Automatic Tag Prediction for Stack OverFlow

- Work by Rohit SVK, Akhil Batra, Nithiya Shree
Group - 35

Problem Statement

- Implementing an automated tag recommendation system for Question - Answers knowledge system like Stack Over Flow.
- Since a Question – Answer site may host millions of questions with tags and other data, this information can be used as a training and test dataset for approaches that automatically suggest tags for new questions based on the historical similarity of the old Question - Answers.

Dataset

- Dataset is initially downloaded and populated from Stack Overflow Data store in our database.
- Database Schema is mentioned in the link below.
  - Http://data.stackexchange.com/stackoverflow/query
- Our dataset consists of 1000 famous tags .
- Each tag has 60 questions corresponding to it.
- These tags and corresponding questions are used as both training data and test data .

Architecture
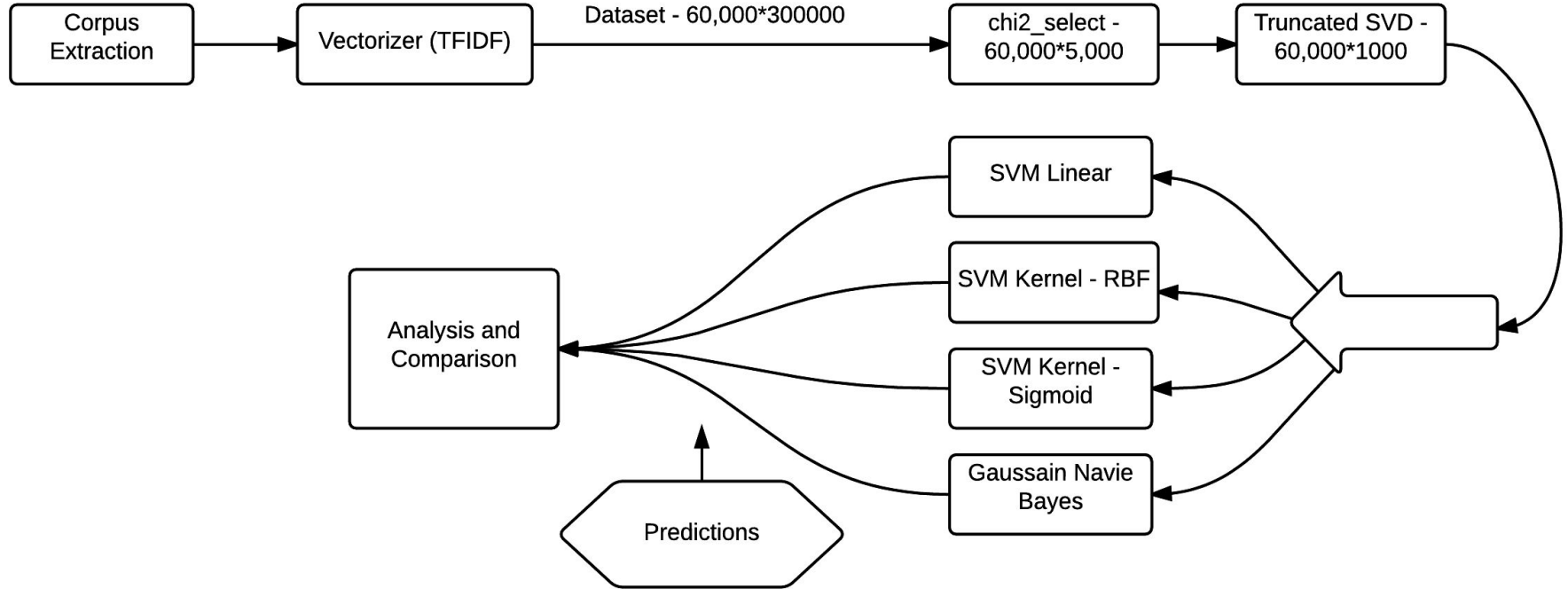
Action Plan

Data Extraction

Querying data from StackExchange dataexplorer.
- Considering 1000 popular tags (tags with most number of post counts).
- 60 questions for each tag.

d the following files.
wish to keep.

sults (2).csv [Origi...
created last time t...
5 3:34 PM

I want to save?

	Id	PostType	Accepte	ParentId	Creation	Deletion	Score	ViewCou	Body	OwnerUs	OwnerDi L
1	Id	PostType	Accepte	ParentId	Creation	Deletion	Score	ViewCou	Body	OwnerUs	OwnerDi L
2	3E+07	1			######		0	11	\<p\>I need to define a layout page. Currently I have index.cshtml and I am using	3E+06	
3	3E+07	1			######		1	45	\<p\>I have three assemblies. A main executable, a common library and an	2E+06	
4	3E+07	1			######		-1	16	\<p\>I have a UserMyAccount class in my model folder and i use this model in	5E+06	
5	3E+07	1			######		0	44	\<p\>Given a database schema with two tables: \<code\>Company\</code\> and	522663	!
6	3E+07	1	3E+07		######		0	15	\<p\>I'm using VB.Net, MVC 5, Visual Studio 2013. I have a question \<a	4E+06	
7	3E+07	1	3E+07		######		0	10	\<p\>I am looking to return a datatable to a gridview, however I dont need all of the	5E+06	
8	3E+07	1			######		2	34	\<p\>I have 3 Excel files to manipulate and I want to generate a single Excel file	5E+06	
9	3E+07	1			######		1	58	\<p\>Why when I run the following example do I only have the Parallel.ForEach	1E+06	
10	3E+07	1			######		1	13	\<p\>Is it possible to get a collection of ITypeSymbol's for the types exported by	3E+06	
11	3E+07	1			######		1	41	\<p\>Honestly, this is may be a dupe of \<a	9970	
12	3E+07	1	3E+07		######		3	63	\<p\>We have a generic \<code\>Job\</code\> class which have an abstract	1E+06	
13	3E+07	1	3E+07		######		1	20	\<p\>I'm trying to make a regex that would match a character at the beginning	128217	
14	3E+07	1	3E+07		######		0	17	\<p\>Is there a way to determine if the paragraph is a standard text or a	5E+06	
15	3E+07	1			######		0	19	\<p\>I have a method which uses IQueryable to get the value from another class,	5E+06	
16	3E+07	1			######		0	22	\<p\>I'm working with some unmanaged type libraries and reference files in a C#	5E+06	
17	3E+07	1			######		-6	29	\<p\>I need to come up with regex pattern for this password condition.	2E+06	
18	3E+07	1			######		0	7	\<p\>I'm using Mandrill Inbound Webhooks to call a method in my WCF API. The	5E+06	
19	3E+07	1			######		-1	21	\<p\>Lets say you have some System.Diagnostic.Trace.WriteLine statements in	9266	
20	3E+07	1			######		0	8	\<p\>I am trying to add a whole section to a particular occurence using GemBox.	4E+06	
21	3E+07	1			######		1	9	\<p\>I have a Visual Studio 2015 solution with 3 C# projects that depend on each	2E+06	
22	3E+07	1			######		0	8	\<p\>In form1 at the top where i declare the variables i have this line:\</p\>	5E+06	
23	3E+07	1			######		0	12	\<p\>This is as weird as it gets.\</p\>	2E+06	
24	3E+07	1			######		-2	29	\<p\>Hi I have a binary file that contains lots of resources and using C# I want to	5E+06	
25	3E+07	1			######		1	53	\<p\>This line:\</p\>	5E+06	
26	3E+07	1	3E+07		######		0	31	\<p\>I'm working on a website with about 500,000 users, the issue im having is	4E+06	
27	3E+07	1			######		0	11	\<p\>I am thinking whether the WeakReferences and WeakEvents are suitable in	5E+06	
28	3E+07	1			######		1	20	\<p\>I have some entities that were generated from a database and I wanted to	336102	
29	3E+07	1			######		1	10	\<p\>In Npgsql V2, I could use the following code to update a record, and return	2E+06	
30	3E+07	1			######		1	28	\<p\>I have a C# desktop application, which needs to make multiple simultaneous	5E+06	
31	3E+07	1			######		2	51	\<p\>I have 2 players, I want each of them to receive 26 random out of this 52	5E+06	
32	3E+07	1			######		0	10	\<p\>I have a string from server that contains content string with html tags and	3E+06	
33	3E+07	1			######		0	20	\<p\>Using VS 2015 Community (but this is also happening in VS2013)	1E+06	
34	3E+07	1			######		0	11	\<p\>I want to upload image on picasa web album.	3E+06	
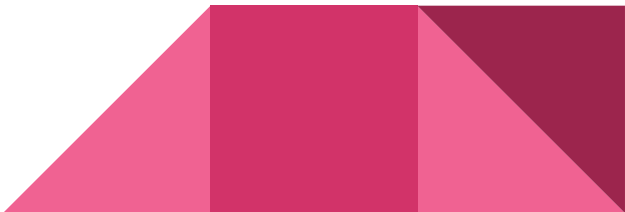
Extraction of Feature Vector

Conversion of each question into a feature vector is done using **'tf-idf vectorizer'** .

- tf-idf stands for term frequency - inverse document frequency.
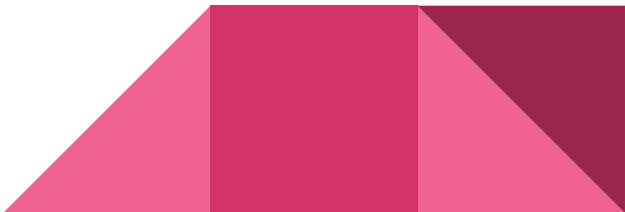- This algorithm is also used for stop-words filtering.

Extraction of Feature Vector

tf-idf vectorizer

- **Term frequency** is to use the raw frequency of a term in a document i.e, the number of times that term t occurs in document d .

$$\text{tf}(t, d) = 0.5 + \frac{0.5 \times f_{t,d}}{\max\{f_{t,d} : t \in d\}}$$

- where $f_{t,d}$ is the raw frequency of t .
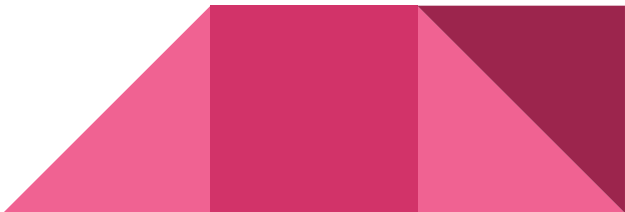
Extraction of Feature Vector

tf-idf vectorizer
- A high weight in tf-idf is reached by a high term frequency and a low document frequency of a term in the whole collection of documents.

$$\text{tfidf}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D)$$

- This weight also tends to filter out common terms.

Extraction of K best features

The algorithm used to get the best k features from a feature vector is **Chi-squared distribution**.

- Chi-squared distribution with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables.

$$Q = \sum_{i=1}^{k} Z_i^2,$$

  - where $z_1, ..., z_k$ are independent and standard normal random variables.
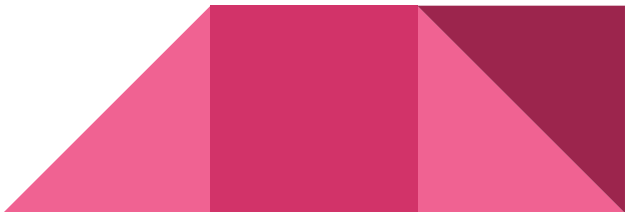
Dimensionality Reduction

Dimensionality reduction is done using **Truncated SVD algorithm**.

- SVD stands for **Singular value Decomposition** is a factorization of real or complex matrix.

$$M = U\Sigma V^*$$
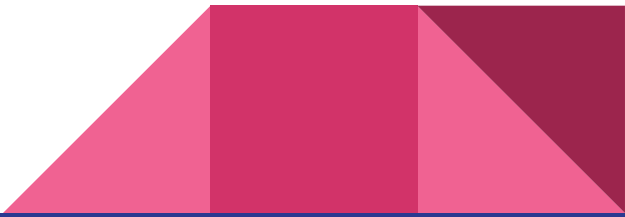
- where **U** is a m X m real or complex unitary matrix.
- **Σ** is a m X n rectangular diagonal matrix with non-negative real numbers on the diagonal.
- **V*** is a n X n real or complex unitary matrix.

Dimensionality Reduction

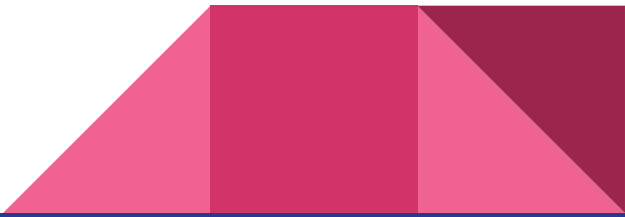- In a truncated SVD , only the *t* column vectors of *U* and *t* row vectors of V* corresponding to the largest *t* largest singular values Σ_t are calculated.

$$\tilde{\mathbf{M}} = \mathbf{U}_t \mathbf{\Sigma}_t \mathbf{V}_t^*$$

- The rest values are discarded.
- Thus, this is more quicker and more economical than the compact SVD if *t << r* .
- The matrix U_t is thus m X t, Σ_t is t X t diagonal, and V_t^* is t X n.
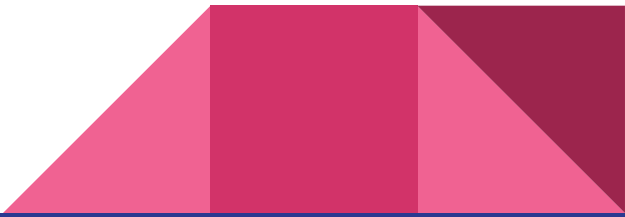
Training the data

For training the data, five different approaches have been employed :

- Linear SVM
- SVM with RBF Kernel
- SVM with Sigmoid kernel
- Multinomial Naive Bayes Classifier.
- Gaussian Naive Bayes Classifier.

Training the data

Linear SVM

- **Support Vector Machines** are supervised learning models with associated learning algorithms that analyze the data and recognize patterns which are useful for classification.
- SVM training algorithm builds a model that assigns new data points into one category or the other making it a non-probabilistic binary linear classifier.
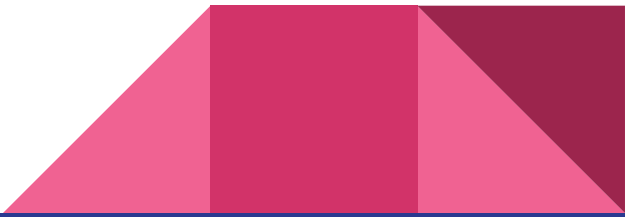- It has given an accuracy of 31.4 percent

Training the data

SVM with RBF Kernel
- RBF is **Radial Basis Function kernel**, is a popular kernel function used in various kernelized learning algorithms.
- The feature space of the kernel has infinite number of dimensions.
- The RBF kernel on 2 samples, is defined as

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||^2}{2\sigma^2}\right)$$

  ○ $||\mathbf{x} - \mathbf{x}'||^2$ is the Euclidian distance between the 2 samples
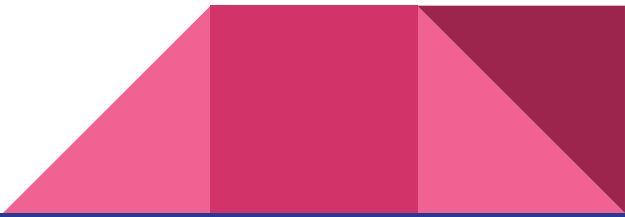 ○ $\square$ is a free parameter
  ○ Accuracy = 39

Training the data

SVM with Sigmoid Kernel

- It uses the following kernel function

$$\tanh(\gamma \langle x, x' \rangle + r)$$

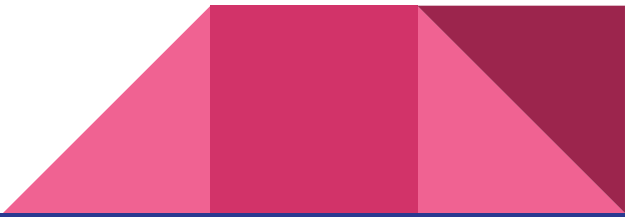- It gives us better accuracies.
- Accuracy - 42.1

Training the data

Gaussian Naive Bayes Classifier

- Gaussian Naive Bayes classifier assumes that the continuous values are associated with each class are distributed according to a Gaussian distribution.

$$p(x = v|c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-\frac{(v-\mu_c)^2}{2\sigma_c^2}}$$

- where is μ_c the mean of values in **x** associated with class c.
- σ_c^2 is variance of values of **x** associated with class c.
- Accuracy - 34

Analysis on mean accuracy based on 5 fold cross validation
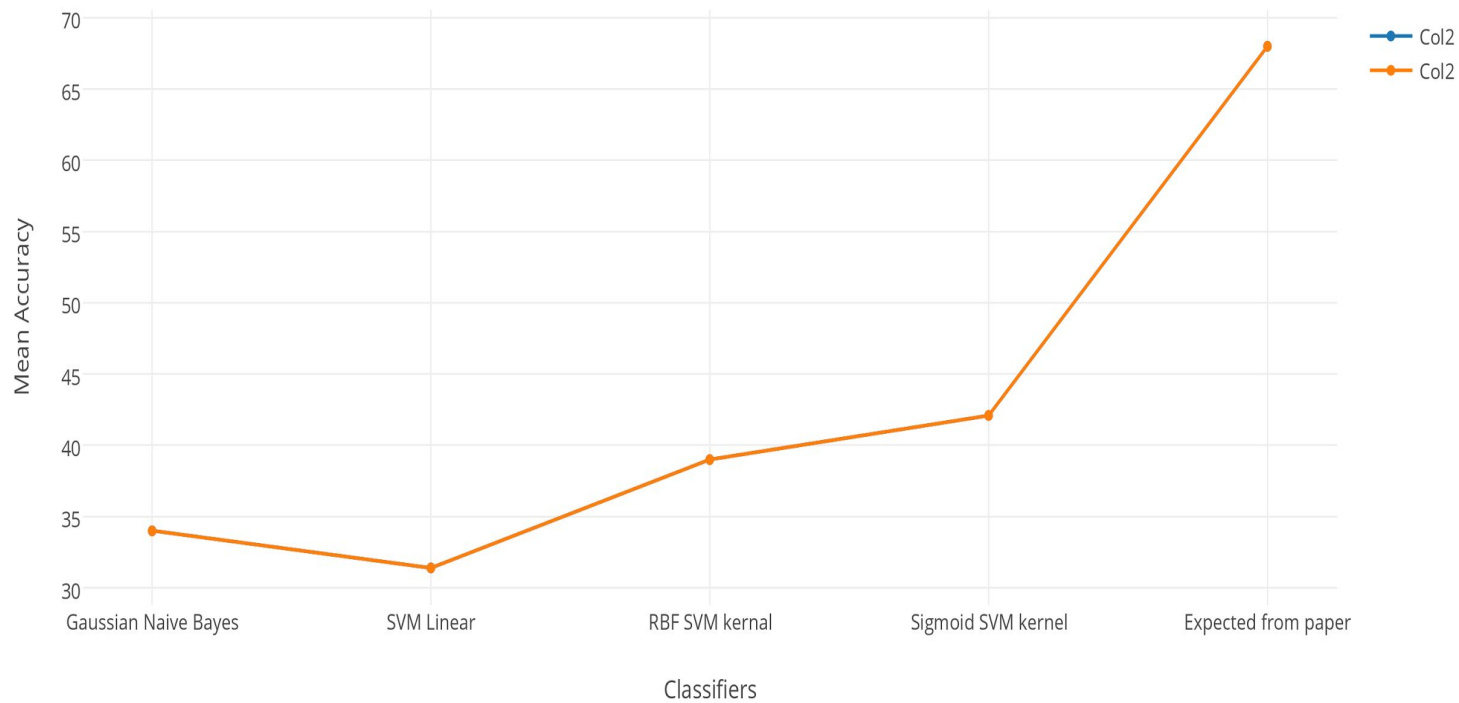
Cross Validation and analysis

- We have taken the test data in 0.16 ratio to that of training data.
- Accuracies Table

Classifier	Accuracy
Linear SVM	31.4
RBF Kernel SVM	39
Sigmoid kernel SVM	42.1
Gaussian Naive Bayes	34

Reference

- " A Discriminative Model Approach for Suggesting Tags Automatically for Stack Overflow Questions" –Avigit K. Saha, Ripon K. Saha, Kevin A. Schneider
  - http://www.cs.usask.ca/~kas/Publications_files/msr13-id175-p-16622-preprint.pdf
  - Conference – 10 th Working Conference on Mining Software Repositories. Mining Challenge – IEEE - 2013
- Will take hints from below paper and adding our approach to improve accuracy for the above approach.
  - " EnTagRec: An Enhanced Tag Recommendation System for Software Information Sites"
  - Conference – ICSME 2014