Capstone Project
Machine Learning Engineer Nanodegree

Foxan Ng
29 December, 2019

Auto Tagging Stack Overflow Questions

Domain Background

Stack Overflow is the popular go-to resource from programming newbies to professionals. There is a joke that a programmer's job is to search for relevant code snippets on Stack Overflow to copy and paste. Over 18 million questions have been asked on this platform, and It is currently the largest question and answer site for various topics in computer programming.

For a platform that has such a high volume of data, it is essential for the questions to be tagged with relevant topics, such as python, apache-spark, c#, etc, so that it allows more effective search and shows the information to users of interest. Correctly tagging Stack Overflow questions can reduce both time to search for an answer and time a question got answered, and thus **increase overall user engagement**.

Natural language processing[1] is a domain of study to program computers to process and analyze large amounts of natural language data. Particularly, **document classification**[2] is a subdomain which deals with problems of assigning a document to one or more classes or categories, which is similar to the problem we are trying to solve. A number of relevant techniques include: **tf-idf** (term frequency-inverse document frequency), **multiple-instance learning**, **latent semantic analysis**, etc.

Problem Statement

The goal of this project is to create an **automatic tagging system for Stack Overflow questions**, i.e. given any question text as input, it will be tagged with **minimum of one and maximum of five most relevant topics**, such as javascript, sql, c#, (i.e. the same rule currently

[1] Natural language processing
[2] Document classification

adopted by Stack Overflow). It will accept question title and question body in json format as input, and return a list of one to five tags as output.

For example:

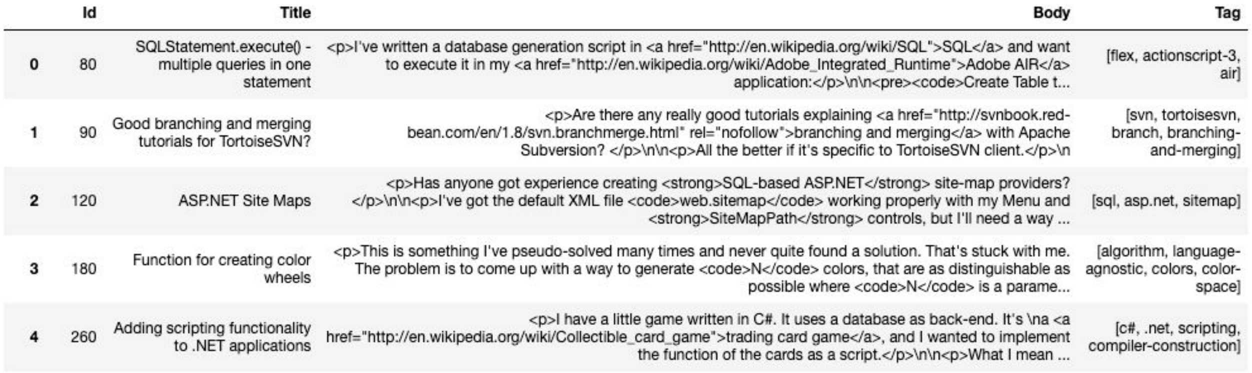
Input:

```
{
   "title": "ASP.NET Site Maps",
   "body": "Has anyone got experience creating SQL-based ASP.NET site-map providers?
I've got the default XML file web.sitemap 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
ASP.NET membership system as well."
}
```

Output:

```
{
   "tags": ["sql", "asp.net", "sitemap"]
}
```

Datasets and Inputs



	Id	Title	Body	Tag
0	80	SQLStatement.execute() - multiple queries in one statement	<p>I've written a database generation script in <a href="http://en.wikipedia.org/wiki/SQL">SQL</a> and want to execute it in my <a href="http://en.wikipedia.org/wiki/Adobe_Integrated_Runtime">Adobe AIR</a> application:</p>\n\n<pre><code>Create Table t...	[flex, actionscript-3, air]
1	90	Good branching and merging tutorials for TortoiseSVN?	<p>Are there any really good tutorials explaining <a href="http://svnbook.red-bean.com/en/1.8/svn.branchmerge.html" rel="nofollow">branching and merging</a> with Apache Subversion? </p>\n\n<p>All the better if it's specific to TortoiseSVN client.</p>\n	[svn, tortoisesvn, branch, branching-and-merging]
2	120	ASP.NET Site Maps	<p>Has anyone got experience creating <strong>SQL-based ASP.NET</strong> site-map providers?</p>\n\n<p>I've got the default XML file <code>web.sitemap</code> working properly with my Menu and <strong>SiteMapPath</strong> controls, but I'll need a way ...	[sql, asp.net, sitemap]
3	180	Function for creating color wheels	<p>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 parame...	[algorithm, language-agnostic, colors, color-space]
4	260	Adding scripting functionality to .NET applications	<p>I have a little game written in C#. It uses a database as back-end. It's \na <a href="http://en.wikipedia.org/wiki/Collectible_card_game">trading card game</a>, and I wanted to implement the function of the cards as a script.</p>\n\n<p>What I mean ...	[c#, .net, scripting, compiler-construction]

Sample data form StackSample: 10% of Stack Overflow Q&A

The dataset is made available by Stack Overflow, and it is released on the machine learning competition platform, Kaggle. It is named “StackSample: 10% of Stack Overflow Q&A”[3], which contains text from 10% of Stack Overflow questions and answers on programming topics.

In this project, I will focus on two of the files from the dataset, which is **Questions.csv** and **Tags.csv**. The remaining file is Answers.csv, which helps predicting tags for the question as well, but would defeat the purpose of predicting tags for questions so that we can show them to relevant users to answer, so I will ignore it.

[3] StackSample: 10% of Stack Overflow Q&A

Questions.csv contains 1.26 million questions, created from 2008 August to 2016 October. The columns I will focus on will be “Title” and “Body”, which corresponds to the question title and the actual content of the question. Tags.csv contains “Id” and “Tag” pair (one tag per row), which can be joined with the questions data using “Id”. There are more than 37,000 unique tags.

Solution Statement

Data preprocessing and cleansing will be done using pandas to cleanse columns such as “Title” and “Body”. Then I will use scikit-learn to extract useful features from the cleansed dataframe. As the problem is a multilabel classification problem, I plan to use **OneVsRestClassifier**[4] from scikit-learn to classify one tag versus all other tags at a time, and the model will be trained by using XGBoost as the machine learning algorithm, as it allows parallel processing and have effective tree pruning, which is great for a big dataset like this. I hope to create a machine learning model that will outperform the benchmark model with a lower hamming loss, which is described in the next section.

Benchmark Model

This problem has been tackled by others with a similar dataset[5], using OneVsRestClassifier and SGDClassifier with tf-idf as features. The author achieved a hamming loss of 0.00277914, when they limited the number of tags to the 500 most popular ones.

Evaluation Metrics

Hamming loss will be used as the evaluation metric, as it is a commonly used loss function in multilabel classification. It is basically the fraction of labels that are incorrectly predicted, so the smaller it is, the better the results are. Precision, recall and F1 score will also be used to support as they are evaluation metrics used to understand model performance in general.

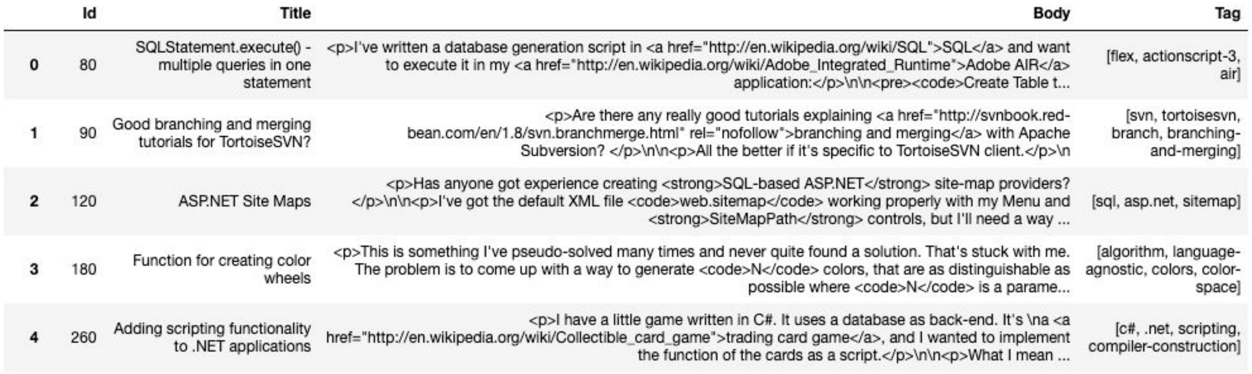
The definition of hamming loss[6] is the fraction of the wrong labels to the total number of labels,

$$\frac{1}{|N| \cdot |L|} \sum_{i=1}^{|N|} \sum_{j=1}^{|L|} \operatorname{xor}(y_{i,j}, z_{i,j})$$

where $y_{i,j}$ is the target and $z_{i,j}$ is the prediction.
This is a loss function, so the optimal value is zero.

[4] sklearn.multiclass.OneVsRestClassifier — scikit-learn 0.22 documentation
[5] Predicting Tags for the Questions in Stack Overflow
[6] Multi-label classification

Project Design



	Id	Title	Body	Tag
0	80	SQLStatement.execute() - multiple queries in one statement	<p>I've written a database generation script in <a href="http://en.wikipedia.org/wiki/SQL">SQL</a> and want to execute it in my <a href="http://en.wikipedia.org/wiki/Adobe_Integrated_Runtime">Adobe AIR</a> application:</p>\n\n<pre><code>Create Table t...	[flex, actionscript-3, air]
1	90	Good branching and merging tutorials for TortoiseSVN?	<p>Are there any really good tutorials explaining <a href="http://svnbook.red-bean.com/en/1.8/svn.branchmerge.html" rel="nofollow">branching and merging</a> with Apache Subversion? </p>\n\n<p>All the better if it's specific to TortoiseSVN client.</p>\n	[svn, tortoisesvn, branch, branching-and-merging]
2	120	ASP.NET Site Maps	<p>Has anyone got experience creating <strong>SQL-based ASP.NET</strong> site-map providers?</p>\n\n<p>I've got the default XML file <code>web.sitemap</code> working properly with my Menu and <strong>SiteMapPath</strong> controls, but I'll need a way ...	[sql, asp.net, sitemap]
3	180	Function for creating color wheels	<p>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 parame...	[algorithm, language-agnostic, colors, color-space]
4	260	Adding scripting functionality to .NET applications	<p>I have a little game written in C#. It uses a database as back-end. It's \na <a href="http://en.wikipedia.org/wiki/Collectible_card_game">trading card game</a>, and I wanted to implement the function of the cards as a script.</p>\n\n<p>What I mean ...	[c#, .net, scripting, compiler-construction]

Sample data form StackSample: 10% of Stack Overflow Q&A

Data preprocessing and cleansing is the first and foremost step for every successful machine learning project, which is no exception to this project as well.

As shown in the sample data above, the “Body” text are in HTML format, hence the HTML tags should be removed. We also need to be careful during text cleansing, as some symbols (such as .NET, C#, C++) are actually meaningful and should not be stripped away.

As title is much more representative in terms of understanding the question, I will give more weight to features derived from it during feature engineering. For example, for id 120, the title is “ASP.NET Site Maps”, it is clear that the question is related to “ASP.NET” and “sitemap”, which is reflected in the tags.

Some exploratory data analysis will also be useful to understand the data set. For example, there are over 37,000 unique tags in the data set. It will be very time-consuming and not practical to train a model on all available tags.

I will not make use of all the available tags, as data set like this tends to have a long tail distribution, i.e. a few popular tags will have a lot of questions (such as javascript, java, c#), each with over 1 million questions asked; while there will be far more unpopular or new tags (such as pyqtgraph, slurm, geom-bar), which have very few questions (a few hundred) asked.

For model training, as the problem is a multilabel classification problem, I will use OneVsRestClassifier from scikit-learn to classify one tag versus all other tags at a time, and the model will be trained by using XGBoost. In case the model performance is not as good as expected, I will also consider some other machine learning algorithms, such as logistic regression, deep neural network, etc.

I will use Amazon SageMaker for model training, as it would be very easy to scale up the machine in case the model training takes a lot of time. Hamming loss, which is a measure of the fraction of labels that are incorrectly predicted, will be used as the metric to evaluate model performance, as it is a commonly used for multilabel classification.

For future work, I will deploy the model using SageMaker, to allow users to submit a computer programming question to the API endpoint for auto tagging.