Tag suggestion system using multi-label text categorization

A report submitted for the J component of

Submitted By

Abhiram Sreekantham - 19BCE2294

NATURAL LANGUAGE PROCESSING
(CSE4022)

Slot - B2

Project Guide

Prof. Rajeshkannan R

School of Computer Science and Engineering

BACHELOR OF TECHNOLOGY
in
COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

Fall Semester, 2021–2022

1. Abstract

Text analysis is a relatively recent field of study. Marketing, product management, academia, and governance are among the domains where the process of analysing and extracting information from textual data is already in use. Text classification or text categorization is the process of assigning appropriate categories to natural language documents from a pre-defined set. In layman's terms, text categorization is the process of extracting generic tags from unstructured textual material. A list of pre-defined groups is used to generate all generic tags. If you organize your content and things into categories, users will find it easier to discover and navigate your website or application.

In this project, we'll be working on a text classification model that analyses a question's textual description and predicts multiple labels for it. We'll develop a multilabel text classification approach for a tag recommendation system using Multi-Label Text Classification, which is a subset of multiple output models.

2. Introduction

Text analysis is a new topic of research in general. The method of evaluating and extracting information from textual data is already being used in fields including marketing, product management, academia, and governance. The action of classifying natural language documents with applicable categories from a pre-set collection is known as text classification or text categorization. Text categorization, in layman's terms, is the technique of obtaining generalized tags from unstructured textual data. All generic tags are drawn from a list of pre-defined groups. Users will find it easier to explore and browse your website or application if you organise your information and items into categories.

One sample can belong to many classes in Multi-Label Text Classification (MLTC). Most MLTC tasks have dependencies or correlations among labels, as observed. Existing techniques frequently overlook the link between labels.

We'll utilise the dataset "StackSample:10% of Stack Overflow Q&A." It's a multilabel text classification algorithm's issue statement. We'll be working on a text classification model which looks at a textual description of a question and predicts numerous labels for it. We'll use Multi-Label Text Classification to create a multilabel text classification method for a tag recommendation system, which is a subset of multiple output models. The text data undergoes text pre-processing, and the cleaned data is loaded for text classification. We'll use text Vectorization on the data, MultilabelBinarizer to encode the tag labels, and Classical classifiers like SGC, MultiNomial Naive Bayes Classifier, Random Forest Classifier, and others to model and compare the results.

Classification is a sort of supervised learning in machine learning. A classification issue is a predictive modelling problem in which a class label for a given input sample is anticipated. It identifies the data point's class and is most useful when the output has both finite and discrete values.

There are four different forms of categorisation:

- ➔ Binary classification
- ➔ Multiclass classification
- ➔ MultiLabel classification
- ➔ Imbalanced classification

When we apply multi-label text classification, such as Keras multi-label text classification, we may improve the algorithm's accuracy. With XLNET and GPT-2 and GPT-3, you can also employ multi-label text categorization.

3. Literature review

Authors and year	***Title (Study)***	***Concept / Theoretical model/ Framework***	***Methodology used/ Implementation***	***Relevant Finding***
Guibin Chen1, Deheng Ye1, Zhenchang Xing2, Jieshan Chen3, Erik Cambria. (2017)	**Ensemble Application of Convolutional and Recurrent Neural Networks for Multi-label Text Categorization**	Multi-label text categorization refers to the task of assigning one or multiple categories (or labels) to a textual document, Multi-label text categorization can generally be divided into two sub-tasks: text feature extraction and multi-label classification. A. Text features Instead of representing the whole text in one step, e.g., tf-idf weighting for a document, most recent works focus on the distributional representation of individual words, which is nevertheless pre-processed and tokenized from the original raw text data. B. Multi-label classification Since text features can be well represented as a feature vector, the next step is to do multi-label classification for such features. Each input instance is a high dimensional feature vector X ∈ Rm, which is assigned to a subset y of the label space Y with L possible labels. The task is to learn a function from the training data.	Our approach consists of two parts: the CNN part for extracting text features and the RNN part for multi-label prediction. CNN is used to extract a global fix-length feature vector for the input text. Then, with these feature vectors, the “initial state” or prior knowledge of the RNN is determined and used to predict a sequence of labels. Two steps of trainings are performed. The first training step is to train a word2vec model by using all the tokenized texts as unlabelled data. Then, this word2vec model is fed into the second training step, that is the supervised training for the CNN-RNN model. Using the softmax classifier as the upper layer of LSTM for labels prediction, the cross-entropy loss is then back-propagated from RNN down to CNN to update the weights of both the CNN-RNN model. One state-of-the art update policy Adam is chosen instead of stochastic gradient decent (SGD) for faster convergence. Besides, for regularization, we apply l2-constraints over all weights in CNN and RNN and use dropout with rate 0.5 in CNN.	Many algorithms have been proposed for representing the features of text. Previous works use either tf-idf weighting or n-gram information for the whole document or local sequence of words. recent developments in multi-label classification algorithms fall into two main categories: problem adaption and algorithm adaption. Problem adaption is about transforming multi-label classification into a simpler problem. For algorithm adaptation popular classification algorithms, e.g., k-nearest-neighbors, decision trees, support vector machines, to solve multi-label classification problems.

Angel Fiallos,Karina Jimenes (2019)	**Using Reddit Data for Multi-label Text Classification of Twitter Users Interests**	1) Multi-label text classification model: The multi-label text classification has been applied to several tasks and applications such as categorizing businesses. indexing of documents collections and detecting sentiment analysis in text Word2vec, is a predictive model, based on two layer surface neural networks that are trained to reconstruct linguistic contexts of words. This model is very efficient to learn word embeddings from raw text. Word2Vec takes a large text corpus and produces a vector space	A methodology is proposed to automatically categorize data by considering Reddit and Twitter data. First, a dataset of 42.100 publications belongs to popular forums site Reddit is collected to train a model with labeled data. Then, a dataset of tweets, an average of 100 tweets per user, from 1573 profiles is collected to predict users' topics of interest with the trained model. Finally, we were able to automatically categorize data with an average precision of 75.62%.	There are several researches focused on inferring user interest profiles considering the tweets that they post. For example, Siehndel [3] builded methods generate user interest profiles based on the extracted entities from a user's tweets, and link these entities to Wikipedia categories. Piao used categories and related entities from DBpedia for inferring user interest profiles due to improve user models for Twitter recommender systems.

Yang Tao,Zhu Cui ,Zhu Wenjun (2018)	**A Multi-Label Text Classification Method Based on Labels Vector Fusion**	Traditional text classification refers to the text of single label classification, in other words, each document belongs to only one category. Due to the complexity of the document, a document often belongs to multiple labels. Single label classification cannot meet the requirements of text classification today. A new text multi-label learning algorithm is proposed, which uses CNN to solve the dimension disaster, local optimal solution and overlearning problem of text vectors, and a more abstract high-level representation is formed by combining the underlying features. The mapping relationship between text and label vector is built through CNN. The network output is used as the label vector of the text to retrieve the nearest neighbor in the word embedding set of labels, and the nearest neighbor as the multi-label of the predicted text	Multi-label text convolutional neural network (ML-TextCNN) is constructed by fusing word embedding of labels. The text matrix formed by word embedding is used as input of ML-TextCNN, the semantic information and position information of adjacent words in the text are extracted by convolution and pooling operations. Then, the output of ML-TextCNN is used as the semantic vector of the predicted labels, and the nearest neighbor is retrieved in the word embedding space of original labels. Finally the nearest neighbor labels are used as the prediction multiple labels of the text. Under the experimental data set, tested the text multi-label.	In order to solve the shortcoming of the VSM, distributed representation of words and phrases through word embedding was proposed [13]. Its main idea is to map every word to the low dimensional and dense common vector space by using the relationship between words, so that the closer the word meaning is, the closer the distance in space.

Haytham Elghazel∗ , Alex Aussem, Ouadie Gharroudi, Wafa Saadaoui (2016)	**Ensemble multi-label text categorization based on rotation forest and latent semantic indexing**	An ensemble multi-label classification method for text categorization based on four key ideas: (1) performing Latent Semantic Indexing; (2) random splitting of the vocabulary; (3) document bootstrapping; and (4) the use of BoosTexter as a powerful multi-label base learner for text categorization.MLRF is a natural extension of the Rotation Forest paradigm to multi-labeled data. MLRF aims at building accurate and diverse multi-label classifiers. The main idea consists in applying feature extraction algorithm on different random splits of the feature set to form a new attributes for each base multi-label classifier in the ensemble. We have chosen LSI because it is considered effective to overcome the problems of lexical matching by deriving conceptual indices instead of individual terms (or words) for retrieval in a collection of documents.	We compared MLRF to several state-of-the-art algorithms, including MLkNN, the Multi-label informed Latent Semantic Indexing and three other ensemble multi-label classification approaches: BoosTexter, RAkEL, and VPCME. To make fair comparisons, the same experimental settings in was adopted here for MLSI, where they were found to yield the most satisfactory performances. As in, the instance-based learning method MLkNN was used as the base classifier for VPCME due to its excellent predictive performance, and the number of nearest neighbor k was set to 10 . For VPCME, the variable pairwise constraint threshold was empirically set to 0.6 as in. Following the experimental settings in , the ensemble size for VPCME and BoosTexter was tuned to 100.	Many methods have been proposed to generate accurate, yet diverse, sets of models. Bagging boosting, Random Subspaces , Random Forest and Rotation Forest are the most popular examples of this methodology. There are many ways to deal with this problem. BoosTexter is powerful approach proposed by Schapire and Singer, In the training phase, BoosTexter maintains a set of weights over both training examples and their labels, where training examples and their corresponding labels that are hard to predict correctly get incrementally higher weights.
Bassam Al-Salemi∗ , Masri Ayob, Shahrul Azman Mohd Noah (2018)	**Feature ranking for enhancing boosting-based multi-label text categorization**	Boosting algorithms have been proved effective for multi-label learning. As ensemble learning algorithms, boosting algorithms build classifiers by composing a set of weak hypotheses. RFBoost was introduced to manage problem based on a rank-and-filter strategy in which it first ranks the training features and then, in each learning iteration, filters and	AdaBoost.MH iteratively builds a set of weak hypotheses and then combines them as a final classifier which is capable of estimating the multiple labels for a given instance. AdaBoost.MH uses binary features to generate the weak hypotheses of decision stumps. To build a weak hypothesis during a specific boosting round, AdaBoost.MH generates a set of weak	Methods such as binary relevance (Boutell et al., 2004), classifier chains (Read et al., 2011), label powerset (Tsoumakas & Vlahavas, 2007), ranking by pairwise comparison (Hüllermeier et al., 2008), and calibrated ranking by pairwise comparison (Fürnkranz et al., 2008) have been

		uses only a subset of the highest-ranked features to construct the weak hypotheses. This step ensures accelerated learning time for RFBoost compared to AdaBoost.MHthis paper presents and investigates seven feature ranking methods in order to improve RFBoost's performance.	hypotheses, equal in number to the training features. The weak hypothesis that minimizes the Hamming loss training error is then selected, and all other hypotheses are eliminated.	introduced and used to solve many multi-label classification problems. multi-label kNN (MLkNN; Zhang & Zhou, 2007) was adapted from the traditional kNN algorithm for multi-label classification and uses the maximum posterior principle t
Sangwoo Han, Chan Lim, Bonggeon Cha, Jongwuk Lee (2021)	**An Empirical Study for Class Imbalance in Extreme Multi-label Text Classification**	Extreme multi-label text classification (XMTC) is the problem of finding the most relevant multi-labels from a text corpus with millions of labels. To overcome the class imbalance problem, existing studies suggested various methods using different loss functions (i.e., focal loss function) and data augmentation (i.e., mix-up).	we investigate focal loss and mix-up with RNN-based and transformer-based deep XMTC models on three datasets for the class imbalance problem. We conduct experiments that use the focal loss to observe improvements of tail label prediction measured by propensity-scores precision. . We evaluate our experiment with two measures. a measure has been proposed to overcome the class imbalance in the multilabel problem. Multiplying P@k with the inverse of propensity score can be interpreted as a means to give an advantage to the tail labels since labels rarely tend to have high inverse propensity score.	1-vs-All method learns classifiers for each label independently. Therefore the computational cost is very high, and the model size is also huge. Tree-based.methods build a decision tree to capture the hierarchical structure of labels or samples. SwiftXML recursively partitions a tree into two child nodes, and each internal node stores input and label features. Embedding-based.label space is large and sparse, the idea of embedding-based methods is to reduce the number of labels using low-rank assumptions. S
Rajni Jindal, Shweta (2018)	**A Novel Method for Efficient Multi-Label Text Categorization of research articles**	Text classification is a significant study area in the realm of text mining. The majority of real-world papers include many labels. They have presented a novel strategy for automating and effectively categorising multi-label text documents in	KNN is one of the most basic data mining classification algorithms. It has several disadvantages that cannot be neglected, despite its effectiveness and efficiency. Furthermore, real-world data is inherently ambiguous. To address this flaw, fuzzy KNN, which is	A Novel Method for Efficient Multi-Label Text Categorization (i e. MFZ-KNN) was used. The results of the IEEE Xplore digital library was compared with their proposed method. The

		this study. The proposed technique is based on semantic and lexical concepts. The standard IEEE taxonomy is used to identify tokens in the text documents. The standard lexical database WordNet is utilised to investigate the semantic links between tokens.	based on fuzzy membership was introduced. However, because the membership is determined during the classification stage, it had a lot of time complexity. To address this, they devised MFZ-KNN, a modified fuzzy-based KNN method in which fuzzy clusters are created during the preprocessing step and the membership of the training data set is estimated with reference to the clusters' centroid. This significantly minimises the complexity of time.	proposed method showed fairly good accuracy of 75%
Genta Indra Winata, Masayu Leylia Khodra (2015)	**Handling imbalanced dataset in multi-label text categorization using Bagging and Adaptive Boosting**	Because classifiers are weighed down by the majority of the data and neglect the minority, multi-label text classification algorithms may not yield the best results. In this research, Bagging and Adaptive Boosting techniques are used to address the problem and increase text categorization performance.	Bagging and adaptive boosting are two ways to address imbalanced datasets that are evaluated and contrasted with multi-label classifiers in this research. Both approaches can help categorizations perform better, notably for J48 and SMO. In terms of subset accuracy and example-based accuracy, SMO with Bagging.ML-LP provides the best results. The micro-averaged f-measure value for SMO with Bagging.ML-BR is the finest of all. The J48 algorithm with AdaBoost.MH, on the other hand, has the lowest hamming loss value.	Fron the results shown in the research we can conclude that bagging has more potential than adaptive boosting in increasing the accuracy of minority labels.
Gulisong Nasierding, Atul Sajjanhar (2013)	**Multi-label Classification with Clustering for Image and Text Categorization**	For multi-semantic picture and text categorization, the research investigates effective multi-label classification algorithms. For the target problem, experimental research of clustering based multi-label classification (CBMLC) is conducted. On	This paper presented experimental evaluation results of the **clustering-based multi-label classification algorithms**. The CBMLC framework is instantiated with six popular MLC algorithms, i.e. BR, LP, RAkEL, CLR, HOMER and	The Experimental conducted in the paper revealed that clustering based multi-label learning algorithms are more effective compared to their

		multi-label picture and text datasets, three commonly used clustering methods and six prominent multi-label classification algorithms are applied and evaluated.	ML-kNN, and three clustering algorithms, i.e. sIB, k-means and EM.	non-clustering counterparts.
Zhiyang He, Ji Wu, Ping Lv (2014)	**Label correlation mixture model for multi-label text categorization**	This study proposes a probabilistic generative label correlation mixing model for multiple labelled document data, which may be utilised for multi-label spoken document categorization as well as multi-label text categorization. In two phases, LCMM models the generating process of numerous labels and words in a given document, corresponding to a label correlation model and a document model. The label correlation network is established and built with the goal of establishing label correlations and estimating the prior of any subset of labels. The words are formed using labels that are represented by the document label, the parameters of which can be learned using the MCE criterion.	This work introduces label correlation mixture model (LCMM), an unique probabilistic generative model for depicting multiplylabeled documents that may be used for multilabel spoken document categorization as well as multi-label text categorization. Labels and themes in LCMM are one-to-one correspondences. Given a new document, in order to find the best subset of, all the possible label subsets should be assessed. However, the number of all the subsets is 2K−1, which makes it impractical to implement, even when is not very large. they use a simplegreedy strategy with the following three steps to solve thisproblem: Step-a: Carry out the fold-in process to calculate theconditional probabilityP(z|y) for each label in the whole label set. Step-b: Discard the labels whose conditional probability is lower than a threshold. Step-c: Compare all the possible subsets that are limited to being chosen from the remainder labels to explore the best subsets.	LCMM provides a general framework for generative models for multiply labelled document data and has advantages in terms of a succinct depiction of the data generation process and the ability to address label correlations. In addition, LCMM can be used to create numerous labelled collections of discrete data sets.

4. Problem Statement

A Stack Overflow question is divided into three sections: Title, Description, and Tags. We should be able to automatically recommend tags relating to the question's subject by using the content in the title and description. These tags are pretty important as they determine the type and field of the question. It also helps in recommending other Stack Overflow users who have already solved many questions related to those tags.

This is crucial to the company's success. The better Stack Overflow can forecast these tags, the better an Ecosystem it can build to get the correct question to the right people.

The task of giving preset categories to free-text texts is known as text categorization (also known as text classification). It has practical real - world applications and may give conceptual representations of document collections. News stories, for example, are often organized by subject categories (topics) or geographic codes; academic papers are commonly classified by technical fields and sub-domains; and patient reports in health-care organizations are frequently indexed from multiple perspectives, utilizing taxonomies of clinical conditions, different kinds of medical procedures, insurance coverage codes, and so on. Spam filtering is another common use of text classification, in which email messages are divided into two categories: spam and non-spam.

Specified tags may assist your blog in a variety of ways, depending on the blog platform you're using. Some blogs include tag clouds plugins, which display every tag which has been assigned to a blog article in the sidebar, with the most frequently used tags in bigger size. A web page for each weblog tag is usually included in most blog platforms. What is the significance of this?

Let's imagine one of your practice areas is vehicle accidents, and one of the tags you commonly use is "motorcycle accidents." Every time you publish a blog article that has that tag, it is added to a page that lists all blog entries that feature "motorcycle crashes" as a tag, increasing the likelihood that your blog will be found by search engines.

The primary purpose of using weblog tags is to categorize your material. Everything on the internet should be nice and tidy, according to Google. Consider blog tags to be distinct buckets that organize entries. However, you would not want to add tags just for adding them.

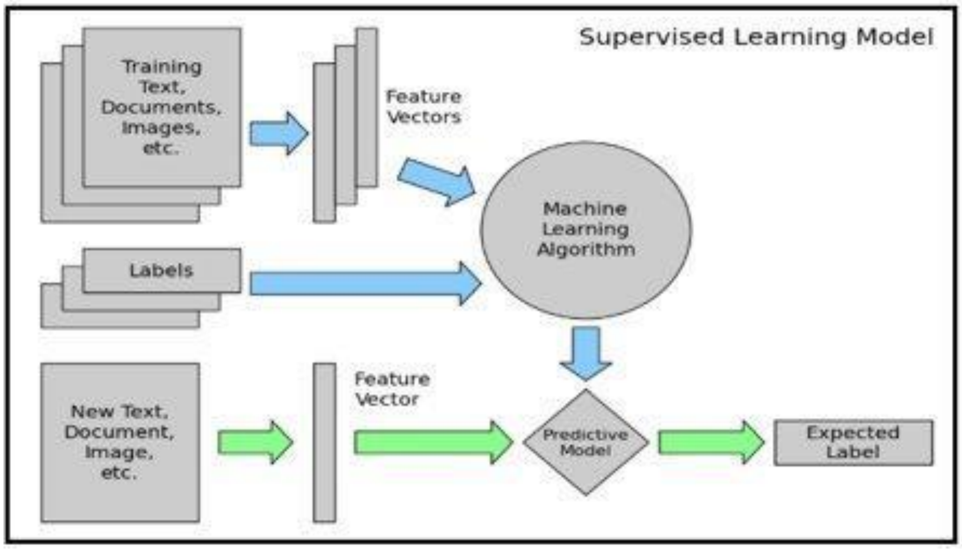
The use of picture tagging in visual marketing may be a very effective social media technique. It can add to the background of your piece and help it get traction. The technique of providing descriptive data to a photo when it is uploaded is known as image tagging.

When you combine image tagging with current keywords in your hashtags, you get a significant competitive edge.

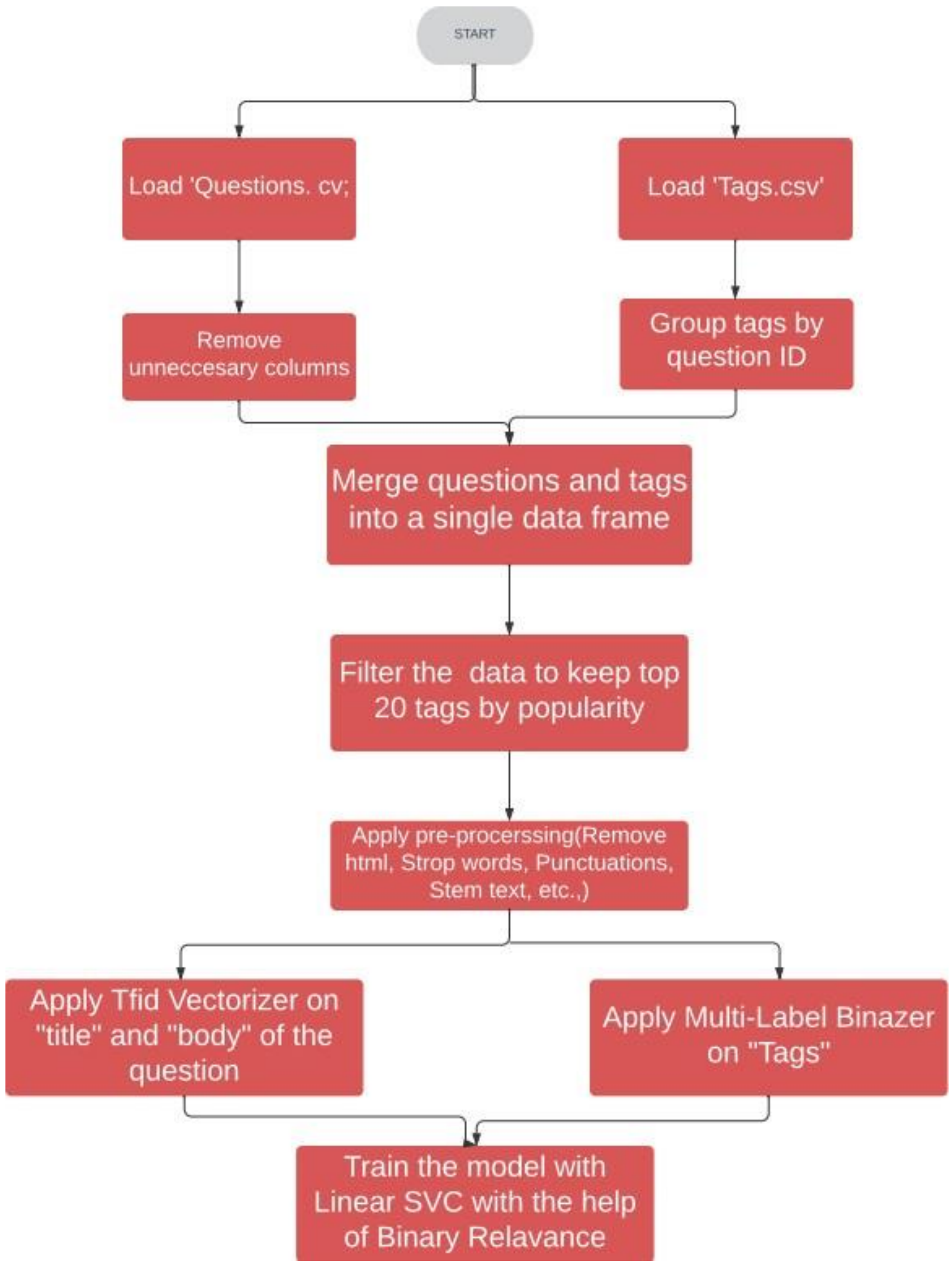
You may construct a list of tags that are presently popular on social media for convenience and quick reference. Using popular hashtags will increase your odds of reaching out to a larger group of people who are already interested.

Today, the finest channel for engaging your target audience is social media. Many educational institutions are taking use of its influence to increase admissions and placements. However, they don't always take full use of digital marketing features like the notion of tagging.
We described how hashtags operate in this post and how institutions may utilize them to successfully improve their social engagement. People are increasingly using hashtag searches to get information these days. It's a novel method of researching, and scholarly brands should be aware of how to make the most of it.

4.1 Architecture Diagram



4.2 Flow Diagram



4.3 PseudocodeExplanation

```
# Import required packages
df_q = readFile("Questions.csv")
df_t = readFile("Tags.csv")

// group all tags given to same question into a single string
grouped_tags = df_t.groupby('Id')
grouped_tags = grouped_tags['Tag'].apply(tags = ' '.join(tags))

// Drop unnecessary columns from questions
df_q.drop(columns=[OwnerUserId, CreationDate, ClosedDate], inplace=True)

// Merge questions and tags into a single dataframe
df = df_q.merge(df_tags_final, on='Id')

// get the most common 20 tags
tag_features = getMostCommn(df, "Tags", 20)

// Filter the tags from the dataset and remove all tags that does not belong to
the tag_features
df['Tags'] = df['Tags'].apply(tags: keep_common(tags))

// apply preprocessing to title
remove_html(df['Title'])
remove_stopwords(df['Title'])
remove_punc(df['Title'])
stem_text(df['Title'])
// apply preprocessing to bodya
remove_html(df['Body'])
remove_stopwords(df['Body'])
remove_punc(df['Body'])
stem_text(df['Body'])

// binarize our tags
binarizer = MultiLabelBinarizer()
y_bin = binarizer.fit_transform(y)

// vectorize
X_title_vect = vectorizer_title.fit_transform(X_title)
X_body_vect = vectorizer_body.fit_transform(X_body)

// train test split
X_train, X_test, y_train, y_test = train_test_split(X, y_bin, test_size =
0.2)

// Develop the model
svc = LinearSVC()
clf = BinaryRelevance(svc)
// fit training data
clf.fit(X_train, y_train)
y_pred = clf.predict(X_test)
print_score(y_test, y_pred)
```

5. Experiment and Results

5.1 Data set (Sample with Explanation)–Share the link source of Dataset.

The dataset "StackSample:10% of Stack Overflow Q&A" will be used. The problem statement for a multilabel text categorization method. The text of 10% of the questions and answers from the Stack Overflow programming Q&A website is included in this dataset. This is broken down into three tables:

- Questions: For all non-deleted Stack Overflow questions with an Id that is a multiple of 10, Questions provides the title, body, creation date, closed date (if applicable), score, and owner ID.
- Answers: Each of the answers to these questions has a body, a creation date, a score, and an owner ID. The ParentId column references the Questions table.
- Tags: Each of these questions has its own set of tags, which are listed under Tags.

Dataset: https://www.kaggle.com/stackoverflow/stacksample?select=Tags.csv

```
# The questions

ENCODING = 'ISO-8859-1'

df_q = pd.read_csv("drive/My Drive/B2_NLP_Team_9/dataset/Questions.csv", encoding=ENCODING)

df_q
```

1 to 25 of 20000 entries Filter

index	Id	OwnerUserId	CreationDate	ClosedDate	Score	Title	Body
0	80	26.0	2008-08-01T13:57:07Z	NaN	26	SQLStatement.execute() - multiple queries in one statement	I've written a database generation script in SQL and want to execute it in my Adobe AIR application: <pre><code>Create Table tRole (roleID integer Primary Key ,roleName varchar(40)); Create Table tFile (fileID integer Primary Key ,fileName varchar(50) ,fileDescription varchar(500) ,thumbnailID integer ,fileFormatID integer ,categoryID integer ,isFavorite boolean ,dateAdded date ,globalAccessCount integer ,lastAccessTime date ,downloadComplete boolean ,isNew boolean ,isSpotlight boolean ,duration varchar(30)); Create Table tCategory (categoryID integer Primary Key ,categoryName varchar(50) ,parent_categoryID integer); ... </code></pre> I execute this in Adobe AIR using the following methods: <pre><code>public static function RunSqlFromFile(fileName:String):void { var file:File = File.applicationDirectory.resolvePath(fileName); var stream:FileStream = new FileStream(); stream.open(file, FileMode.READ) var strSql:String = stream.readUTFBytes(stream.bytesAvailable); NonQuery(strSql); } public static function NonQuery(strSQL:String):void { var sqlConnection:SQLConnection = new SQLConnection(); sqlConnection.open(File.applicationStorageDirectory.resolvePath(DBPATH); var sqlStatement:SQLStatement = new SQLStatement(); sqlStatement.text = strSQL; sqlStatement.sqlConnection = sqlConnection; try { sqlStatement.execute(); } catch (error:SQLError) { Alert.show(error.toString()); } } </code></pre> No errors are generated, however only <code>tRole</code> exists. It seems that it only looks at the first query (up to the semicolon- if I remove it, the query fails). Is there a way to call multiple queries in one statement?
1	90	58.0	2008-08-01T14:41:24Z	2012-12-26T03:45:49Z	144	Good branching and merging tutorials for TortoiseSVN?	Are there any really good tutorials explaining branching and merging with Apache Subversion? All the better if it's specific to TortoiseSVN client.
2	120	83.0	2008-08-01T15:50:08Z	NaN	21	ASP.NET Site Maps	Has anyone got experience creating SQL-based ASP.NET site-map providers? I've got the default XML file <code>web.sitemap</code> working properly with my Menu and SiteMapPath controls, but I'll need a way for the users of my site to create and modify pages dynamically. I need to tie page viewing permissions into the standard <code>ASP.NET</code> membership system as well.
3	180	2089740.0	2008-08-01T18:42:19Z	NaN	53	Function for creating color wheels	This is something I've pseudo-solved many times and never quite found a solution. That's stuck with me. The problem is to come up with a way to generate <code>N</code> colors, that are as distinguishable as possible where <code>N</code> is a parameter.
4	260	91.0	2008-08-01T23:22:08Z	NaN	49	Adding scripting functionality to .NET applications	I have a little game written in C#. It uses a database as back-end. It's a trading card game, and I wanted to implement the function of the cards as a script. What I mean is that I essentially have an interface, <code>ICard</code>, which a card class implements (<code>public class Card056 : ICard</code>) and which contains function that are called by the game. Now, to make the thing maintainable/moddable, I would like to have the class for each card as source code in the database and essentially compile it on first use. So when I have to add/change a card, I'll just add it to the database and tell my application to refresh, without needing any assembly deployment (especially since we would be talking about 1 assembly per card which means hundreds of assemblies). Is that possible? Register a class from a source file and then instantiate it, etc. <pre><code>ICard Cards[current] = new MyGame.CardLibrary.Card056(); Cards[current].OnEnterPlay(ref currentGameState); </code></pre> The language is C#, but extra bonus if it's possible to write the script in any .NET language.
5	330	63.0	2008-08-02T02:51:36Z	NaN	29	Should I use nested classes in this case?	I am working on a collection of classes used for video playback and recording. I have one main class which acts like the public interface, with methods like <code>play()</code>, <code>stop()</code>, <code>pause()</code>, <code>record()</code> etc... Then I have workhorse classes which do the video decoding and video encoding. I just learned about the existence of nested classes in C++, and I'm curious to know what programmers think about using them. I am a little wary and not really sure what the benefits/drawbacks are, but they seem (according to the book I'm reading) to be used in cases such as mine. The book suggests that in a scenario like mine, a good solution would be to nest the workhorse classes inside the interface class, so there are no separate files for classes the client is not meant to use, and to avoid any possible naming conflicts? I don't know about these justifications. Nested classes are a new concept to me. Just want to see what programmers think about the issue.
6	470	71.0	2008-08-02T15:11:47Z	2016-03-26T05:23:29Z	13	Homegrown consumption of web services	I've been writing a few web services for a .net app, now I'm ready to consume them. I've seen numerous examples where there is homegrown code for consuming the service as opposed to using the auto generated methods Visual Studio creates when adding the web reference. Is there some advantage to this?
7	580	91.0	2008-08-	NaN	21	Deploying SQL Server	<p>I wonder how you guys manage deployment of a database between 2 SQL Servers, specifically SQL Server 2005. Now, there is a development and a live one. As this should be part of a buildscript (standard windows batch, even do with current complexity of those scripts, i might switch to PowerShell or so later), Enterprise Manager/Management Studio Express do not count.</p> <p>Would you just copy the .mdf File and attach it? I am always a bit careful when working with binary data, as this seems to be a compatiblity issue (even though development and live should run the same version of the server at all time).</p> <p>Or - given the lack of "EXPLAIN CREATE TABLE" in T-SQL - do you do something that exports an existing database into SQL-Scripts which you can run on the target server? If yes, is there a tool that can automatically dump a given Database into SQL Queries and that runs off the command line? (Again, Enterprise Manager/Management Studio Express do not

```python
# the tags
df_t = pd.read_csv("drive/My Drive/B2_NLP_Team_9/dataset/Tags.csv", encoding=ENCODING)
df_t
```

1 to 25 of 20000 entries Filter

index	Id	Tag
0	80	flex
1	80	actionscript-3
2	80	air
3	90	svn
4	90	tortoisesvn
5	90	branch
6	90	branching-and-merging
7	120	sql
8	120	asp.net
9	120	sitemap
10	180	algorithm
11	180	language-agnostic
12	180	colors
13	180	color-space
14	260	c#
15	260	.net
16	260	scripting
17	260	compiler-construction
18	330	c++
19	330	oop
20	330	class
21	330	nested-class
22	470	.net
23	470	web-services
24	580	sql-server

5.1.1 Explain methodology with the dataset.

- Load Questions.csv and Tags.csv

```
import numpy as np
import pandas as pd
```

```
[7] # The questions

ENCODING = 'ISO-8859-1'

df_q = pd.read_csv("drive/My Drive/B2_NLP_Team_9/dataset/Questions.csv", encoding=ENCODING)

df_q.head()
```

	Id	OwnerUserId	CreationDate	ClosedDate	Score	Title	Body
0	80	26.0	2008-08-01T13:57:07Z	NaN	26	SQLStatement.execute() - multiple queries in o...	<p>I've written a database generation script i...
1	90	58.0	2008-08-01T14:41:24Z	2012-12-26T03:45:49Z	144	Good branching and merging tutorials for Torto...	<p>Are there any really good tutorials explain...
2	120	83.0	2008-08-01T15:50:08Z	NaN	21	ASP.NET Site Maps	<p>Has anyone got experience creating <strong>...
3	180	2089740.0	2008-08-01T18:42:19Z	NaN	53	Function for creating color wheels	<p>This is something I've pseudo-solved many t...
4	260	91.0	2008-08-01T23:22:08Z	NaN	49	Adding scripting functionality to .NET applica...	<p>I have a little game written in C#. It uses...

```
# the tags

df_t = pd.read_csv("drive/My Drive/B2_NLP_Team_9/dataset/Tags.csv", encoding=ENCODING)

df_t.head()
```

	Id	Tag
0	80	flex
1	80	actionscript-3
2	80	air
3	90	svn
4	90	tortoisesvn

- Group all tags given to the same question into a single string as the data is not grouped by ID.

```
df_t['Tag'] = df_t['Tag'].astype(str)

# group all tags given to same question into a single string
grouped_tags = df_t.groupby('Id')['Tag'].apply(lambda tags: ' '.join(tags))

grouped_tags.head()
```

```
Id
80                          flex actionscript-3 air
90        svn tortoisesvn branch branching-and-merging
120                             sql asp.net sitemap
180     algorithm language-agnostic colors color-space
260            c# .net scripting compiler-construction
Name: Tag, dtype: object
```

```
[10] # reset index for simplicity
grouped_tags.reset_index()

df_tags_final = pd.DataFrame({'Id': grouped_tags.index, 'Tags': grouped_tags.values})

df_tags_final.head()
```

	Id	Tags
0	80	flex actionscript-3 air
1	90	svn tortoisesvn branch branching-and-merging
2	120	sql asp.net sitemap
3	180	algorithm language-agnostic colors color-space
4	260	c# .net scripting compiler-construction

- Drop unnecessary columns from questions and Merge questions and tags into a single dataframe

```
[11] # Drop unnecessary columns from questions
     df_q.drop(columns=['OwnerUserId', 'CreationDate', 'ClosedDate'], inplace=True)
```

```
# Merge questions and tags into a single dataframe
df = df_q.merge(df_tags_final, on='Id')

del df_q
del df_t

df.head()
```

	Id	Score	Title	Body	Tags
0	80	26	SQLStatement.execute() - multiple queries in o...	<p>I've written a database generation script i...	flex actionscript-3 air
1	90	144	Good branching and merging tutorials for Torto...	<p>Are there any really good tutorials explain...	svn tortoisesvn branch branching-and-merging
2	120	21	ASP.NET Site Maps	<p>Has anyone got experience creating <strong>...	sql asp.net sitemap
3	180	53	Function for creating color wheels	<p>This is something I've pseudo-solved many t...	algorithm language-agnostic colors color-space
4	260	49	Adding scripting functionality to .NET applica...	<p>I have a little game written in C#. It uses...	c# .net scripting compiler-construction

- Remove questions with scores lower than 5 to avoid outliers and split tags into list

```
# remove questions with score lower than 5
df = df[df['Score'] > 5]

print(df.shape)
```

```
(72950, 5)
```

```
# split tags into list
df['Tags'] = df['Tags'].apply(lambda tags: tags.lower().split())

df.head()
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

	Id	Score	Title	Body	Tags
0	80	26	SQLStatement.execute() - multiple queries in o...	<p>I've written a database generation script i...	[flex, actionscript-3, air]
1	90	144	Good branching and merging tutorials for Torto...	<p>Are there any really good tutorials explain...	[svn, tortoisesvn, branch, branching-and-merging]
2	120	21	ASP.NET Site Maps	<p>Has anyone got experience creating <strong>...	[sql, asp.net, sitemap]
3	180	53	Function for creating color wheels	<p>This is something I've pseudo-solved many t...	[algorithm, language-agnostic, colors, color-s...
4	260	49	Adding scripting functionality to .NET applica...	<p>I have a little game written in C#. It uses...	[c#, .net, scripting, compiler-construction]

- Get the most common 20 tags without the count and filter the dataset to only consider the data which has top 20 tags.

```python
# get the most common 20 tags without the count
tag_features = list(map(lambda x: x[0], tag_freq.most_common(20)))

print(tag_features)
```

```
['c#', 'java', 'javascript', 'android', 'python', 'c++', 'php', 'jquery', '.net', 'ios', 'html', 'css', 'c', 'iphone', 'objective-c', 'ruby-on-rails', 'sql', 'asp.net', 'mysql', 'ruby']
```

```python
# Filter the tags from the dataset and remove all tags that does not belong to the tag_features
def keep_common(tags):

    filtered_tags = []

    # filter tags
    for tag in tags:
        if tag in tag_features:
            filtered_tags.append(tag)

    # return the filtered tag list
    return filtered_tags

# apply the function to filter in dataset
df['Tags'] = df['Tags'].apply(lambda tags: keep_common(tags))

# set the Tags column as None for those that do not have a most common tag
df['Tags'] = df['Tags'].apply(lambda tags: tags if len(tags) > 0 else None)

# Now we will drop all the columns that contain None in Tags column
df.dropna(subset=['Tags'], inplace=True)

df.shape
```

```
(48976, 5)
```

- Apply data pre-processing by:
    - **Remove HTML**
    - **Remove stopwords**
    - **Remove special characters**
    - **Convert to lowercase**
    - **Stemming**

Html can be handled by the use of **Regular Expressions** and the rest can be done by the help of the **nltk** liberty

```python
def remove_html(text):
    # Remove html and convert to lowercase
    return re.sub(r"\<[^\>]\>", "", text).lower()

def remove_stopwords(text):
    # tokenize the text
    words = tokenizer.tokenize(text)

    filtered = [w for w in words if not w in stop_words]
    return ' '.join(map(str, filtered))

def remove_punc(text):
    #tokenize
    tokens = tokenizer.tokenize(text)

    # remove punctuations from each token
    tokens = list(map(lambda token: re.sub(r"[^A-Za-z0-9]+", " ", token).strip(), tokens))

    # remove empty strings from tokens
    tokens = list(filter(lambda token: token, tokens))

    return ' '.join(map(str, tokens))

def stem_text(text):
    #tokenize
    tokens = tokenizer.tokenize(text)

    # stem each token
    tokens = list(map(lambda token: stemmer.stem(token), tokens))

    return ' '.join(map(str, tokens))
```

```python
# apply preprocessing to title and bodya
df['Title'] = df['Title'].apply(lambda x: remove_html(x))
df['Title'] = df['Title'].apply(lambda x: remove_stopwords(x))
df['Title'] = df['Title'].apply(lambda x: remove_punc(x))
df['Title'] = df['Title'].apply(lambda x: stem_text(x))

df.head()
```

	Title	Body	Tags
2	asp net site map	<p>Has anyone got experience creating <strong>...	[sql, asp.net]
4	ad script function net applic	<p>I have a little game written in C#. It uses...	[c#, .net]
5	use nest class case	<p>I am working on a collection of classes use...	[c++]
6	homegrown consumpt web servic	<p>I've been writing a few web services for a ...	[.net]
8	automat updat version number	<p>I would like the version property of my app...	[c#]

```python
# apply preprocessing to title and body
df['Body'] = df['Body'].apply(lambda x: remove_html(x))
df['Body'] = df['Body'].apply(lambda x: remove_stopwords(x))
df['Body'] = df['Body'].apply(lambda x: remove_punc(x))
df['Body'] = df['Body'].apply(lambda x: stem_text(x))

df.head()
```

	Title	Body	Tags
2	asp net site map	anyon got experi creat strong sql base asp net...	[sql, asp.net]
4	ad script function net applic	littl game written c use databas back end a hr...	[c#, .net]
5	use nest class case	work collect class use video playback record o...	[c++]
6	homegrown consumpt web servic	write web servic net app readi consum them see...	[.net]
8	automat updat version number	would like version properti applic increment b...	[c#]

- Apply **MultiLabelBinarizer** on Tags and **TfidfVectorizer** on Questions

```python
from sklearn.preprocessing import MultiLabelBinarizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from scipy.sparse import hstack
```

```python
[24] X_title = df['Title']
X_body = df['Body']
y = df['Tags']

del df

# binarize our tags
binarizer = MultiLabelBinarizer()
y_bin = binarizer.fit_transform(y)
```

```python
# vectorize
vectorizer_title = TfidfVectorizer(
    analyzer = 'word',
    strip_accents = None,
    encoding = 'utf-8',
    preprocessor=None,
    max_features=10000)

vectorizer_body = TfidfVectorizer(
    analyzer = 'word',
    strip_accents = None,
    encoding = 'utf-8',
    preprocessor=None,
    max_features=10000)

X_title_vect = vectorizer_title.fit_transform(X_title)
X_body_vect = vectorizer_body.fit_transform(X_body)

X = hstack([X_title_vect, X_body_vect])
```

- The most applicable machine learning algorithm for our problem is **Linear SVC**.
- The objective of a Linear SVC (Support Vector Classifier) is to fit to the data you provide, returning a "best fit" hyperplane that divides or categorizes, your data. From there, after getting the hyperplane, you can then feed some features to your classifier to see what the "predicted" class is. This makes this specific algorithm rather suitable for our uses, though you can use this for many situations.
- For trying the model we decide to go with **Linear SVC** after a lot of testing as it constantly gave better results when compared to other classical classification methods like SGD Classifier and other Regression models.
- We also used **Binary Relevance** from the scikit-multilearn lib. Binary Relevance transforms a multi-label classification problem with L labels into L single-label separate binary classification problems using the same base classifier provided in the constructor. The prediction output is the union of all per label classifiers.

```
# Develop the model
from sklearn.svm import LinearSVC
from skmultilearn.problem_transform import BinaryRelevance # gives better precision

svc = LinearSVC()
clf = BinaryRelevance(svc)

# fit training data
clf.fit(X_train, y_train)
```

```
BinaryRelevance(classifier=LinearSVC(), require_dense=[True, True])
```

```
# Metrics
from sklearn.metrics import accuracy_score, precision_score, recall_score, hamming_loss, f1_score

# make prediction
y_pred = clf.predict(X_test)

print_score(y_test, y_pred)
```

```
Jacard score: 0.6670273581053491
----
Recall: 0.6732329084588644
----
Precision:  0.8087120124028273
----
Hamming Loss (%):  2.9374234381380155
----
F1 Score:  0.7452753664819695
----
```

5.1.2 Partial output

Link to Google Collab Notebook:

https://colab.research.google.com/drive/18CMI_zLSWoW-11HUIx30VEMPZA4Nta_w?usp=sharing

```
# Metrics
from sklearn.metrics import accuracy_score, precision_score, recall_score, hamming_loss, f1_score

# make prediction
y_pred = clf.predict(X_test)

print_score(y_test, y_pred)
```

```
Jacard score: 0.6670273581053491
----
Recall: 0.6732329084588644
----
Precision:  0.8087120124028273
----
Hamming Loss (%):  2.9374234381380155
----
F1 Score:  0.7452753664819695
----
```

Test Data:

```
df_x = pd.concat([X_title, X_body], axis= 1)
X_train1, X_test1, y_train1, y_test1 = train_test_split(df_x, y, test_size = 0.2, random_state = 0)
df_x = pd.DataFrame(X_test1)
df_x.reset_index(drop=True, inplace=True)
```

```
p = binarizer.inverse_transform(y_pred)
df_p = pd.DataFrame({'Predducted':p})

yTest = binarizer.inverse_transform(y_test)
df_y = pd.DataFrame({'Tags':yTest})

df_p = pd.concat([X_test1, df_y, df_p], axis= 1)
df_p
```

	Title	Body	Tags	Predducted
0	hibern nosuchfielderror instanc strut 1	new java hibern rail c develop anyway test pro...	(java,)	(java,)
1	unit test bad form unit test call unit test	unit test call code testmakeavalidcal code tes...	(c#,)	()
2	argument unpack use iter item get	use python 2 7 3 p consid dummi class custom a...	(python,)	(python,)
3	list paramet caus error	code p pre code public bool somemethod list lt...	(c#,)	(c#,)
4	xsd xml document generat c	anyon know xsd file somewher use valid xml doc...	(c#,)	(c#,)
...	...	...	...	...
9791	pgp python generat key encrypt decrypt	make program python distribut window user via ...	(python,)	(python,)
9792	best practic implement secur rememb	sometim come across certain web develop framew...	(asp.net,)	()
9793	broadcast receiv check internet connect androi...	hello develop android broadcast receiv check i...	(android, java)	(android,)
9794	html code caus caption comment pop mous roll h...	peic html code allow provid comment link user ...	(html,)	(html,)
9795	check variabl null undefin javascript	blockquot strong possibl duplic strong br a hr...	(javascript, jquery)	(javascript,)

9796 rows × 4 columns

index	Title	Body	Tags	Predducted
2475	chang android l keyboard enter key color	new android l keyboard use system theme coloracc background color enter key match app custom theme way chang that p would assum theme style keyboard find theme materi xml style found android keyboardviewstyl give error 9 21 resourc found match given name attr android keyboardviewstyl p img src http i stack imgur com jnhb4 png width 400 p	android	android
2476	branchless memori manag	anyon thought write memori manag c complet branch free written pool stack queue link list alloc pool wonder plausibl write branch free general memori manag p help make realli reusabl framework solid concurr in order cpu cach friend develop p edit branchless mean without direct indirect function call without use if think probabl implement someth first chang request size zero fals call realli got much that feel imposs aspect exercis profil said unfriend processor see worth tri hard avoid branch p	c++	c++
2477	detect held mous button picturebox	need fire event mous picturebox mous button alreadi click held down p problem p mousedown mouseent event handler work togeth well p instanc mous button click held c fire mousedown event handler cursor move picturebox mouseent event fire mous button realeas p	c#	c#
2478	heavi javascript page gap 15 second respons page load	page heavi javascript page leav page go page b take long time go page b differ page realli fast someth page probabl javascript p run network profil develop tool ie 9 show gap 15 second respons domcontentload event p page heavi javascript run xopus editor rich text xml editor p anybodi idea could either analys gap happen could make page unload faster p	javascript	javascript
2479	datepick popup format work valu set initi scope	use angular ui bootstrap date picker popup use custom direct plunker a href http plnkr co edit 053vjym1mpzuikwftfrt p preview http plnkr co edit 053vjym1mpzuikwftfrt p preview a p pre code modul var usermodul angular modul usermodul ui bootstrap control usermodul control samplecontrol scope function scope scope mindat new date direct code usermodul direct datepick function datefilt return restrict e requir ngmodel scope ngmodel ngreadon mindat maxdat dtprequir dateopt templat lt p class input group gt lt input type text style cursor pointer class form control datepick popup format ng model ngmodel is open open min date mindat max date maxdat datepick option dateopt date disabl disabl date mode ng requir dtprequir close text close ng readon ngreadon ng click openpopup gt lt span class input group btn gt lt button type button class btn btn default ng click openpopup event gt lt class fa fa calendar gt lt i gt lt button gt lt span gt lt p gt control function scope check defin set default scope dateopt scope dateopt 124 124 formatyear yy startingday 1 showweek fals scope openpopup function event event undefin event stoppropag scope open true scope format dd mmmm yyyi link function scope element attr control remov default formatt input direct prevent conflict control formatt shift code pre work fine date format fine select date calendar popup howev set date ng model control date format dd mmmm yyyi return date string like sat oct 01 2016 01 00 00 gmt 0100 gmt daylight time howev plunker abl set date control format fine html date picker p pre code lt date picker ng model startdat valu datepick option dateopt min date mindat ng readon true gt lt date picker gt code pre control code startdat valu new date code p sure problem could be imag show get back p a href http i stack imgur com brvl4 png img src http i stack imgur com brvl4 png alt wrong date a p	html,javascript	javascript
2480	php syntax highlight	search strong php syntax highlight engin strong custom i e provid strong own token strong new languag handl sever languag em simultan em i e output page engin work well togeth strong css class strong i e format output insert code lt span gt code element adorn code class code attribut bonus point extens schema p em not em search client side syntax highlight script javascript p far stuck a href http qbnz com highlight geshi a unfortun geshi fail abysm sever reason main reason differ languag file defin complet differ inconsist style work hour tri refactor differ languag definit common denomin sinc definit file quit bad final like switch p ideal like api similar a href http coderay rubychan de coderay a a href http pygment org pygment a javascript a href http code googl com p syntaxhighlight dp syntaxhighlight a p h2 clarif h2 look code highlight softwar written em in em php em for em php sinc need use insid php p	php	php
2481	issu touch scroll io focus input	issu scrolleabl div io tri scroll touch outsid input scroll ok without problem tri scroll touch input start scroll lot chanc happen div lot input scroll whole window instead scroll div problem either desktop android found similar question a href http stackoverflow com question 14333410 io html input tag stop scroll in scrollabl element io html input tag stop scroll scrollabl element a answer either find good solut decid prevent event touchmov user touch input exact want p may someon alreadi face problem help would realli appreci it p thank advanc p regard p marcelo p	html,ios	ios,javascript
2482	scale drawabl setimagelevel dictat scale	creat small test app tri scale drawabl p defin drawabl xml save code scale up xml code code main xml code layout file code imageview code use code android src drawabl scale up code imageview use drawabl creat xml activ use code myimageview setimagelevel level code set level imageview defin code main xml code well imageview display drawabl defin scale up p confus follow p ul li chang code scaleheight code code scalewidth code code scale up xml code make perceiv differ level set 10000 li li chang level code setimagelevel code 0 10000 chang size imag realli perceiv around 5000 higher valu scaleheight scalewidth smaller imag depend valu code setimagelevel code li ul suppos question s p ol li is correct use code setimagelevel code pass level 0 10000 activ control size drawabl defin xml file li li what relationship scaleheight scalewidth code setimagelevel code see may well set scaleheight scalewidth 100 use code setimagelevel 5000 code get 50 scale imag i e increas reduc code setimagelevel 0 10000 code chang scale make code scaleheight code code scalewidth code pretti pointless li ol appreci clarif perhap exampl use scale drawabl defin xml correct p	android	android
2483	calcul x 509 certif sha 1 fingerprint c c object c	background p write client util capabl connect remot server use ssl tls client use openssl perform ssl tls transact would like allow user specifi author ca cert case self sign cert privat ca setup use sign server certif plan use cert fingerprint common name valid date allow user quick view cert client use valid server p question p calcul sha1 hash fingerprint x509 cert store within pem file use c c object c p day search experi found solut post answer howev welcom better correct solut p	c	
2484	question regard c non pod union	c 11 gave us possibl use non pod type within union say follow piec code p pre code union one v two uni code pre somewher within class one member activ time question rather simpl p ol li what default valu uni undefin li li whenev class destruct member within union destruct ul li suppos std typeinfo keep track activ member call destructor explicit member destructor li ul li li doe anyon link languag propos chang union accept non pod type li ol	c++	c++
2485	reason overhead memori usag array java	java charact data type code char code repres 2 byte array code n code charact code char code repres code 2n 24 code byte p general overhead code 24 code byte store array code n code object least object primit type p need addit 24 byte use p strong edit strong juli 2nd 2015 strong strong brought attent comment answer question offer a href http programm stackexchang com question 162546 whi the overhead when alloc object array in java here a programm stackexchang p	java	java
2486	python decor simpl recurs standard librari elsewher	look python decor make function recurs find write lot function like p pre code def xyz data isinst data typethatdenotessingular return map xyz data return singular xyz data code pre figur must decor somewher standard librari slim notat tad p pre code recurs typethatdenotessingular def xyz data return singular xyz data code pre search seem get anywher perhap miss essenti terminolog p thank point right direct p	python	python

index	Title	Body	Tags	Predducted
50	display activ indic wkwebview use swift	work follow code tri show activ indic view whilst page load p tri implement strong wknavigationdeleg strong method fail noth show p suggest fix this p set supportwebview view strong deleg strong anywher know swift p pre code import uikit import webkit class supportwebview uiviewcontrol wknavigationdeleg iboutlet var containerview uiview nil var webview wkwebview overrid func loadview super loadview self webview wkwebview self view self webview overrid func viewdidload super viewdidload var datamanag datamanag shareddatamanag var url datamanag myvalidurl var req nsurlrequest url url self webview loadrequest req overrid func didreceivememorywarn super didreceivememorywarn dispos resourc recreat func webview webview wkwebview didstartprovisionalnavig navig wknavig uiapplic sharedappl networkactivityindicatorvis true func webview webview wkwebview didfinishnavig navig wknavig uiapplic sharedappl networkactivityindicatorvis fals code pre	ios	ios
51	know calle answer call phone state lift call	tri know alert calle lift call use code phonestatelisten code along code broadcastreceiv code p general three state code call state idl call state offhook call state ring code p code call state offhook code state call call connect state three state call calle answer call p broadcastreceiv p pre code public class phonestatebroadcastreceiv extend broadcastreceiv overrid public void onrec context context intent intent telephonymanag telephonymanag telephonymanag context getsystemservic context telephoni servic telephonymanag listen new customphonestatelisten context phonestatelisten listen call state public class customphonestatelisten extend phonestatelisten context context context make toast requir activitymanag activitymanag public customphonestatelisten context context super this context context overrid public void oncallstatechang int state string incomingnumb super oncallstatechang state incomingnumb log v phonestatebroadcastreceiv oncallstatechang state state switch state case telephonymanag call state idl toast maketext context call state idl toast length long show break case telephonymanag call state offhook toast maketext context call state offhook toast length long show break case telephonymanag call state ring toast maketext context call state ring toast length long show break default break code pre seen applic record voic call accept want know state accept call p state listen know calle answer call p	android	android
52	submit file jqueri ajax give typeerror	tri submit file form use jqueri code ajax code method p pre code var ofil document getelementbyid imag file 0 var formdata new formdata formdata append imag ofil ajax url element save element data formdata type post code pre result error code typeerror append call object implement interfac formdata code p caus error happen actual code formdata append code insid jqueri p	javascript,jquery	javascript,jquery
53	convert variabl string valu express javascript	tri convert string valu express use condit like p pre code var stringval 20 gt 18 amp amp yes yes if stringval code pre possibl pleas suggest me p thank p	javascript	javascript
54	differ size t std size t	differ code size t code code std size t code term declar use differenti featur p	c++	c++
55	forc nvidia gpu programmat optimus laptop	program directx game run optimus laptop intel gpu use result horribl perform forc nvidia gpu use context menu renam execut bf3 exe famous game execut name perform expect br obvious neither accept solut redistribut game way programmat forc laptop use nvidia gpu br br alreadi tri use directx enumer adapt idirect3d9 getadaptercount idirect3d9 getadapteridentifi work 1 gpu report one use p	c++	
56	pass object paramet constructor function copi properti new object	javascript constructor like p pre code function box obj this obj obj code pre want pass object paramet like p pre code var box new box prop1 prop2 b prop3 c code pre give someth like p pre code box obj prop1 box obj prop2 box obj prop3 code pre would like properti direct object like p pre code box prop1 box prop2 box prop3 code pre know could someth like p pre code function box obj this prop1 obj prop1 this prop2 obj prop2 this prop3 obj prop3 code pre good constructor would know name properti object paramet would like abl pass differ object paramet assign properti direct properti new custom object creat constructor get code box propx code code box obj propx code hope make clear mayb meas someth obvious newbi pleas need help p thank advanc p	javascript,jquery	c#,objective-c
57	write stereo wav file python	follow code write simpl sine frequenc 400hz mono wav file code chang order produc strong stereo strong wav file second channel differ frequenc p pre code import math import wave import struct freq 440 0 data size 40000 fname wavetest wav frate 11025 0 framer float amp 64000 0 multipli amplitud sine list x x rang data size sine list x append math sin 2 math pi freq x frate wav file wave open fname w nchannel 1 sampwidth 2 framer int frate nframe data size comptyp none compnam compress wav file setparam nchannel sampwidth framer nframe comptyp compnam sine list x write audio frame file wav file writefram struct pack h int s amp 2 wav file close code pre	python	python
58	differ 01 1 javascript date	differ date 2015 10 01 2015 10 1 javascript p pre code new date 2015 10 1 code pre return code thu oct 01 2015 00 00 00 gmt 0300 code p pre code new date 2015 10 01 code pre return code wed sep 30 2015 21 00 00 gmt 0300 code p	javascript	javascript
59	java increas xmx dynam runtim	jvm server machin want 2 apserv mine sit machin howev want standbi one realli low amount memori alloc xmx passiv one main server activ goe want alloc memori passiv server alreadi without restart much xmx note would consum memori startup cant allow possibl outofmemori p want passiv low xmx activ goe want passiv receiv much xmx p way achiev that thank p	java	java
60	fcm aw sns	use code aw code resourc android project plan add push notif servic project code aw sns code there question bother much find question regard except one two unclear explan p 1 doe code aw code support code fcm code code sns code work code gcm code googl recommend use fcm instead code gcm code find code aw code support code fcm code p 2 do code aw code store messag data databas even send push notif p 3 i tri put fcm api key sns applic platform show invalid paramet whi p	android	
61	stop webcam stream getusermedia without page refresh	tri close webcam javascript function close receiv ajax respons seem imposs close without refresh page method close like video src null video paus etc work browser uniqu way close stream pass paramet success function way use object outsid function success close webcam p know question ask a href http stackoverflow com question 15137069 stop close webcam use getusermedia and rtcpeerconnect chrome 25 stop close webcam use getusermedia rtcpeerconnect chrome 25 a need differ would need help solv problem p thank p edit work code tri close webcam p pre code navig getusermedia navig getusermedia 124 124 navig mozgetusermedia 124 124 navig webkitgetusermedia 124 124 navig msgetusermedia if navig getusermedia var video constraint mandatori maxheight 480 maxwidth 640 option var self self navig getusermedia audio fals video video constraint self onsuccess onerror els alert error occur start webcam stream pleas revis instruct fix problem function onsuccess stream var video document getelementbyid webcam if navig webkitgetusermedia 124 124 navig mozgetusermedia video src window url createobjecturl stream els if navig msgetusermedia futur implement internet explor els video src stream self localstream stream video play function onerror alert problem retriev stream allow access function closewebcamconnect this localstream stop code pre uff realli complic post code xd p	javascript	javascript

Example Application:

Example data :

'Title':

- 'How to handle or avoid a stack overflow in C++',
- "Python getopt module error NameError: name 'opts' is not defined after importing",
- "get append element but got undefined",
- "I have a problem with pip",
- "Best way to center a <div> on a page vertically and horizontally? "],

'Body':

- 'In C++ a stack overflow usually leads to an unrecoverable crash of the program. For programs that need to be really robust, this is an unacceptable behaviour, particularly because stack size is limited. A few questions about how to handle the problem. Is there a way to prevent stack overflow by a general technique. (A scalable, robust solution, that includes dealing with external libraries eating a lot of stack, etc.) Is there a way to handle stack overflows in case they occur? Preferably, the stack gets unwound until theres a handler to deal with that kinda issue. There are languages out there, that have threads with expandable stacks. Is something like that possible in C++? Any other helpful comments on the solution of the C++ behaviour would be appreciated.',

- "I'm trying to take in two arguments from the console. The following code seems to have worked on my colleague's computer, so I'm not sure why it is giving me an error when trying to run it on mine. I am on a Mac.import getoptimport sysquestion_id= Nonearg_student = Noneargv = sys.argv[1:]print('test')try:opts, args = getopt.getopt(argv, 'i:s:', ['question_id=','arg_student='])except:print('Error')for opt, arg in opts:if opt in ['-i', '--question_id']:question_id = argelif opt in ['-s', '--arg_student']:arg_student = argprint('Question Number: ' + question_id)print('Student response: ' + arg_student)This is the error I am getting:ErrorTraceback (most recent call last):File '/Users/ailanysmacbook/github/AutomatedEssayGrading/AutomatedEssayGrading/input.py', line 1, in <module>import getoptFile '/Users/ailanysmacbook/github/AutomatedEssayGrading/AutomatedEssayGrading/getopt.py', line 20, in <module>for opt, arg in opts:NameError: name 'opts' is not definedIt seems to be happening right after I try importing it. Do I need to install something? I'm not sure what's missing.",
- "I have append element on my page, and after the element was append i want to send the element value with ajax in javascrypt. But i got response undefinedappend element jquery",
- "I am facimg problems in installing packages from pip",
- "Best way to center a <div> element on a page both vertically and horizontally?I know that margin-left: auto; margin-right: auto; will center on the horizontal, but what is the best way to do it vertically, too?"]

Example Application

```
[51] # Actual Application
df_app = pd.DataFrame({
    'Title':['How to handle or avoid a stack overflow in C++',
             "Python getopt module error NameError: name 'opts' is not defined after importing",
             "get append element but got undefined",
             "I have a problem with pip",
             "Best way to center a <div> on a page vertically and horizontally? "],
    'Body':['In C++ a stack overflow usually leads to an unrecoverable crash of the program. For programs that need to be really robust, this is an unacceptable behaviour, particularly because stack size is limited. A few questions a
            "I'm trying to take in two arguments from the console. The following code seems to have worked on my colleague's computer, so I'm not sure why it is giving me an error when trying to run it on mine. I am on a Mac.import g
            "I have append element on my page, and after the element was append i want to send the element value with ajax in javascrypt. But i got response undefinedappend element jquery",
            "I am facimg problems in installing packages from pip",
            "Best way to center a <div> element on a page both vertically and horizontally?I know that margin-left: auto; margin-right: auto; will center on the horizontal, but what is the best way to do it vertically, too?"]
    })
```

```
# preprocessing title and body
df_app['Title'] = df_app['Title'].apply(lambda x: remove_html(x))
df_app['Title'] = df_app['Title'].apply(lambda x: remove_stopwords(x))
df_app['Title'] = df_app['Title'].apply(lambda x: remove_punc(x))
df_app['Title'] = df_app['Title'].apply(lambda x: stem_text(x))

df_app['Body'] = df_app['Body'].apply(lambda x: remove_html(x))
df_app['Body'] = df_app['Body'].apply(lambda x: remove_stopwords(x))
df_app['Body'] = df_app['Body'].apply(lambda x: remove_punc(x))
df_app['Body'] = df_app['Body'].apply(lambda x: stem_text(x))

df_app
```

	Title	Body
0	handl avoid stack overflow c	c stack overflow usual lead unrecover crash pr...
1	python getopt modul error nameerror name opt d...	tri take two argument consol follow code seem ...
2	get append element got undefin	append element page element append want send e...
3	problem pip	facimg problem instal packag pip
4	best way center div page vertic horizont	best way center div element page vertic horizo...

```
[53] x_title = vectorizer_title.transform(df_app['Title'])
     x_body = vectorizer_body.transform(df_app['Body'])

     P_app = hstack([x_title, x_body])
     P_app.shape

(5, 20000)
```

```
[54] py= clf.predict(P_app)
     binarizer.inverse_transform(py)

[('c++',), ('python',), ('javascript', 'jquery'), ('python',), ('css', 'html')]
```

6. Conclusion

Initially, we used Pandas Data Frame to load the Text Pre-Processed Dataset, then we analyzed the String Tags.AST Module and encoded the tags using Multilabelbinarizer.
Following that, we used TfidfVectorizer to do text vectorization on the question sort dataset. Consequently, we've tested the model against a variety of classifiers, including SGDClassifier, LinearSVC and Supplying Regression For Multi-Label Classification, and we've compared the results to real data.
We found that LinearSVC gave the best results among the classification models tested.

7. Reference

1. Y. Tao, Z. Cui and Z. Wenjun, "A Multi-Label Text Classification Method Based on Labels Vector Fusion," 2018 International Conference on Promising Electronic Technologies (ICPET), 2018, pp. 80-85, doi: 10.1109/ICPET.2018.00021.
2. A. Fiallos and K. Jimenes, "Using Reddit Data for Multi-Label Text Classification of Twitter Users Interests," 2019 Sixth International Conference on eDemocracy & eGovernment (ICEDEG), 2019, pp. 324-327, doi: 10.1109/ICEDEG.2019.8734365.https://ieeexplore.ieee.org/document/7078554
3. G. Chen, D. Ye, Z. Xing, J. Chen and E. Cambria, "Ensemble application of convolutional and recurrent neural networks for multi-label text categorization," 2017 International Joint Conference on Neural Networks (IJCNN), 2017, pp. 2377-2383, doi: 10.1109/IJCNN.2017.7966144.
4. S. Han, C. Lim, B. Cha and J. Lee, "An Empirical Study for Class Imbalance in Extreme Multi-label Text Classification," 2021 IEEE International Conference on Big Data and Smart Computing (BigComp), 2021, pp. 338-341, doi: 10.1109/BigComp51126.2021.00073.
5. Al-Salemi, B., Ayob, M., & Noah, S. A. M. (2018). Feature ranking for enhancing boosting-based multi-label text categorization. *Expert Systems with Applications*, *113*, 531–543. https://doi.org/10.1016/j.eswa.2018.07.024
6. Elghazel, H., Aussem, A., Gharroudi, O., & Saadaoui, W. (2016). Ensemble multi-label text categorization based on rotation forest and latent semantic indexing. *Expert Systems with Applications*, *57*, 1–11. https://doi.org/10.1016/j.eswa.2016.03.041
7. R. Jindal and Shweta, "A Novel Method for Efficient Multi-Label Text Categorization of research articles," 2018 International Conference on Computing, Power and Communication Technologies (GUCON), 2018, pp. 333-336, doi: 10.1109/GUCON.2018.8674985.
8. G. I. Winata and M. L. Khodra, "Handling imbalanced dataset in multi-label text categorization using Bagging and Adaptive Boosting," *2015 International Conference on Electrical Engineering and Informatics (ICEEI)*, 2015, pp. 500-505, doi: 10.1109/ICEEI.2015.7352552.
9. G. Chen, D. Ye, Z. Xing, J. Chen and E. Cambria, "Ensemble application of convolutional and recurrent neural networks for multi-label text categorization," 2017 International Joint Conference on Neural Networks (IJCNN), 2017, pp. 2377-2383, doi: 10.1109/IJCNN.2017.7966144.
10. Z. He, J. Wu and P. Lv, "Label correlation mixture model for multi-label text categorization," 2014 IEEE Spoken Language Technology Workshop (SLT), 2014, pp. 83-88, doi: 10.1109/SLT.2014.7078554.
11. https://towardsdatascience.com/multi-label-text-classification-with-scikit-learn-30714b7819c5
12. https://www.section.io/engineering-education/multi-label-classification-with-scikit-multilearn/
13. https://www.coursera.org/learn/classification-vector-spaces-in-nlp?action=enroll
14. https://www.geeksforgeeks.org/understanding-tf-idf-term-frequency-inverse-document-frequency/
15. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html
16. https://scikit-learn.org/stable/modules/generated/sklearn.multiclass.OneVsRestClassifier.html
17. https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html