnashaik0706@gmail.com

Amritha R: ar9011@srmist.edu.in

INTRODUCTION

PROJECT OVERVIEW

To begin, platforms for exchanging information and doing Q&A sessions are becoming increasingly popular. Quora, Stack Overflow, Reddit are a few of the many examples. Tags play a crucial role in organizing and categorizing questions, enabling users to search and find relevant information effectively. Currently, tags on Stack Overflow are assigned manually by users, which can be time-consuming and inconsistent. Moreover, as the platform grows, the number of questions increases rapidly, making it challenging to keep up with the tagging process. In order to address these challenges and enhance the user experience, the objective is to create an automated solution using machine learning to accurately predict and assign appropriate tags to Stack Overflow questions.

PURPOSE

The purpose of Stack Overflow is to serve as an online community and knowledge-sharing platform for programmers and developers. It was created to provide a platform where developers can ask questions, share knowledge, and collaborate with their peers in solving programming-related problems.

Stack Overflow aims to create a vast repository of programming knowledge. Questions and answers posted on the platform become accessible to anyone who may encounter similar problems in the future, helping them find solutions and learn from the collective expertise of the community.

Users can follow specific tags or topics of interest and participate in discussions through comments, votes, and edits. Developers often encounter challenges and errors while working on projects.

1. LITERATURE SURVEY

EXISTING PROBLEM

Question and answer sessions are now frequently held on information-sharing platforms. Reddit, StackOverflow, Quora, and OpenEDX are a few examples. Although the amount of information on these websites has multiplied, there is no effective, automatic method for classifying data as such. The majority of these websites need users to tag their inquiries, which is not a natural way to do so.

PROPOSED SOLUTION

Here we aim to develop an autonomous system that can accurately and efficiently assign relevant tags to Stack Overflow questions without manual intervention. The system should take into account the question text and other contextual information to identify the most relevant tags. By leveraging machine learning techniques, the autonomous tagging system will be able to accurately predict relevant tags for Stack Overflow questions. This automated solution will enhance the organization and searchability of questions, facilitating a smoother user experience for developers seeking information on the platform.

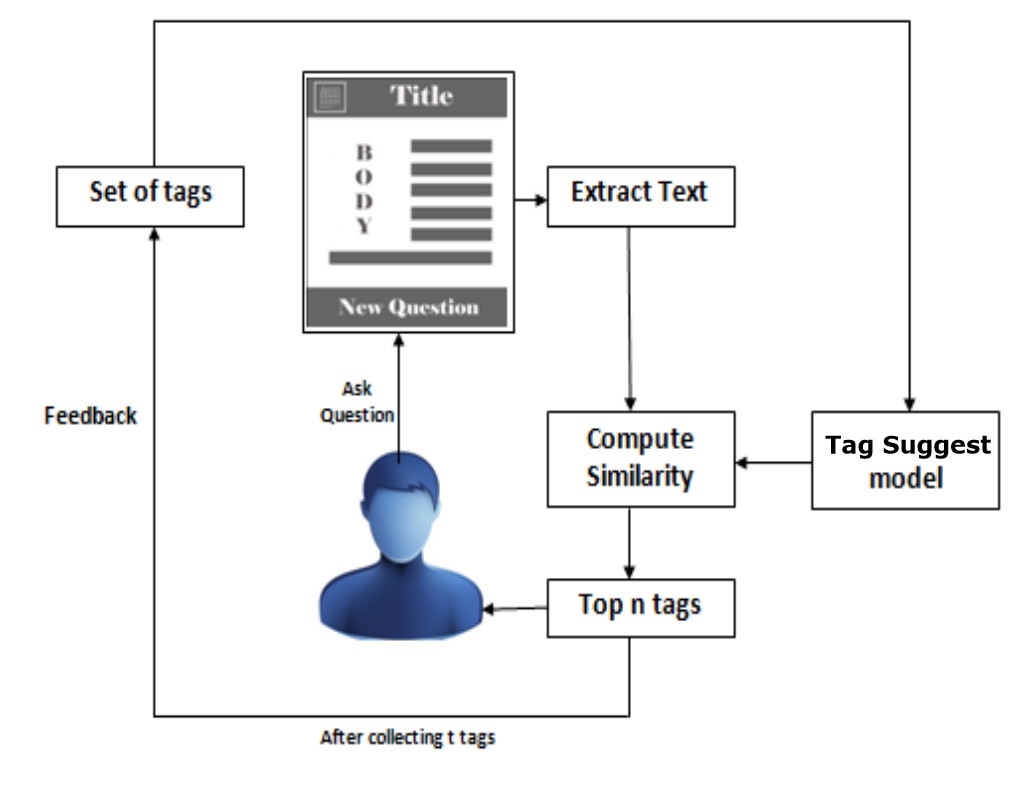
THEORETICAL ANALYSIS

**BLOCK DIAGRAMS:**

SOLUTION ARCHITECTURE DIAGRAM:

This solution architecture outlines a possible approach to building an automated system for predicting tags in StackOverflow. It leverages machine learning techniques and natural language processing to suggest relevant tags for questions based on their content.

1. Set of tags: The first step is to define a set of tags that the system can use to suggest tags for questions. This can be done by analyzing the most commonly used tags in StackOverflow and selecting the most relevant ones for the system.
2. Extract Title: The title of the question is extracted and used as input for the next step.
3. Extract Text: The text from the title is extracted and preprocessed to remove any irrelevant information such as stop words, punctuation, etc.
4. Compute Similarity: The similarity between the preprocessed text and the set of tags is computed using a similarity measure such as cosine similarity or Jaccard similarity.
5. Top n tags: The top n tags with the highest similarity scores are selected as the suggested tags for the question.
6. Tag Suggest model: A machine learning model is trained on historical data to predict the most relevant tags for a given question. The model takes as input the preprocessed text and outputs a list of suggested tags.
7. Feedback: The system collects feedback from users on the accuracy of the suggested tags and uses this feedback to improve the performance of the tag suggestion model.



HARDWARE AND SOFTWARE REQUIREMENTS:

1. HARDWARE REQUIREMENTS:

1) **Real time detection** - A method of detecting stack overflows is to create a canary space at the end of each task. This space is filled with some known data. If this data is ever modified, then the application has written past the end of the stack.

2) **Accuracy** - The content quality of shared knowledge in Stack Overflow (SO) is crucial in supporting software developers with their programming problems.

3) **Data storage and analysis** - Stack Overflow helps people find the answers they need, when they need them. We're best known for our public Q&A platform that over 100 million people visit every month to ask questions, learn, and share technical knowledge. We are implementing it thorough the IBM cloud platform.

1. SOFTWARE REQUIREMENTS:

**Python**: A versatile programming language that is widely used in the field of data science and machine learning. Other languages such as HTML, CSS, JavaScript

**Data collection**: Stack Overflow question data will be collected and pre-processed to obtain a clean and structured dataset for training the machine learning models.

**Feature extraction**: Various features, such as question text, keywords, and syntax, will be extracted to represent the input data in a numerical format suitable for machine learning algorithms.

**Machine learning model development**: A combination of supervised learning techniques will be employed to train models that can accurately predict relevant tags based on the input features.

**Model evaluation and refinement**: The trained models will be evaluated using appropriate metrics and refined to improve their performance and accuracy.

**Web Frameworks- Flask**: These frameworks provide a foundation for developing web applications and will assist in building the user interface for our autonomous tagging system.

**Deployment on IBM Watson Studio**: Deploying the final system on IBM Watson Studio for seamless integration with the Stack Overflow platform.

1. EXPERIMENTAL INVESTIGATIONS

### Dataset Selection: For the experimental investigations, a diverse and representative dataset of user-generated questions was used. The dataset includes questions from various domains, each associated with a set of actual tags. This dataset was used to train and test the machine learning model for tag prediction.

Dataset Used: https://www.kaggle.com/stackoverflow/stacksample

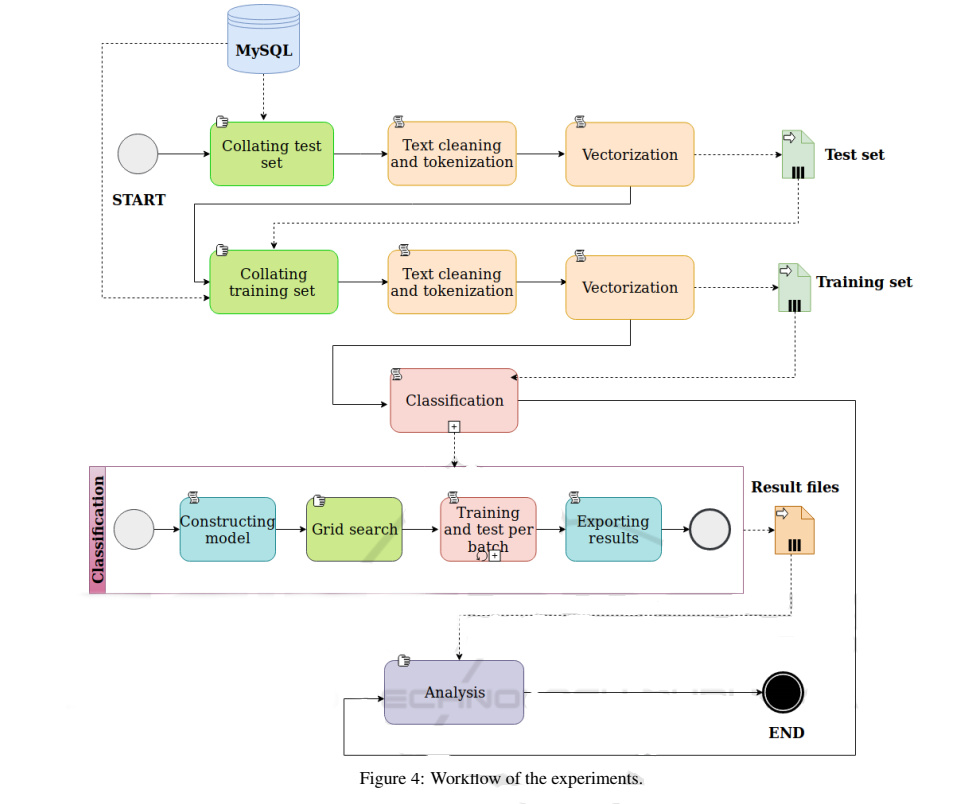
### Machine Learning Model Evaluation

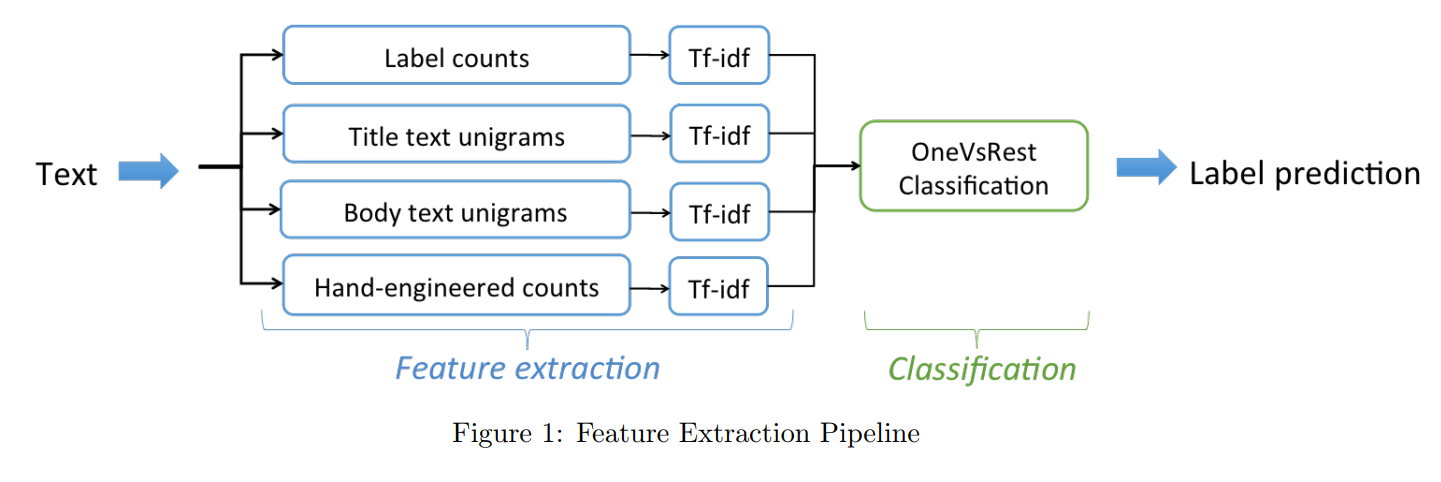
The trained machine learning model, specifically a Linear Support Vector Machine (LinearSVM), was evaluated using standard evaluation metrics. These metrics include precision, recall, F1-score, and accuracy. The model's performance was assessed on both training and testing datasets to ensure generalizability.

### Stack Overflow API Integration

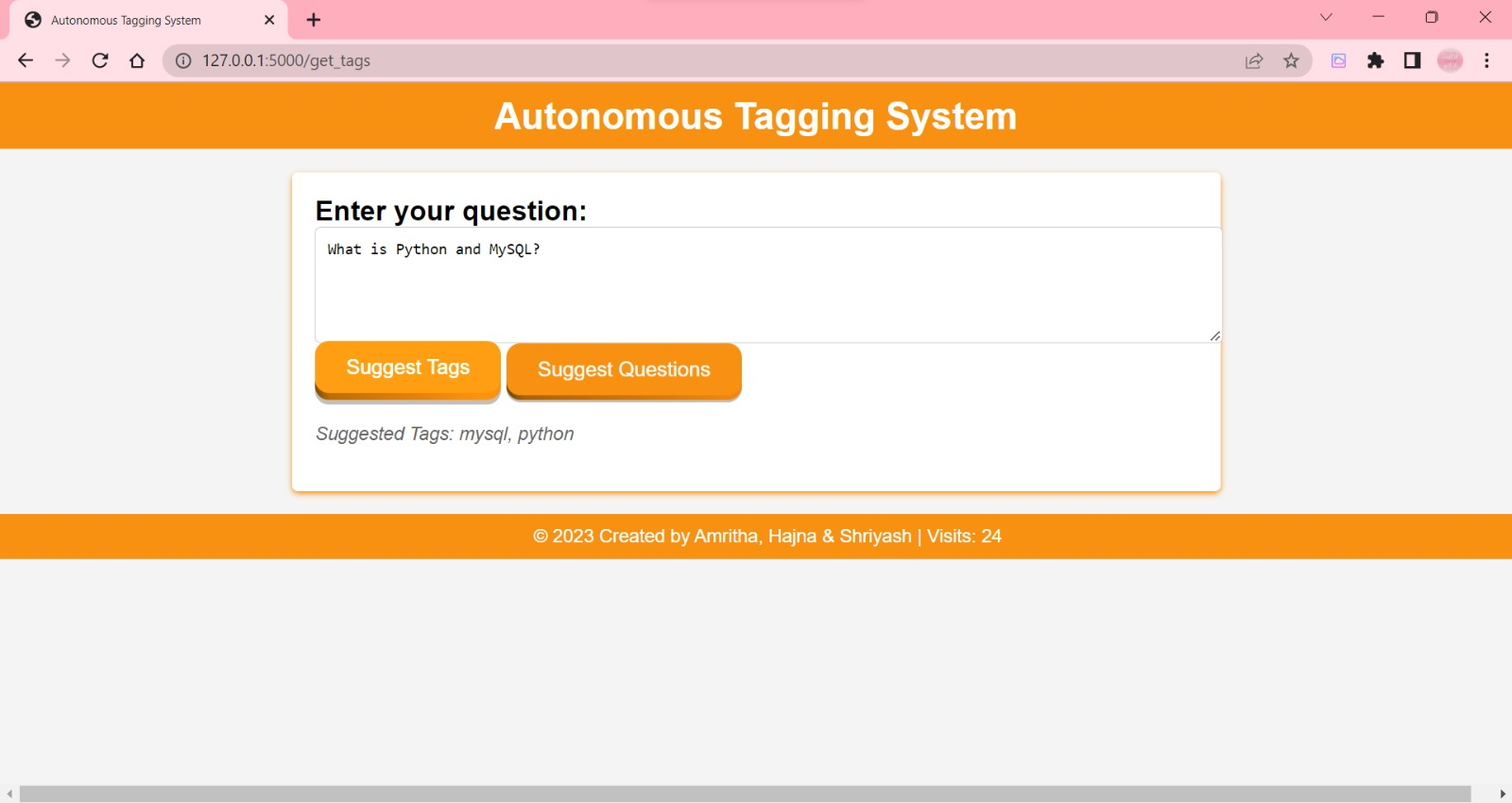
The integration of the Stack Overflow API was tested by querying the API to retrieve relevant questions and tags. Specifically, the system was tested to ensure that it could effectively search for frequently asked questions (FAQs) by utilizing the 'suggest tags’ and ‘suggest questions’ API functions.

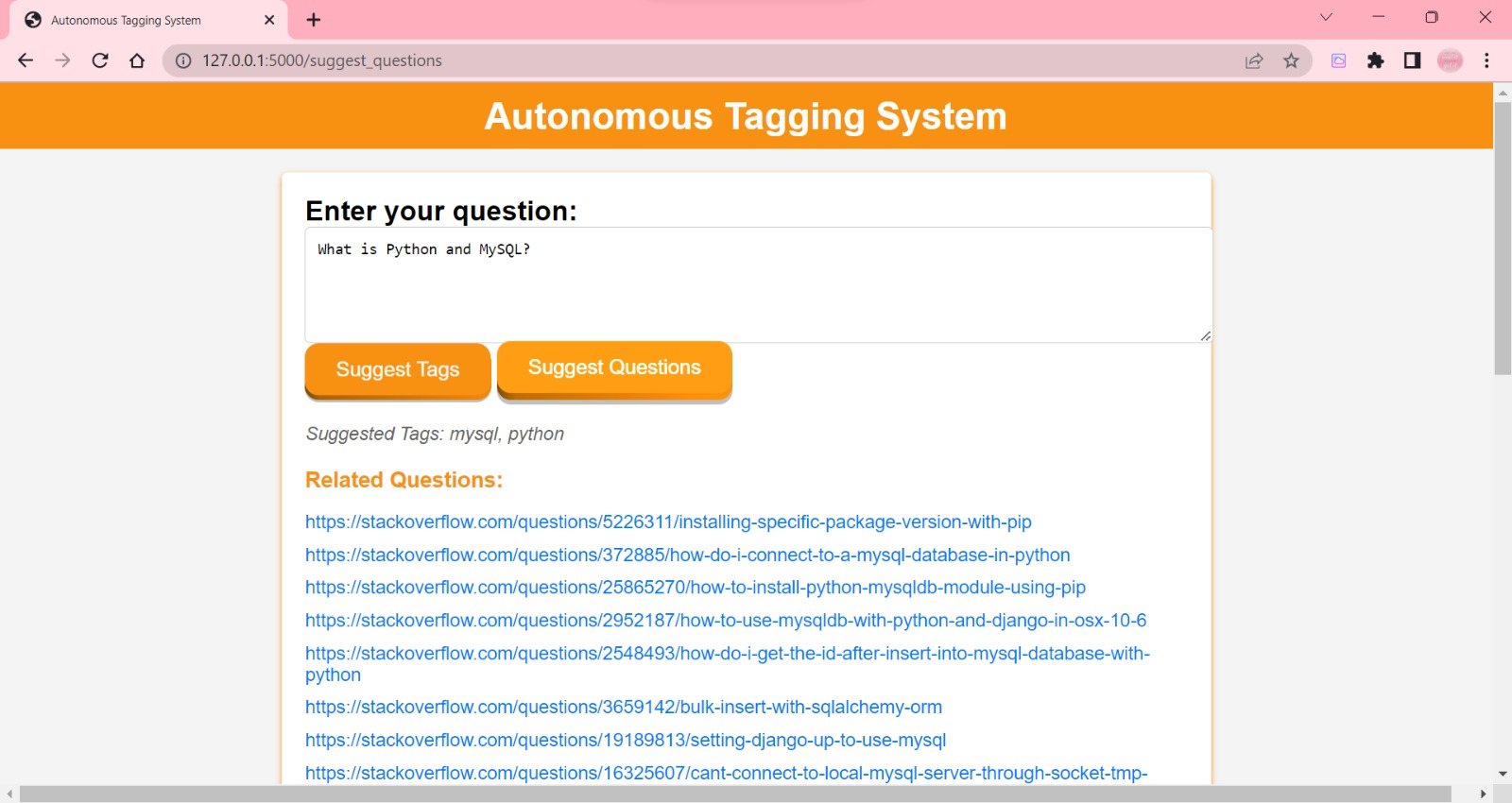
1. FLOWCHART



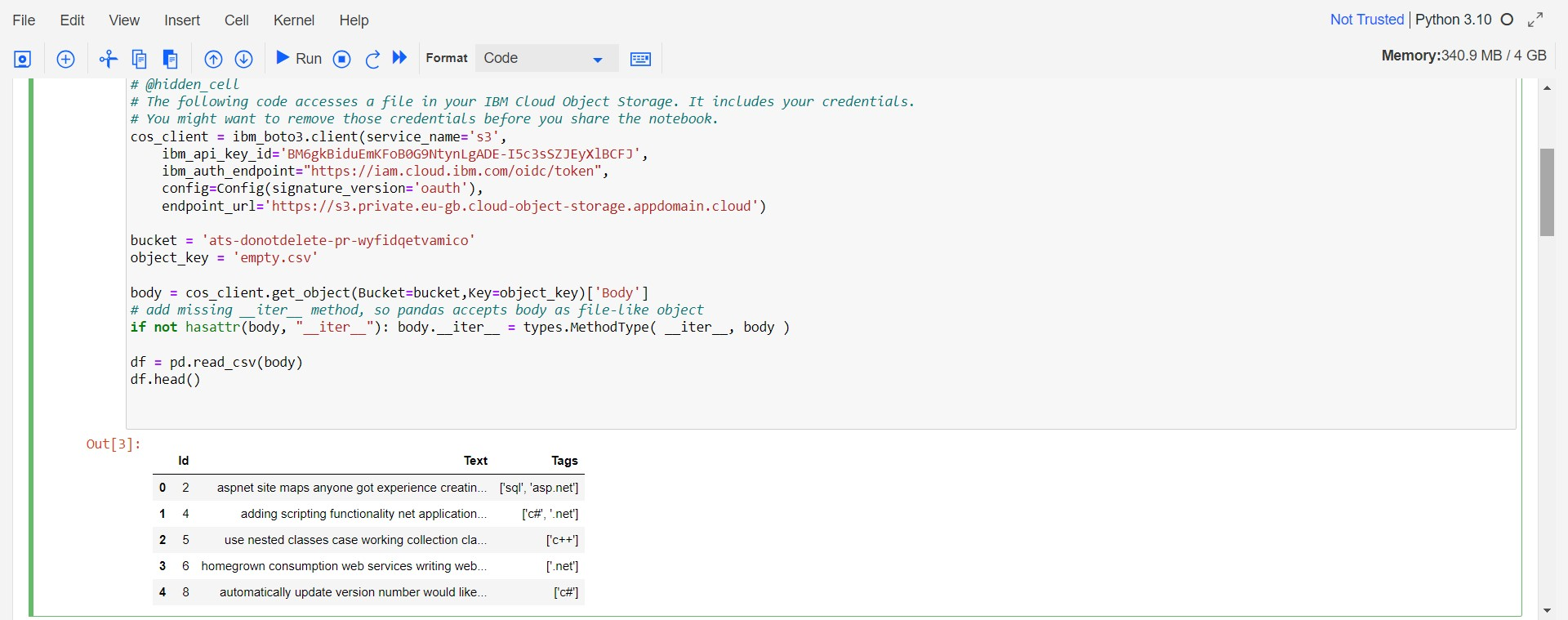


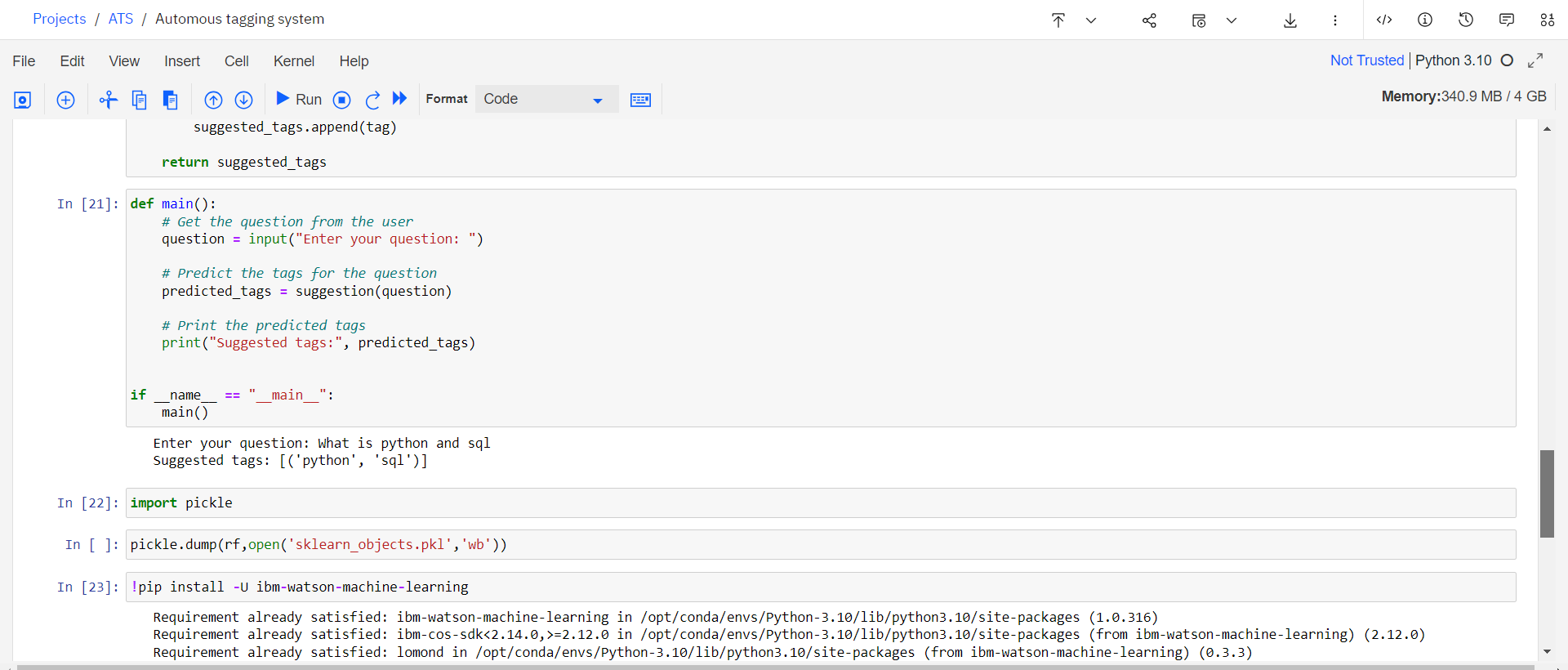
1. RESULT

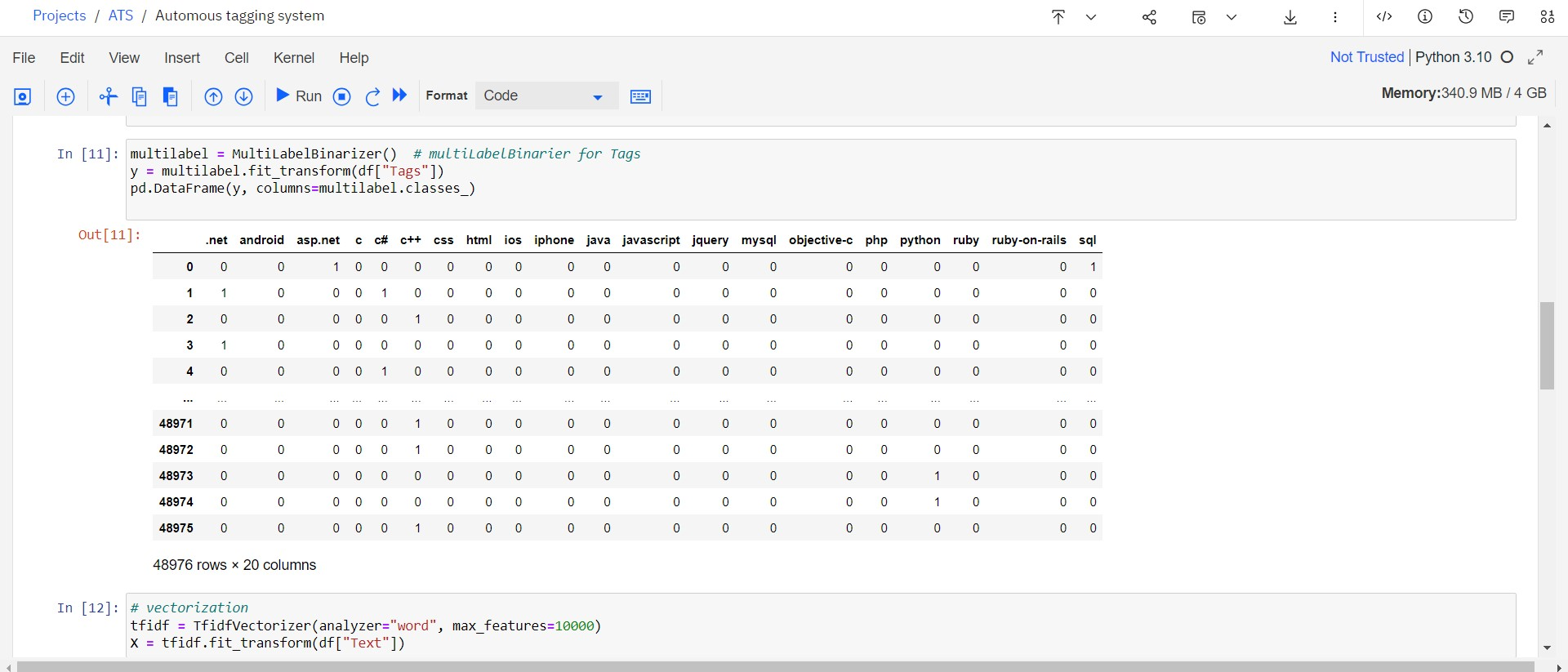




IBM Watson







ADVANTAGES:

1. Improved Tag Accuracy: Autonomous tagging leverages machine learning and natural language processing techniques to analyze the content of questions and assign appropriate tags.
2. Time and Effort Saving: Autonomous tagging automates the process of assigning tags to questions, eliminating the need for manual tagging by users.
3. Enhanced Discoverability: Accurate and relevant tagging improves the discover-ability of questions.
4. Reduced Duplicate Questions: Autonomous tagging can help identify similar or duplicate questions by analyzing their content and assigned tags.
5. Enriched User Engagement: By providing accurate tags, autonomous tagging encourages experts and users with relevant knowledge to discover questions within their areas of expertise.

DISADVANTAGES:

1. Tagging Errors: Autonomous tagging systems may occasionally assign incorrect or irrelevant tags to questions.
2. Lack of Human Interpretation: Autonomous tagging lacks the human judgment and interpretation that manual tagging provides.
3. Challenges with Ambiguous or Uncommon Topics: Autonomous tagging can struggle with questions related to ambiguous or less common topics.
4. Inability to Capture Changing Language Trends: Programming languages, frameworks, and technologies evolve over time, and new ones emerge.
5. Over Reliance on Tagged Questions: Autonomous tagging systems heavily rely on previously tagged questions for training.

APPLICATIONS

Autonomous Tagging System has a wide range of applications across various domains due to its ability to automatically suggest relevant tags for user-generated content. Here are some areas where this system can be applied:

1. **Content Management Systems** (CMS): The system can be integrated into CMS platforms, helping content creators and editors assign accurate tags to articles, blog posts, and other content, leading to better organization and discover-ability.
2. **Online Forums and Communities**: In discussion forums and community platforms, the system can assist users in tagging their questions or posts appropriately, ensuring that the content reaches the right audience and is easier to find.
3. **E-commerce Platforms**: On e-commerce websites, the system can automatically suggest product tags based on product descriptions, enhancing the search and filtering capabilities for customers.
4. **Educational Portals**: The system can aid educators in tagging educational resources, such as lecture notes, videos, and quizzes, making it simpler for students to locate relevant materials.
5. **Programming and Development Communities**: In platforms focused on programming and development, the system can assist developers in accurately tagging their questions, making it easier for other developers to find and offer solutions.

CONCLUSION

In conclusion, autonomous tagging of Stack Overflow questions offers several advantages and brings efficiency to the platform. It improves tag accuracy, saves time and effort, enhances discover-ability, reduces duplicate questions, fosters user engagement, allows for continuous improvement, and ensures scalability.

However, there are also potential disadvantages to consider, such as tagging errors, the lack of human interpretation, challenges with ambiguous or uncommon topics, difficulty capturing changing language trends.

These drawbacks can be addressed through continuous refinement, user feedback, and a combination of autonomous tagging with human moderation.

By striking the right balance between automation and human intervention, Stack Overflow can leverage autonomous tagging to improve the user experience, facilitate knowledge sharing, and promote effective collaboration within the programming community.

FUTURE SCOPE

The future scope for autonomous tagging of Stack Overflow involves advancements and enhancements in several areas. Here are some potential future directions:

1. **Improved Tagging Models**: Continued research and development in natural language processing (NLP) and machine learning can lead to more advanced tagging models.
2. **Integration of User Feedback**: Leveraging user feedback and interactions to improve the tagging system is crucial.
3. **Enhanced Topic Coverage**: Autonomous tagging can expand its coverage to include a broader range of programming languages, frameworks, and emerging technologies.
4. **Customization and Personalization**: Providing users with the ability to customize and personalize the tagging system can further enhance the user experience.

The future of autonomous tagging on Stack Overflow lies in a combination of advanced machine learning techniques, by harnessing these advancements, Stack Overflow can enhance the accuracy, efficiency, and overall user experience of its autonomous tagging system.

BIBLIOGRAPHY

https://ibm.github.io/watson-machine-learning-sdk/

https://api.stackexchange.com/docs

https://www.kaggle.com/datasets/stackoverflow/stacksample

APPENDIX

# predicting the model with samples

def suggestion(question):

x=[]

x.append(question)

xt=tfidf.transform(x)

answ = multilabel.inverse\_transform(clf.predict(xt))

suggested\_tags = [ ]

for tags in answ:

for tag in tags:

suggested\_tags.append(tag)

return suggested\_tags

def main():

# Get the question from the user

question = input("Enter your question: ")

# Predict the tags for the question

predicted\_tags = suggestion(question)

# Print the predicted tags

print("Suggested tags:", predicted\_tags)

# Function to search Stack Overflow and get relevant question links

def search\_stack\_overflow(tags):

base\_url = "https://api.stackexchange.com/2.3/search/advanced"

params = {

"order": "desc",

"sort": "votes",

"q": " ".join(tags),

"site": "stackoverflow"

}

response = requests.get(base\_url, params=params)

data = response.json()

question\_links = []

for item in data.get("items", []):

question\_links.append(item["link"])

return question\_links

# Function to update visit count in the session

def update\_visit\_count(session):

# Initialize visit count if not present in session

if 'visit\_count' not in session:

session['visit\_count'] = 0

# Increment visit count

session['visit\_count'] += 1

if \_\_name\_\_ == "\_\_main\_\_":

main()