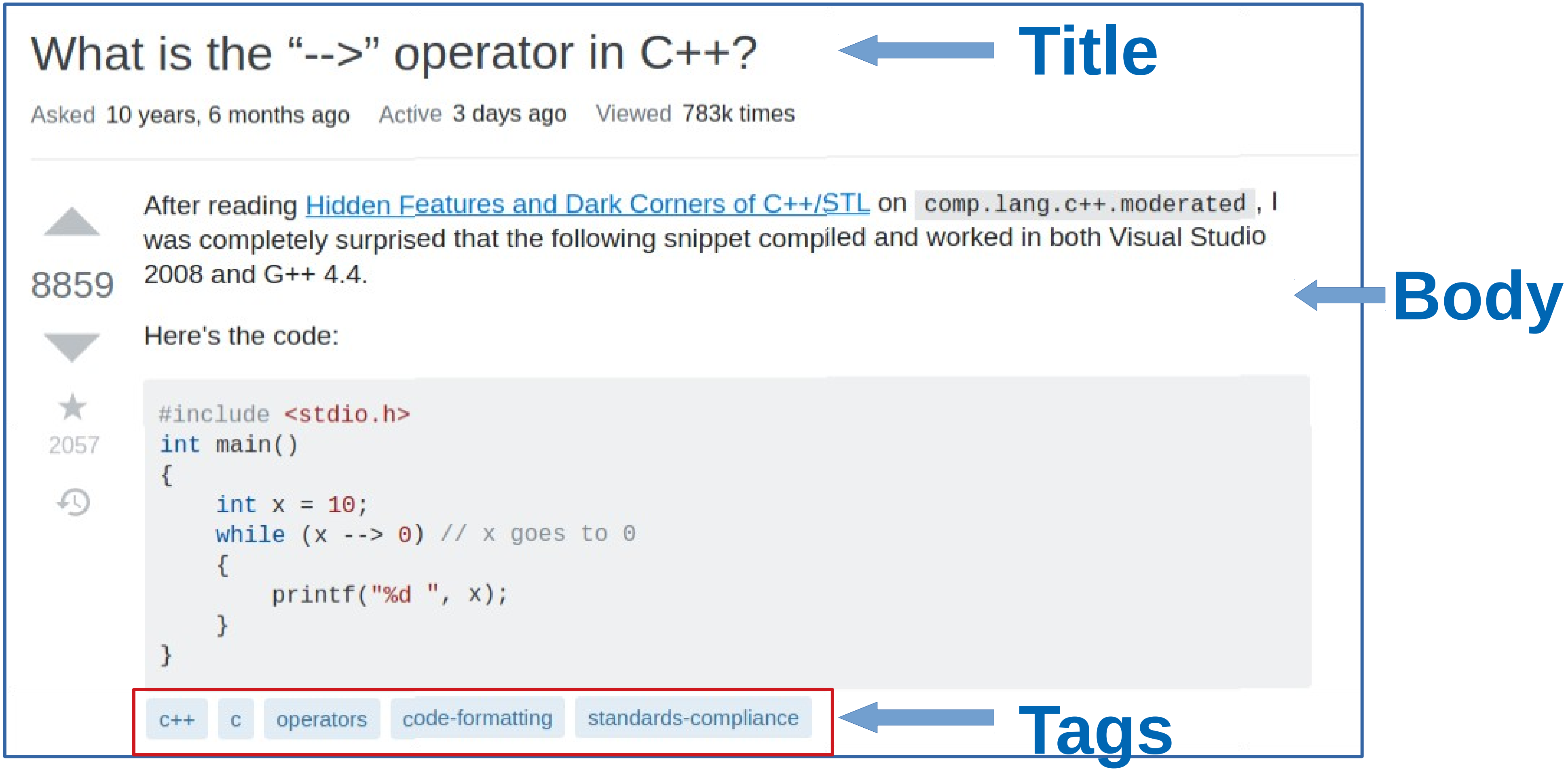
Stackoverflow Questions Classification

Business Problem

How to auto classify the thousands of questions posted every day?



What is the "-->" operator in C++?

Asked 10 years, 6 months ago Active 3 days ago Viewed 783k times

8859

2057

After reading Hidden Features and Dark Corners of C++/STL on `comp.lang.c++.moderated`, I was completely surprised that the following snippet compiled and worked in both Visual Studio 2008 and G++ 4.4.

Here's the code:

```
#include <stdio.h>
int main()
{
    int x = 10;
    while (x --> 0) // x goes to 0
    {
        printf("%d ", x);
    }
}
```

c++ c operators code-formatting standards-compliance

Title

Body

Tags

Business Problem

- **Multi-label Classification** problem
- Incorrect predictions lead to poor customer experience
- No latency requirements

Tools

- AWS – EC2 Extra large
- SQL – 6 million rows with over 8 GB of data
- NLP tools for stemming and stop words
- Remote Jupyter notebook
- Flask

Data Exploration

Schema

Id – Unique identifier for a Question

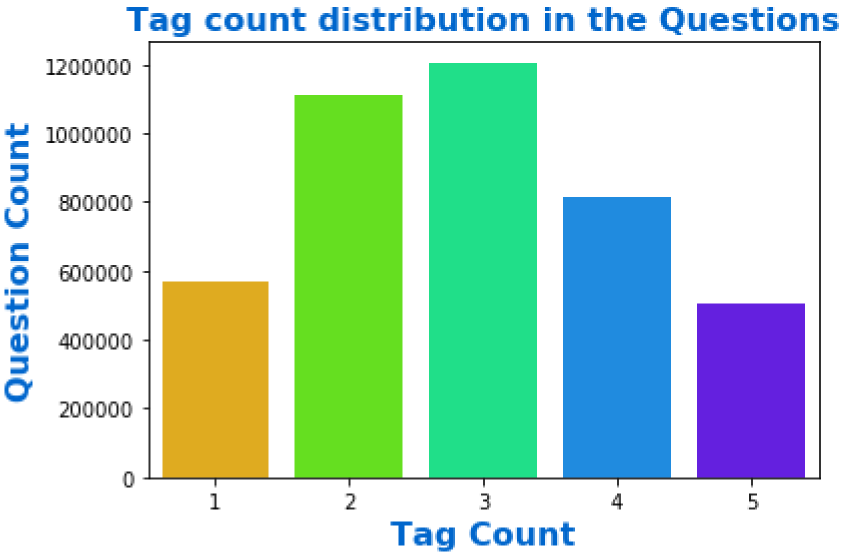
Title – Brief overview of Question

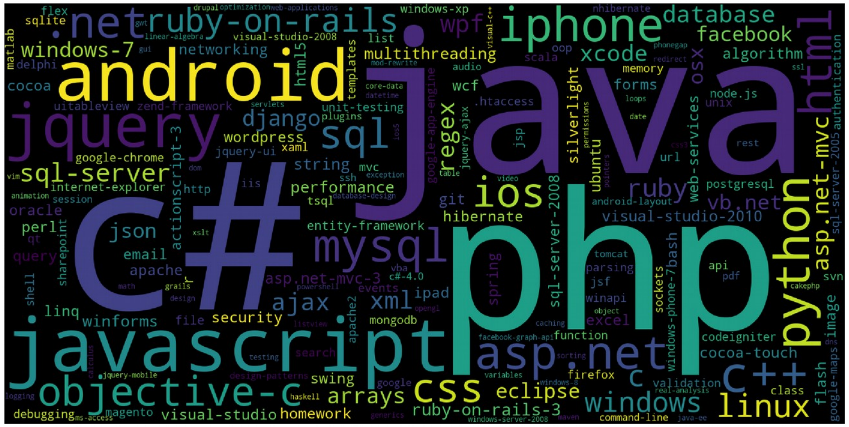
Body – Detailed description my contain Code

Tags – Output classes to be predicted

Data Exploration

- 20% Duplicate entries
- 42k unique tags & 4 million questions
- 2.7 Average tags per Question

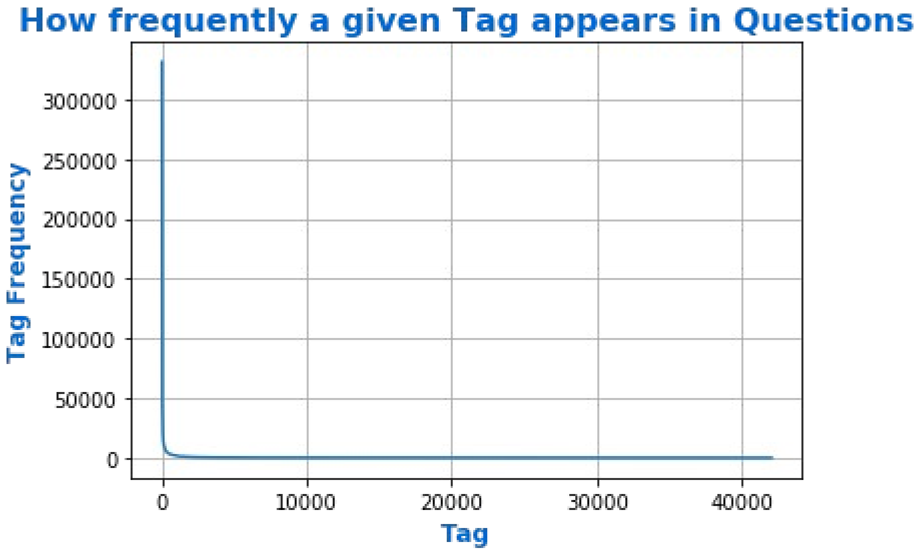


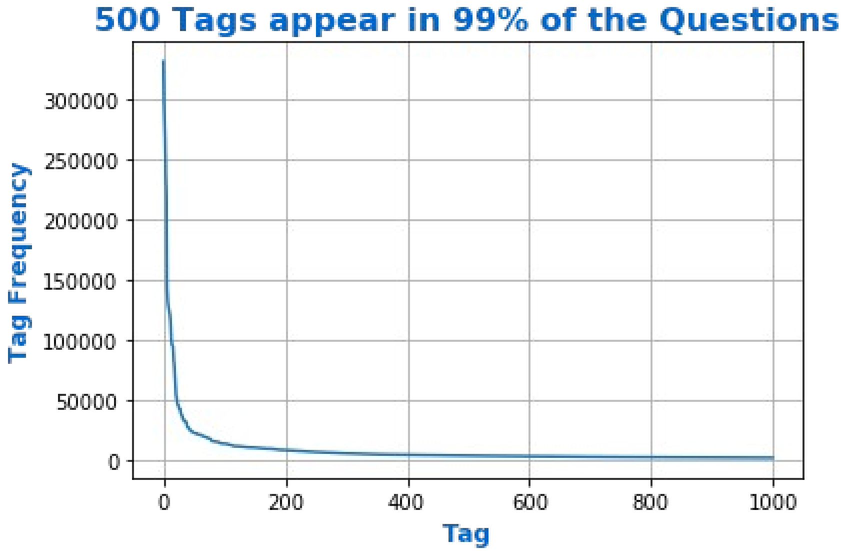


Feature Engineering

Pareto Distribution to the rescue!

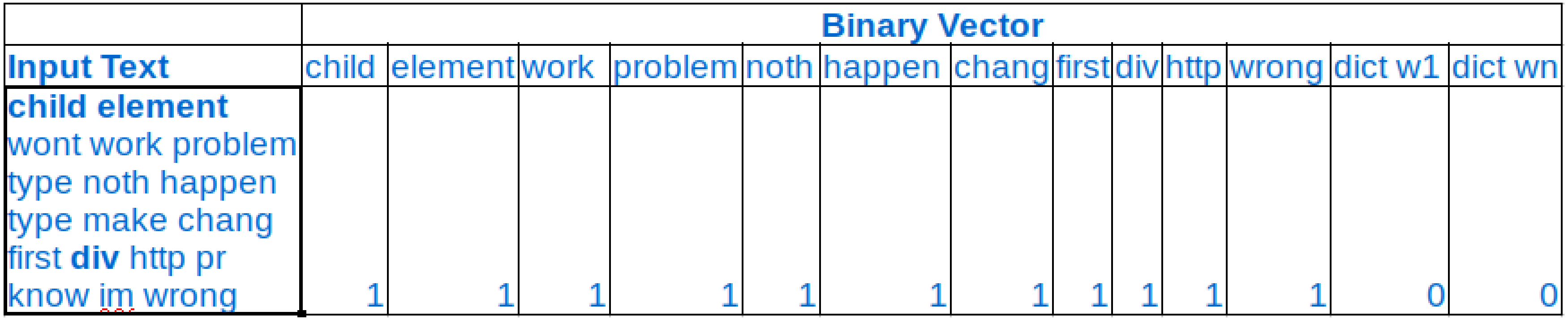
Very small number of Tags used across a majority of the Questions


How frequently a given Tag appears in Questions


500 Tags appear in 99% of the Questions

Input Transformation

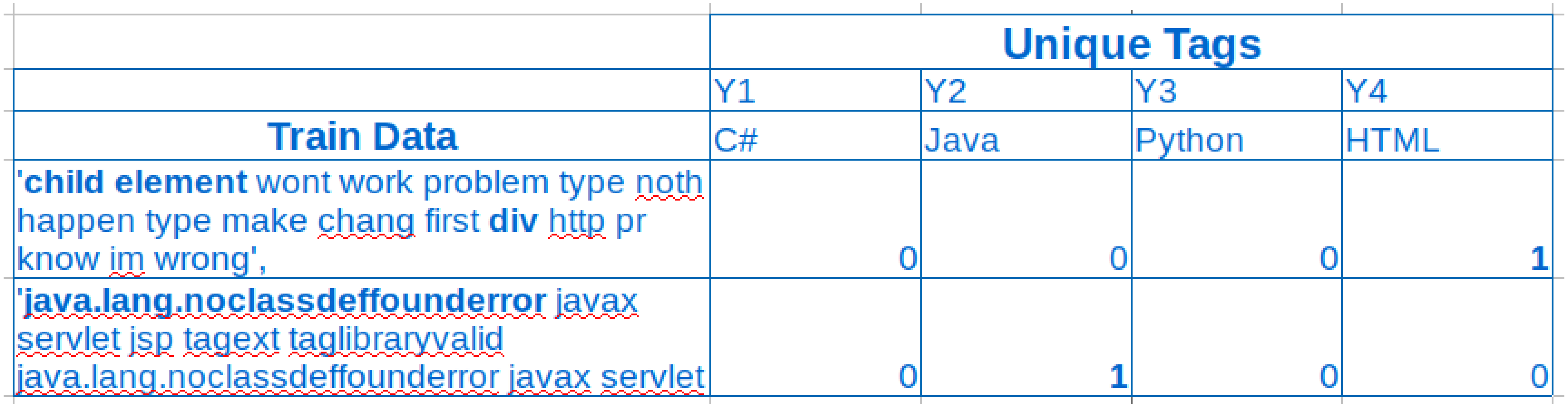
- Merge Title & Body columns to a single column
- Perform Stemming and remove stop words
- TFIDFVectorizer to transform input text to a Bag of Words



	Binary Vector												
Input Text	child	element	work	problem	noth	happen	chang	first	div	http	wrong	dict w1	dict wn
child element wont work problem type noth happen type make chang first **div** http pr know im wrong	1	1	1	1	1	1	1	1	1	1	1	0	0

Output Transformation

CountVectorizer to transform Tags into a 500 dimension **binary vector**.



	Unique Tags			
	Y1	Y2	Y3	Y4
Train Data	C#	Java	Python	HTML
'**child element** wont work problem type noth happen type make chang first **div** http pr know im wrong',	0	0	0	**1**
'**java.lang.noclassdeffounderror** javax servlet jsp tagext taglibraryvalid java.lang.noclassdeffounderror javax servlet	0	**1**	0	0

Model

OneVsRestClassifier

- Multilabel problem broken down into n binary classification problems
- Highly compute intensive
- > 6 hours to build model
- Wraps around other classification models like KNN & Logistic Regression

model = OneVsRestClassifier(**SGDClassifier**(max_iter=1000), n_jobs=4)

Scores

	KNN (n=20)	SGD Classifier
F1 score	0.3693	0.448
Recall	0.26	0.33
Precision	0.63	0.72
Accuracy	0.202	0.2361