WEB PAGE CLASSIFICATON

100042773 – Can Okan TAŞKIRAN

100042970 – Alperen KÖYLÜ

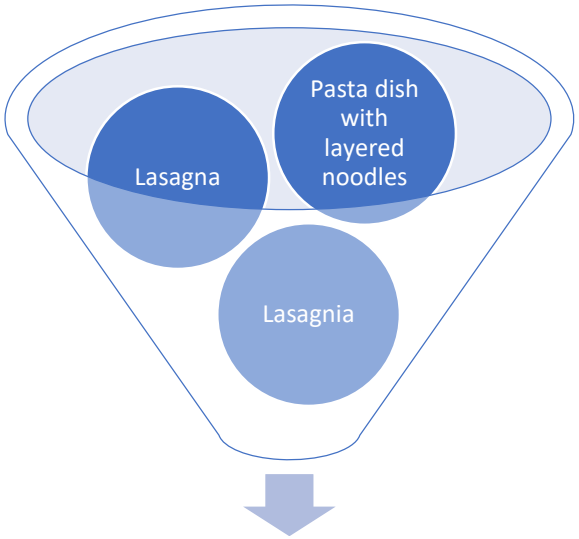
1. Introduction

What is the problem ?

- Past decade we have witnessed an explosive growth on the Internet, with millions of web pages on millons of topic.
- But the Internet is not a suitable tool for locating or organizing the mass of information.
- Tools like search engines assist users in locating information on the Internet.
- They perform excellently in locating but provide limited ability in organizing the web pages.

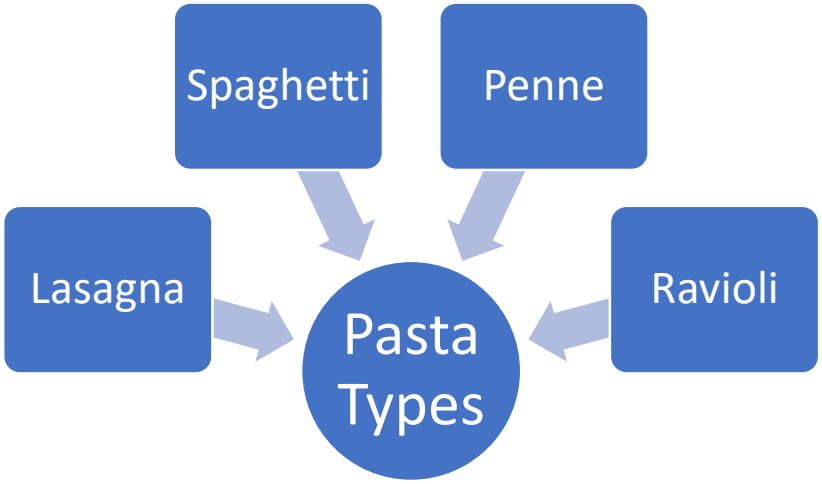
Difference between locating and classification

What Google do …



Lasagna

What we are trying …



What is web page classification ?

- Web page classification, aka categorization, defined as the task of determining whether a web page belongs to a category or categories.
- Let **C = {c1, …, cK}** be a set of predefined categories, **D = {d1, …, dN}** be a set of web pages to be classified, and **A = D×C** be a decision matrix.

A diagram for showing the decision matrix

Web Pages	Categories				
	C1	…	Cj	…	Ck
D1	A11	…	A1j	…	A1k
⋮	…	…	…	…	…
Di	Ai1	…	Aij	…	Aik
⋮	…	…	…	…	…
Dn	An1	…	Anj	…	Ank

2. Experimental Setup

Our working space

- 7 classes
  - Basic materials sector
  - Energy sector
  - Financial sector
  - Healthcare sector
  - Technology sector
  - Transportation sector
  - Utilities sector

- 4581 individual html files
- More than 1461640 dimensions in matrix (no reduction or selection algorithm applied on)

Pure HTML view

```
<!DOCTYPE HTML PUBLIC "-//IETF//DTD HTML 2.0//EN">
<HTML>
<HEAD>
<TITLE>Kaydon Bearings</TITLE>
<META  name="description" content="Kaydon bearings are the designer's choice for difficult and sensitive applications in material handling, construction equipment, robotic and other key industrial positions.">
<META  name="keywords" content="Bearings, Reali-Slim, Thin Section, Ball, Roller, Custom, Assembly, Hybrid, Stainless-steel, Vacuum, Robots, Large bore, Turntable, Slewing Rings, Worm Drive, Construction Equipment, Manlifts">

</HEAD><body background="/kaydon/backgrounds/bgmenu.gif">
<table width="500">
   <td width="160" align="left" valign="top">
<center><A HREF="/kaydon/default.htm"><IMG SRC="/kaydon/graphics/logo100.gif" border=none></A>
<IMG SRC="/kaydon/bearings/graphics/bearings3.jpg" width="100">
<A HREF="/kaydon/bearings/default.htm"><IMG SRC="/kaydon/bearings/graphics/bearings.gif" border=none></A></center>
<P>
<B><center><A HREF="/kaydon/bearings/index.htm">Bearings Site Index</A></center></B>
  </td>
  <td width=340>
<center><IMG SRC="/kaydon/bearings/graphics/bearingslogo.gif"></center>
    <table width=100% border=1>
    <td>
<center><IMG SRC="/kaydon/bearings/graphics/bearingspix.jpeg"></center>
    </td>
    <td>
<center><B>2860 McCracken St<BR>
Muskegon, MI 49441<BR>
(616) 755-3741<BR>
 FAX: (616) 759-4102 <BR>
Telex: 228436<BR>
<BR>
<hr>
Engineer Services <BR>
Hotline <BR>
(800) 514-3066
</center>
    </td>
  </table>
<center>
<P>
<B><A HREF="/kaydon/bearings/cap.htm">Capabilities</A> |
<A HREF="/kaydon/bearings/app.htm">Applications</A> |
<A HREF="/kaydon/bearings/products.htm">Products</A> |
<A HREF="/kaydon/bearings/interface.htm">Interface</A> |
<A HREF="/kaydon/bearings/faqs.htm">FAQs</A></B>
</center>
<P>
<hr>
<B>Kaydon bearings</B> are the designer's choice for difficult and sensitive applications in material handling, construction equipment, robotic and other key industrial positions.  Carefully engineered solutions are available promptly and completely.  Prototype
<center>
<P>
Other Kaydon Divisions at your service:<BR>

<A HREF="/kaydon/cooper/default.htm">Cooper Split Roller Bearings</A> |
<A HREF="/kaydon/electro/default.htm">Electro-Tec</A> |
<A HREF="/kaydon/filtration/default.htm">Filtration</A> |
<A HREF="/kaydon/fluid/default.htm">Fluid Power</A> |
<A HREF="/kaydon/iti/default.htm">Industrial Tectonics Inc</A> |
<A HREF="/kaydon/ringseal/default.htm">Kaydon Ring & Seal</A>
```

How to interpret HTML files ?

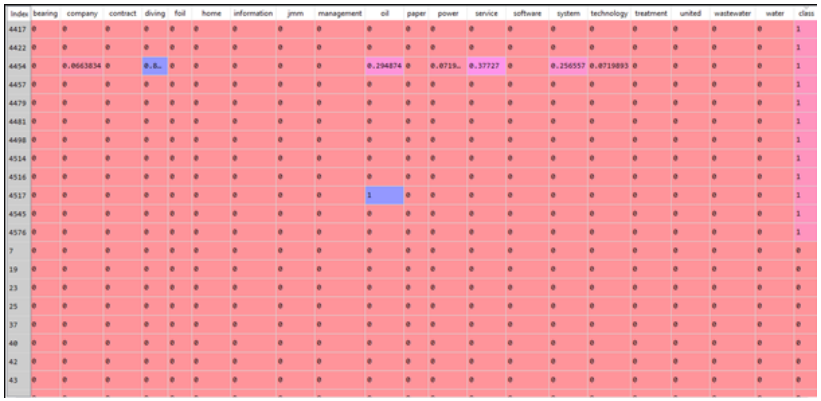
- External writed C# code that removes html tags and styles in pages, and gives you only texts.

- BeautifulSoup4
  - P tags
  - Meta tags
  - All tags

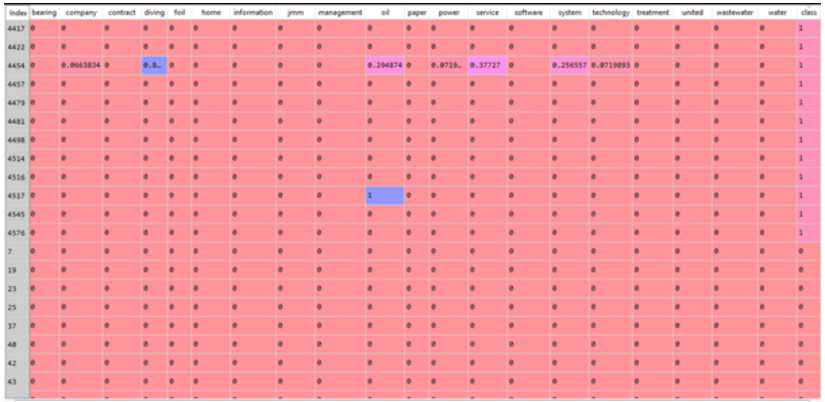
- P.S. : By the way cross-validation applied all algorithms.

Challenges

Sparse Matrix

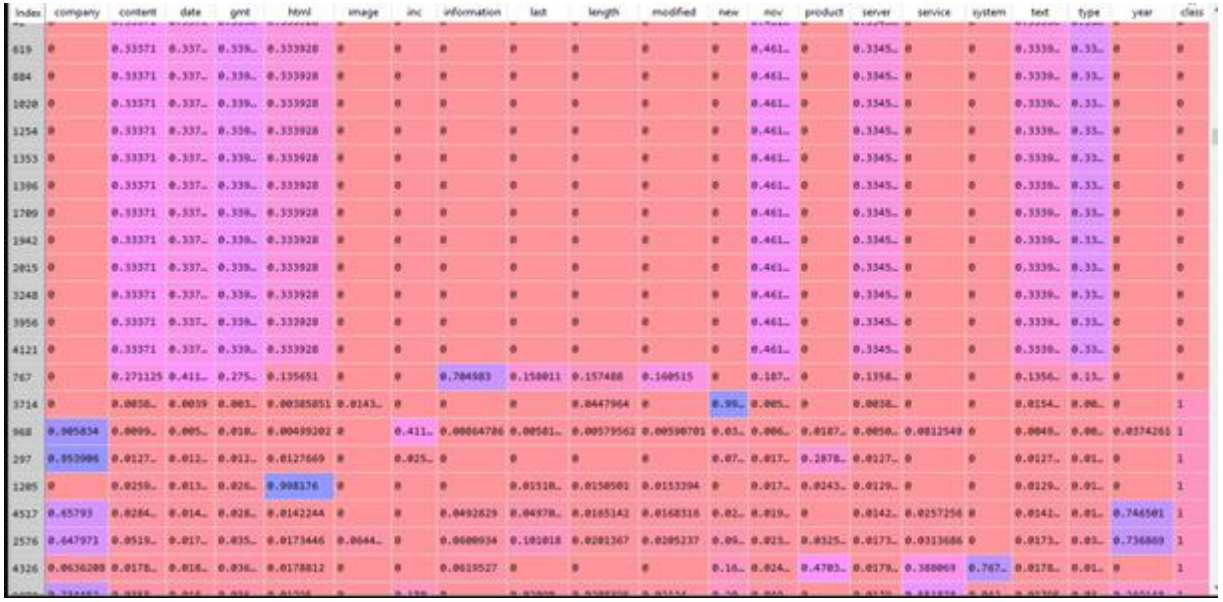


Only P Tags



Solution

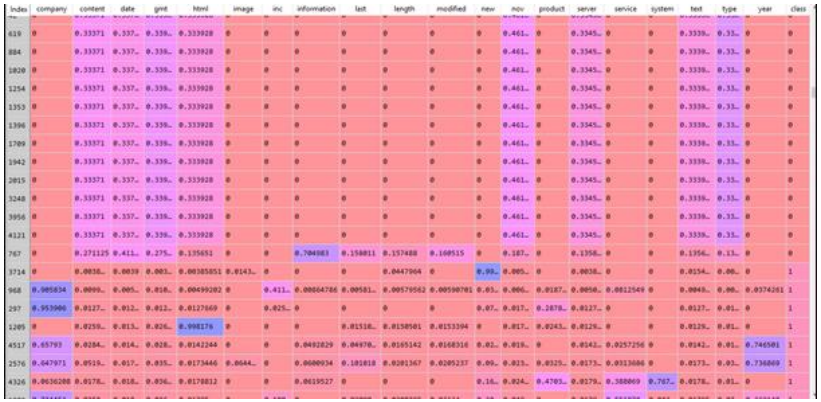
- Final matrix that all learning algorithm applied on



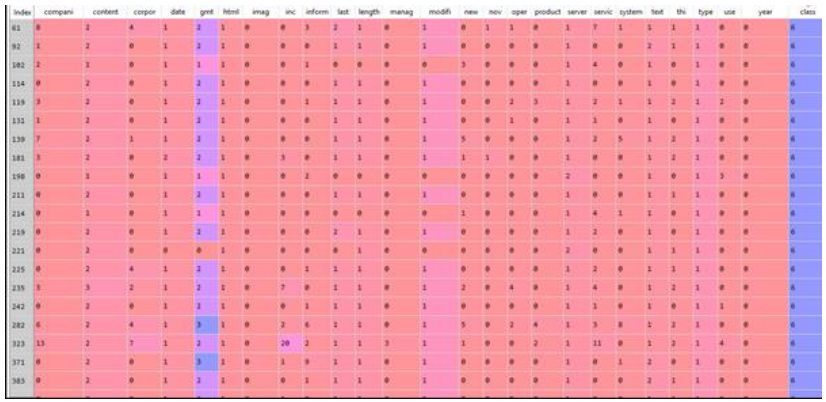
Index	company	content	date	gmt	html	image	inc	information	last	length	modified	new	nov	product	server	service	system	text	type	year	class
619	0	0.33371	0.337…	0.339…	0.333928	0	0	0	0	0	0	0	0.461…	0	0.3345…	0	0	0.3338…	0.33…	0	0
884	0	0.33371	0.337…	0.339…	0.333928	0	0	0	0	0	0	0	0.461…	0	0.3345…	0	0	0.3338…	0.33…	0	0
1020	0	0.33371	0.337…	0.339…	0.333928	0	0	0	0	0	0	0	0.461…	0	0.3345…	0	0	0.3338…	0.33…	0	0
1254	0	0.33371	0.337…	0.339…	0.333928	0	0	0	0	0	0	0	0.461…	0	0.3345…	0	0	0.3338…	0.33…	0	0
1353	0	0.33371	0.337…	0.339…	0.333928	0	0	0	0	0	0	0	0.461…	0	0.3345…	0	0	0.3338…	0.33…	0	0
1396	0	0.33371	0.337…	0.339…	0.333928	0	0	0	0	0	0	0	0.461…	0	0.3345…	0	0	0.3338…	0.33…	0	0
1709	0	0.33371	0.337…	0.339…	0.333928	0	0	0	0	0	0	0	0.461…	0	0.3345…	0	0	0.3338…	0.33…	0	0
1942	0	0.33371	0.337…	0.339…	0.333928	0	0	0	0	0	0	0	0.461…	0	0.3345…	0	0	0.3338…	0.33…	0	0
2015	0	0.33371	0.337…	0.339…	0.333928	0	0	0	0	0	0	0	0.461…	0	0.3345…	0	0	0.3338…	0.33…	0	0
3248	0	0.33371	0.337…	0.339…	0.333928	0	0	0	0	0	0	0	0.461…	0	0.3345…	0	0	0.3338…	0.33…	0	0
3956	0	0.33371	0.337…	0.339…	0.333928	0	0	0	0	0	0	0	0.461…	0	0.3345…	0	0	0.3338…	0.33…	0	0
4121	0	0.33371	0.337…	0.339…	0.333928	0	0	0	0	0	0	0	0.461…	0	0.3345…	0	0	0.3338…	0.33…	0	0
767	0	0.271125	0.411…	0.275…	0.135651	0	0	0.784583	0.158011	0.157488	0.160515	0	0.187…	0	0.1358…	0	0	0.1356…	0.13…	0	0
1714	0	0.0036…	0.0039	0.003…	0.00385851	0.0143…	0	0	0	0.0447964	0	0.99…	0.005…	0	0.0038…	0	0	0.0154…	0.00…	0	1
968	0.905834	0.0099…	0.005…	0.010…	0.00499202	0	0.411…	0.00064786	0.00581…	0.00579562	0.00590701	0.03…	0.006…	0.0187…	0.0050…	0.0012549	0	0.0048…	0.00…	0.0374261	1
297	0.953906	0.0127…	0.012…	0.012…	0.0127669	0	0.025…	0	0	0	0	0.07…	0.017…	0.2878…	0.0127…	0	0	0.0127…	0.01…	0	1
1205	0	0.0259…	0.013…	0.026…	0.998176	0	0	0	0.01518…	0.0158501	0.0153394	0	0.017…	0.0243…	0.0129…	0	0	0.0129…	0.01…	0	1
4517	0.65793	0.0284…	0.014…	0.028…	0.0142244	0	0	0.0492829	0.04978…	0.0165142	0.0168316	0.02…	0.019…	0	0.0142…	0.0257256	0	0.0141…	0.01…	0.746501	1
2576	0.647973	0.0519…	0.017…	0.035…	0.0173446	0.0644…	0	0.0600934	0.101018	0.0201367	0.0205237	0.09…	0.023…	0.0325…	0.0173…	0.0313686	0	0.0173…	0.01…	0.736869	1
4326	0.0636289	0.0178…	0.018…	0.036…	0.0178812	0	0	0.0619527	0	0	0	0.16…	0.024…	0.4703…	0.0179…	0.388069	0.767…	0.0178…	0.01…	0	1

Lemmitazing vs Stemming

Lemmitazing



Stemming



P.S. : Stopwords and regex are eliminated before we process on data on every algorithm.

Future represantation

- TF-IDF
- Count Vectorizer + TF-IDF

- N Gram
  - BiGram

TOP 5 COUNT+TF-IDF

MAX_FEATURES	MIN_DF	MAX_DF	DTREE
100	600	250	0.713787086
90	600	90	0.712041885
100	600	310	0.710296684
90	500	230	0.709424084
90	500	310	0.709424084

MAX_FEATURES	MIN_DF	MAX_DF	KNN
70	750	90	0.620418848
80	750	90	0.620418848
90	750	90	0.620418848
100	750	90	0.620418848
60	750	90	0.620418848

MAX_FEATURES	MIN_DF	MAX_DF	BAYES
90	450	110	0.533158813
90	450	510	0.533158813
90	450	150	0.533158813
90	450	310	0.533158813
90	450	330	0.533158813

Obtained from **accuracy_VECTORIZER&TFIDF.csv** file and this file has been created after 23 hours of live working code.

TOP 5 TF-IDF

MAX_FEATURES	MIN_DF	MAX_DF	DTREE
140	450	70	0.725130890
130	450	90	0.721640489
140	450	110	0.721640489
140	450	250	0.721640489
140	450	370	0.719895288

MAX_FEATURES	MIN_DF	MAX_DF	KNN
150	450	90	0.623909250
140	450	90	0.623909250
70	750	90	0.620418848
100	750	90	0.620418848
90	750	90	0.620418848

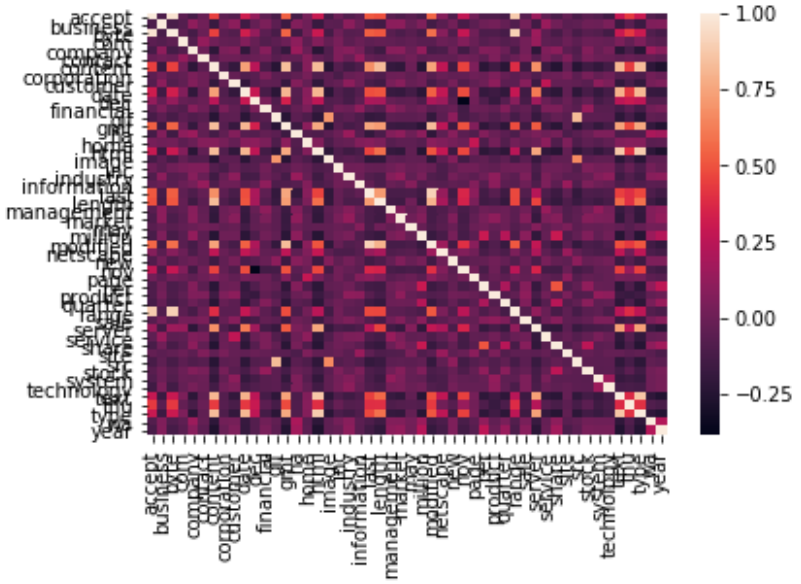
MAX_FEATURES	MIN_DF	MAX_DF	BAYES
130	450	270	0.538394415
130	450	110	0.538394415
130	450	470	0.538394415
130	450	390	0.538394415
130	450	370	0.538394415

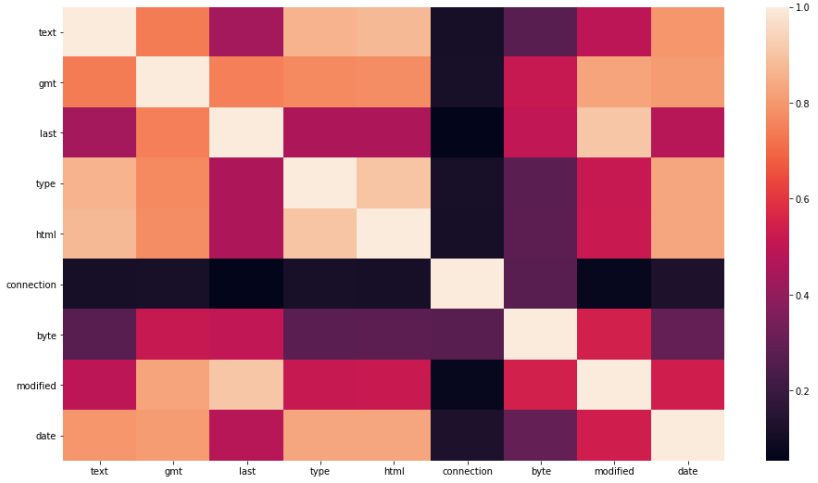
Obtained from **accuracy_TFIDFonly.csv** file and this file has been created after 11 hours of live working code.

Features Selection and Feature Extraction

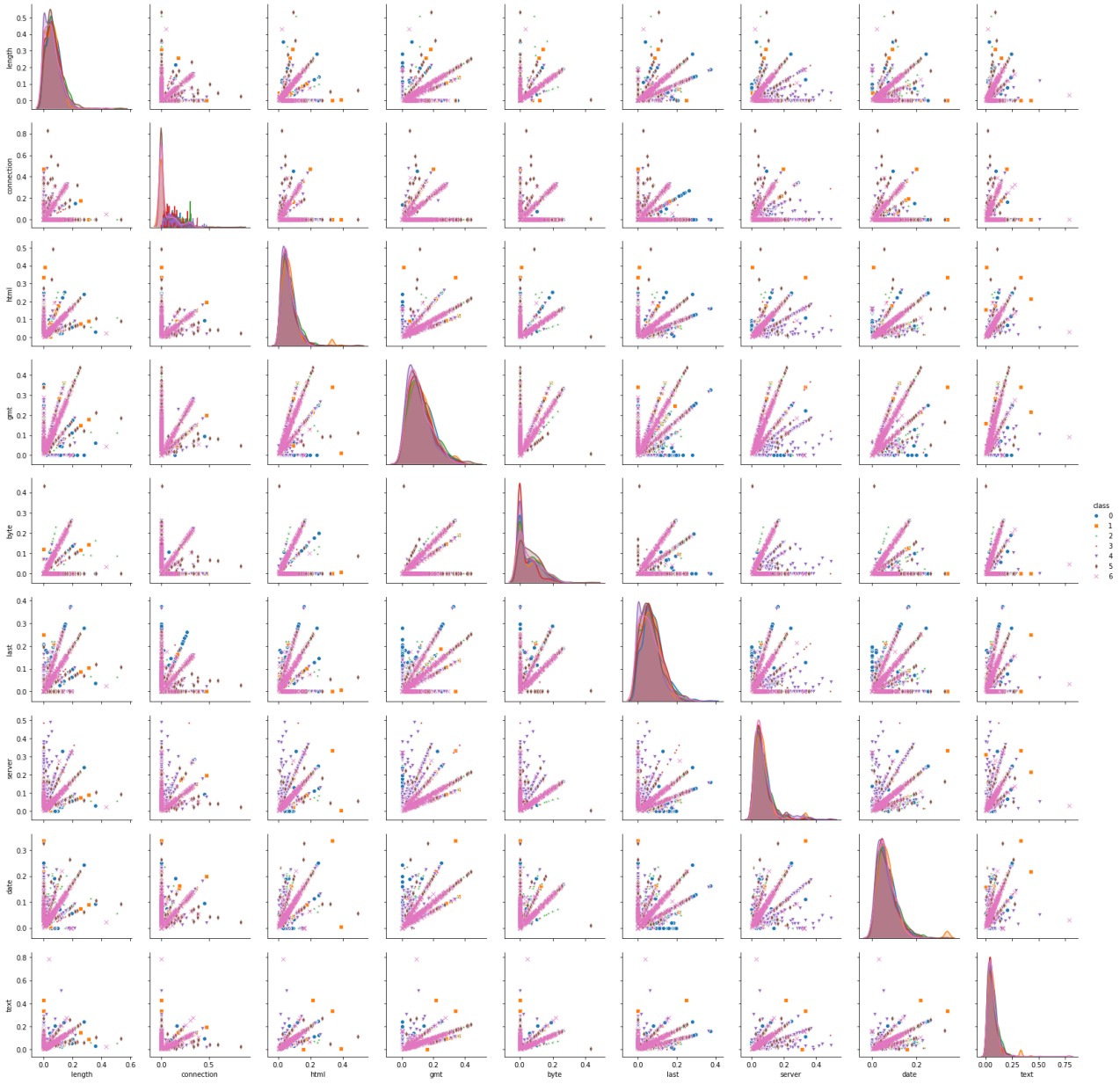
- Feature Selection
  - Filter Methods
    - Correlation
    - Information Gain (Mutual Inf.)
    - Relief
  - Wrapper Methods
    - Sequential Feature Selection
- Feature Reduction
  - Principal Component Analysis (PCA)
  - Linear Discriminant Analysis (LDA)
  - Latent Semantic Analysis (LSA)

Correlation matrix





Cross of Features and Correlation



Latent Semantic Analysis (SVD)

- Unsupervised
- Decompose it into 3 seperated matrix.
- Term-Document Matrix

$$A = U S V^T$$

$$A_{m \times n} = U_{m \times m} S_{m \times n} V^T_{n \times n}$$

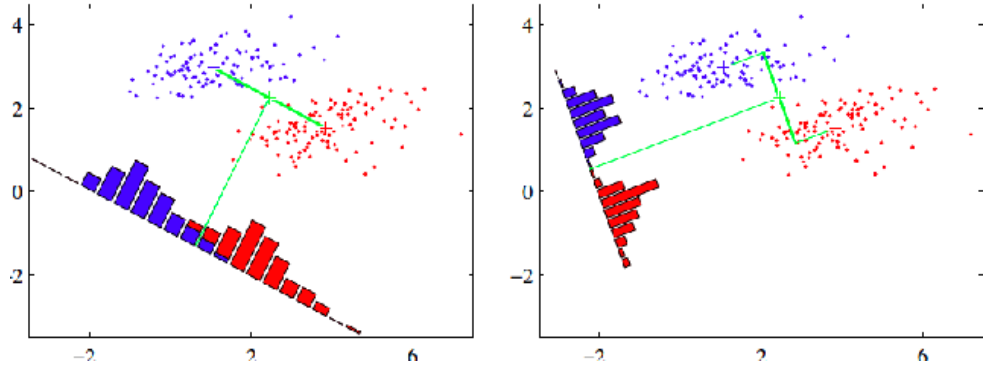
$$\underset{m \times n}{\mathrm{A}} = \underset{m \times m}{\mathrm{U}} \; \underset{m \times n}{\mathrm{S}} \; \underset{n \times n}{\mathrm{V^T}}$$

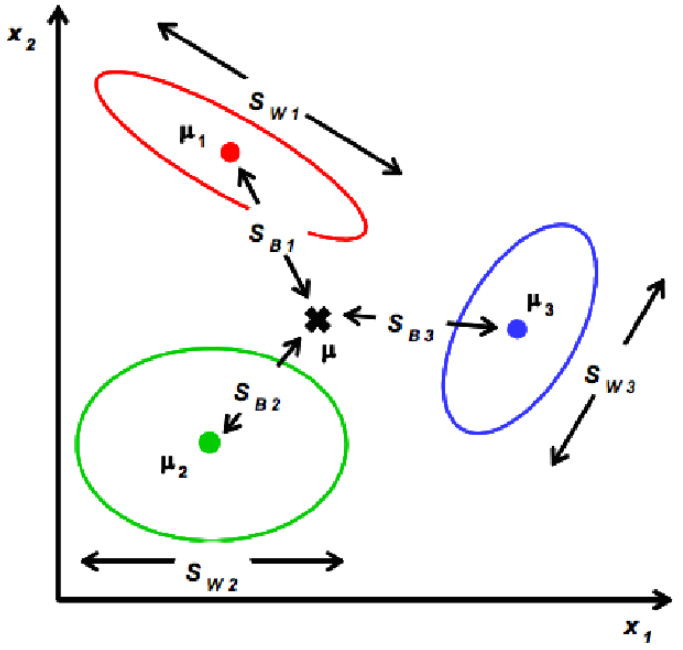
$$\begin{pmatrix} x_{11} & x_{12} & & x_{1n} \\ & & \ddots & \\ x_{m1} & & & x_{mn} \end{pmatrix} = \begin{pmatrix} u_{11} & & u_{m1} \\ & \ddots & \\ u_{1m} & & u_{mm} \end{pmatrix} \begin{pmatrix} \sigma_1 & \cdots & 0 \\ & \sigma_r & \\ 0 & & 0 \end{pmatrix} \begin{pmatrix} v_{11} & & v_{1n} \\ & \ddots & \\ v_{n1} & & v_{nn} \end{pmatrix}$$

https://math.stackexchange.com/questions/3869/what-is-the-intuitive-relationship-between-svd-and-pca

Linear Discriminant Analysis (LDA)

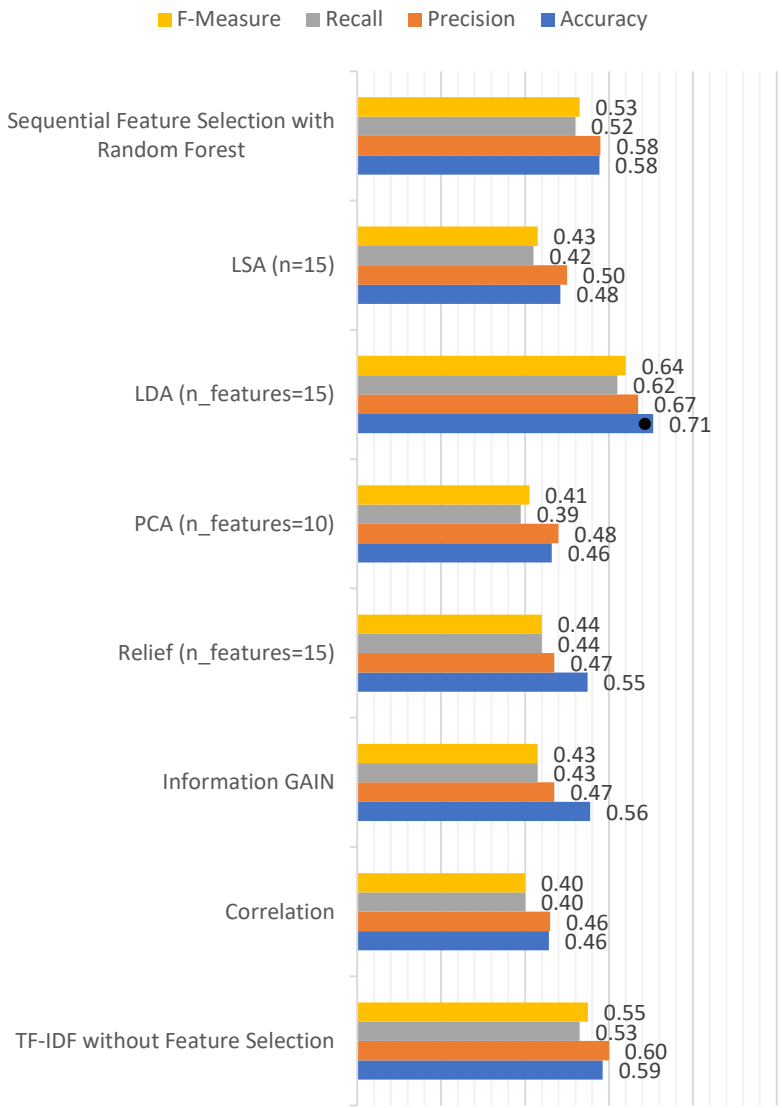
- Supervised
- Preserving class discrimination Information
- Maximize distance

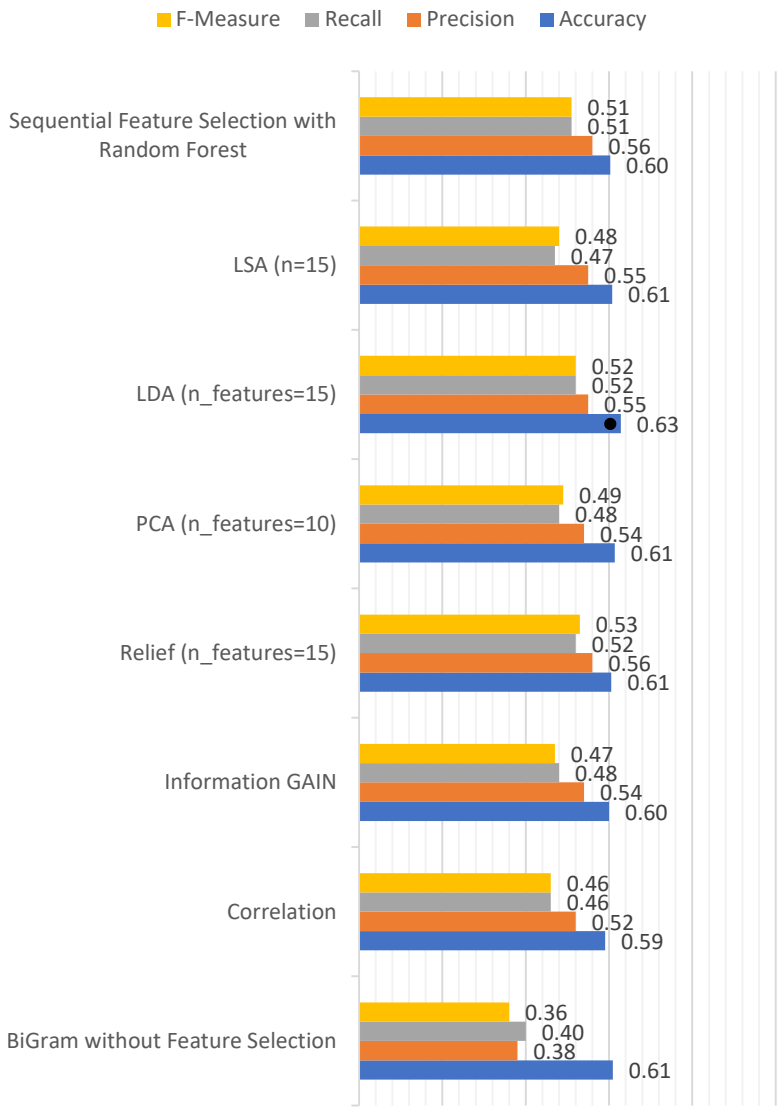




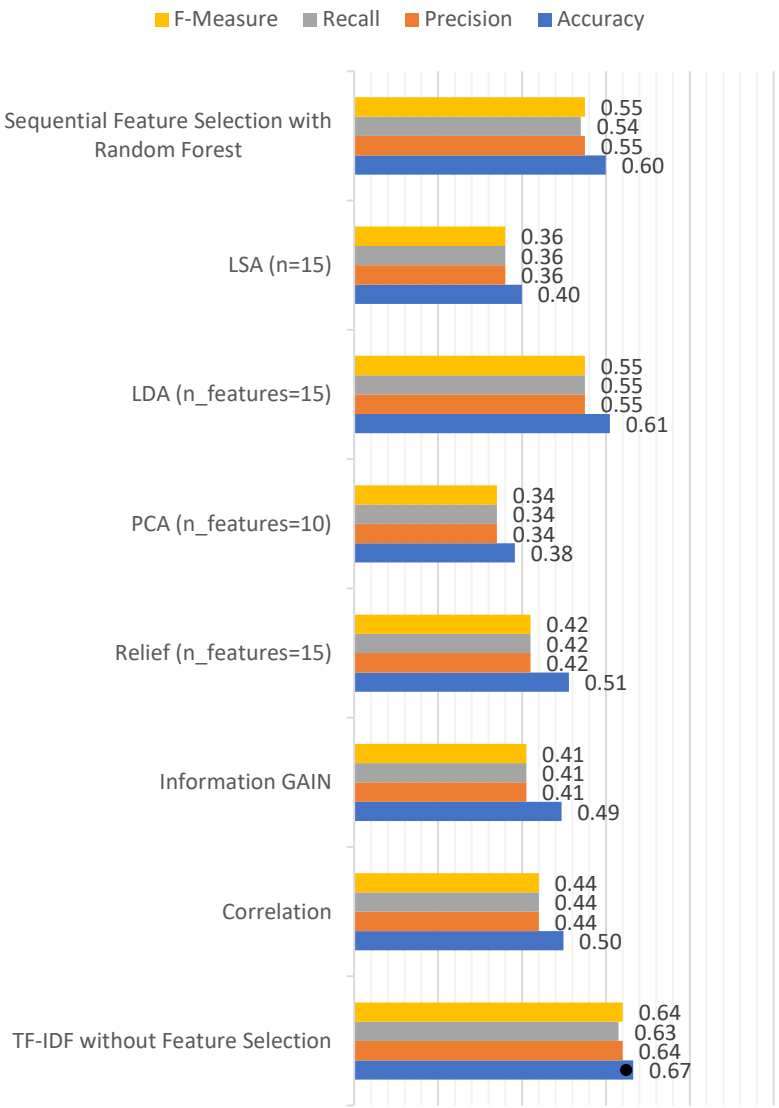
(Fisher Linear Discriminant Analysis,Cheng Lİ 2014)

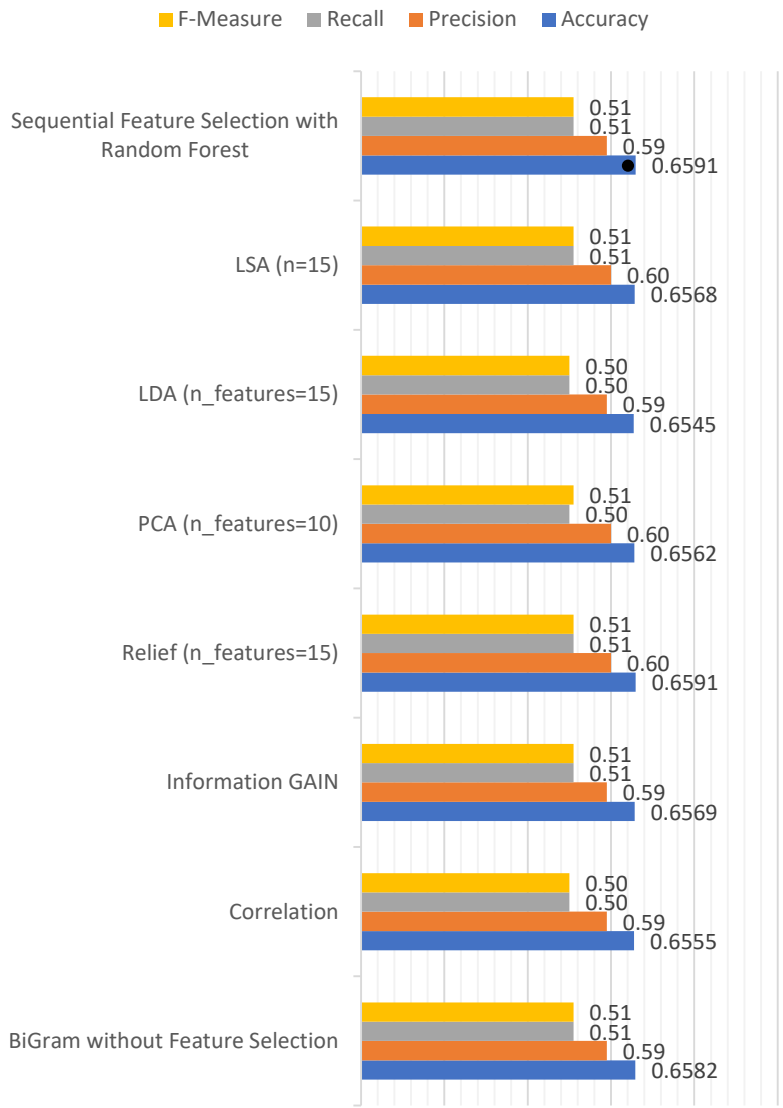
Count Vectorizer vs. Bi Gram on KNN



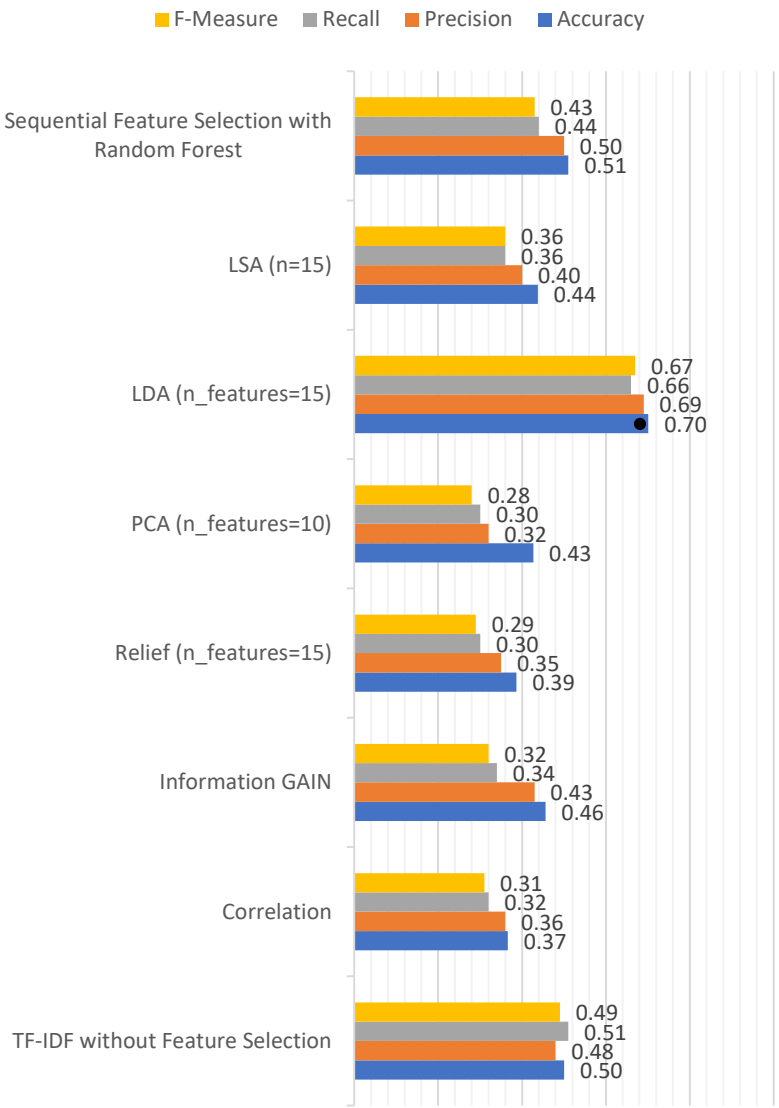


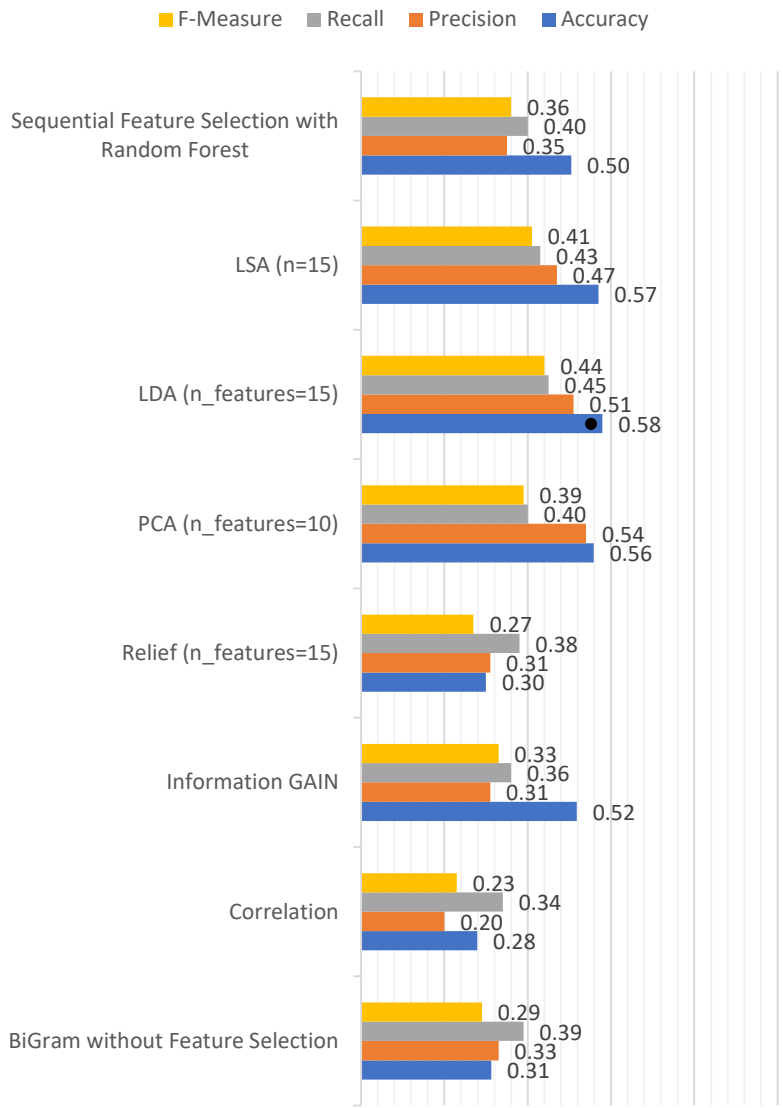
Count Vectorizer vs. Bi Gram on Decision Tree



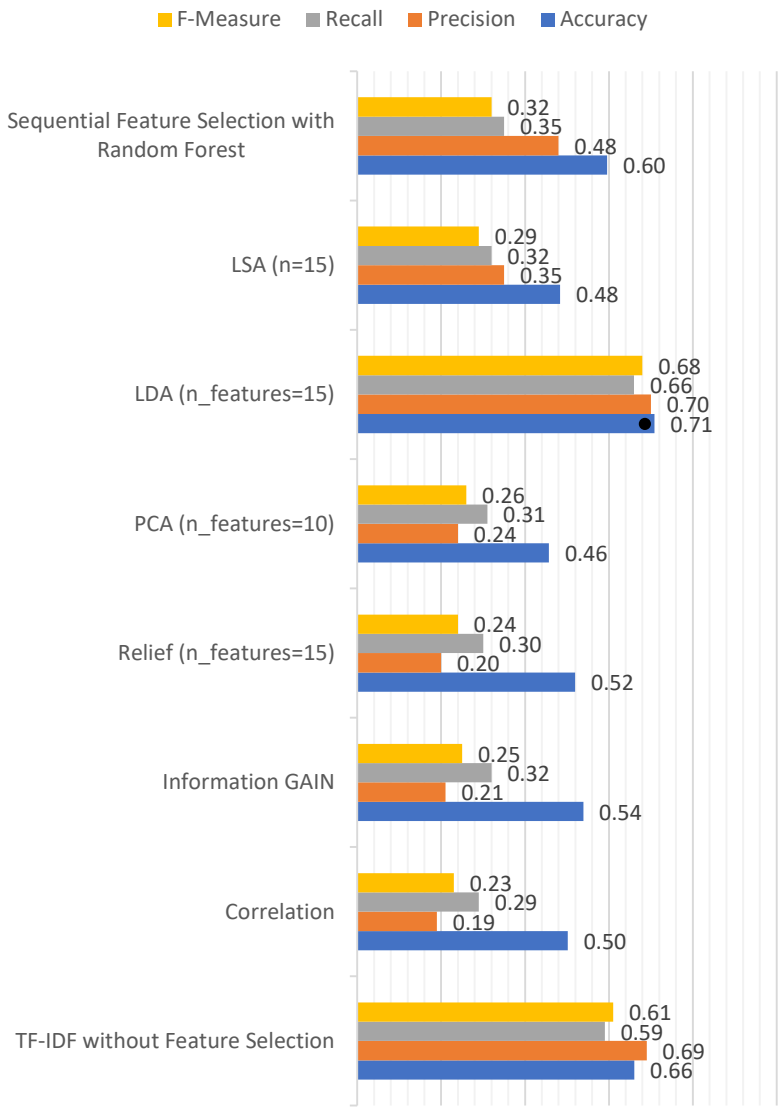


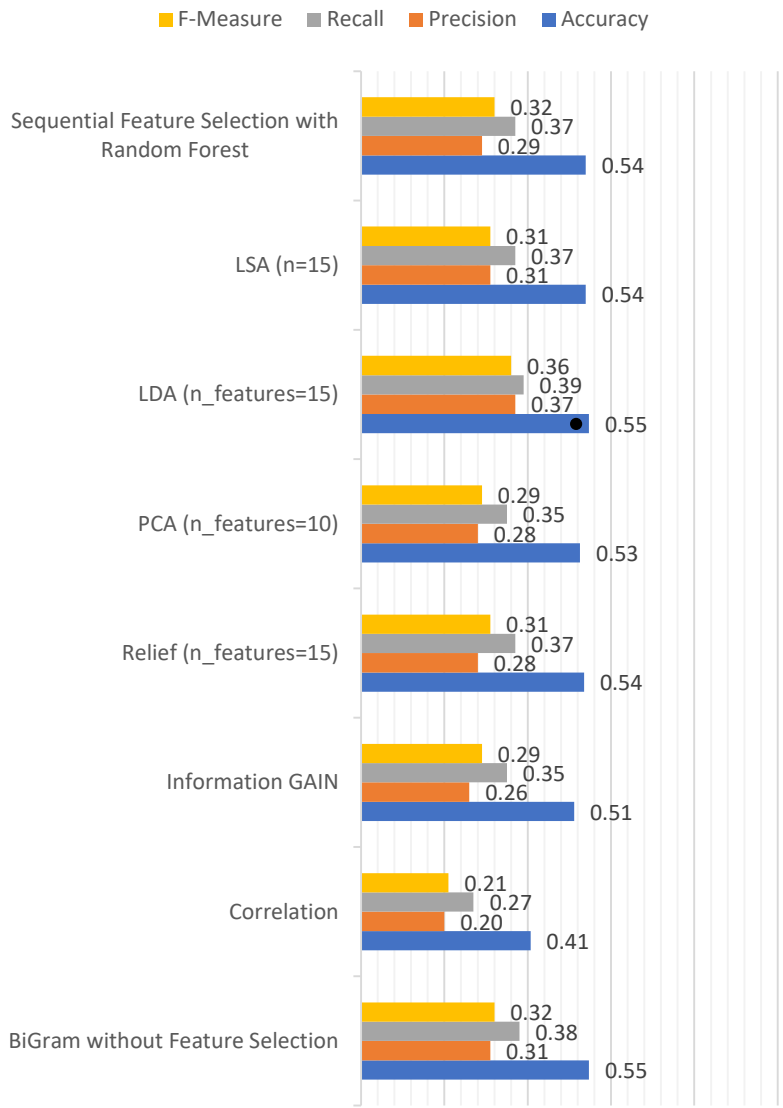
Count Vectorizer vs. Bi Gram on Naive Bayes



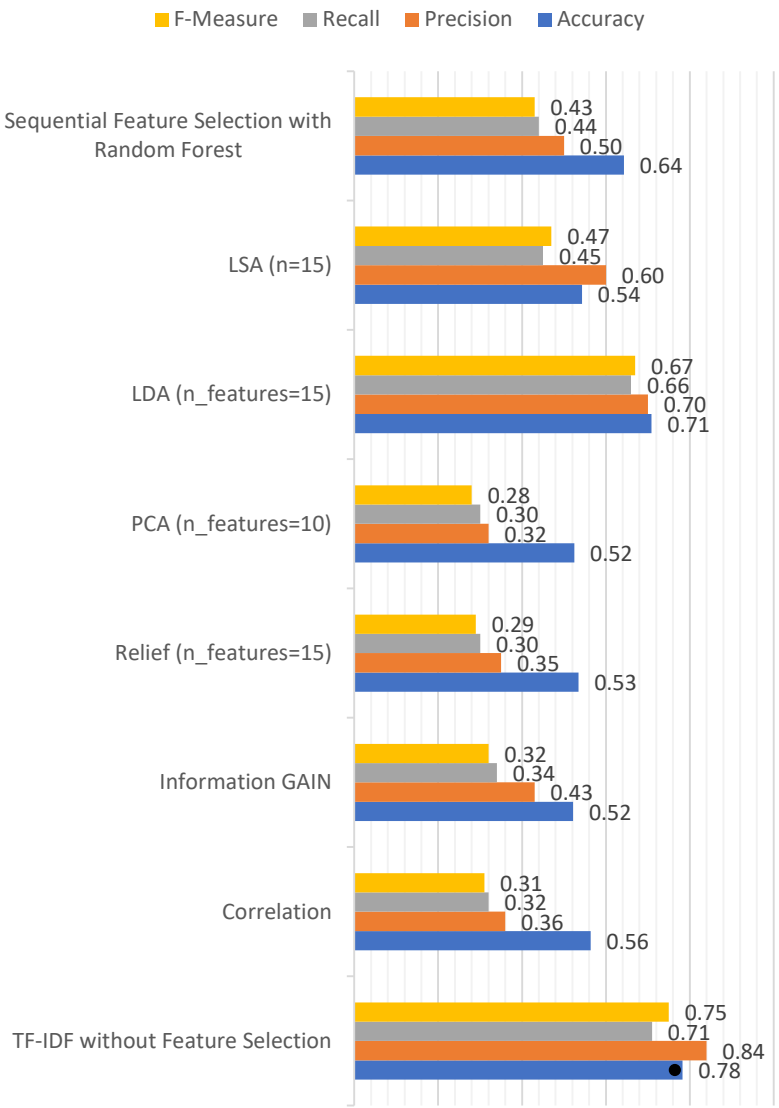


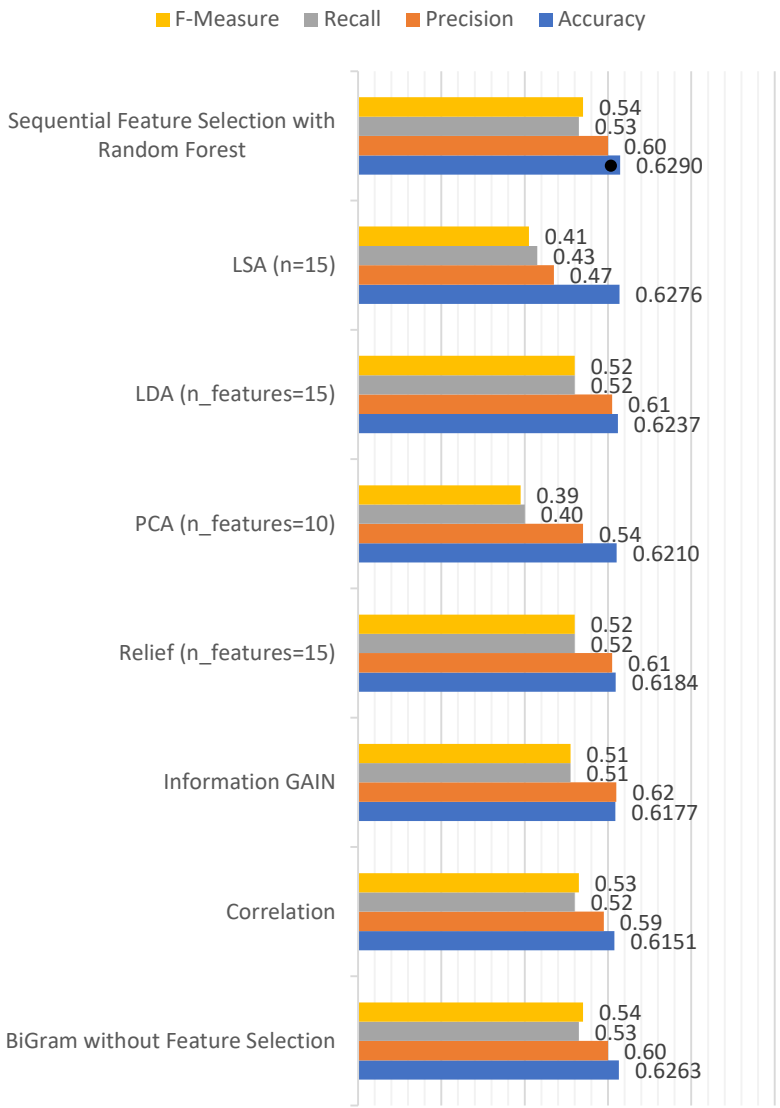
Count Vectorizer vs. Bi Gram on SVM



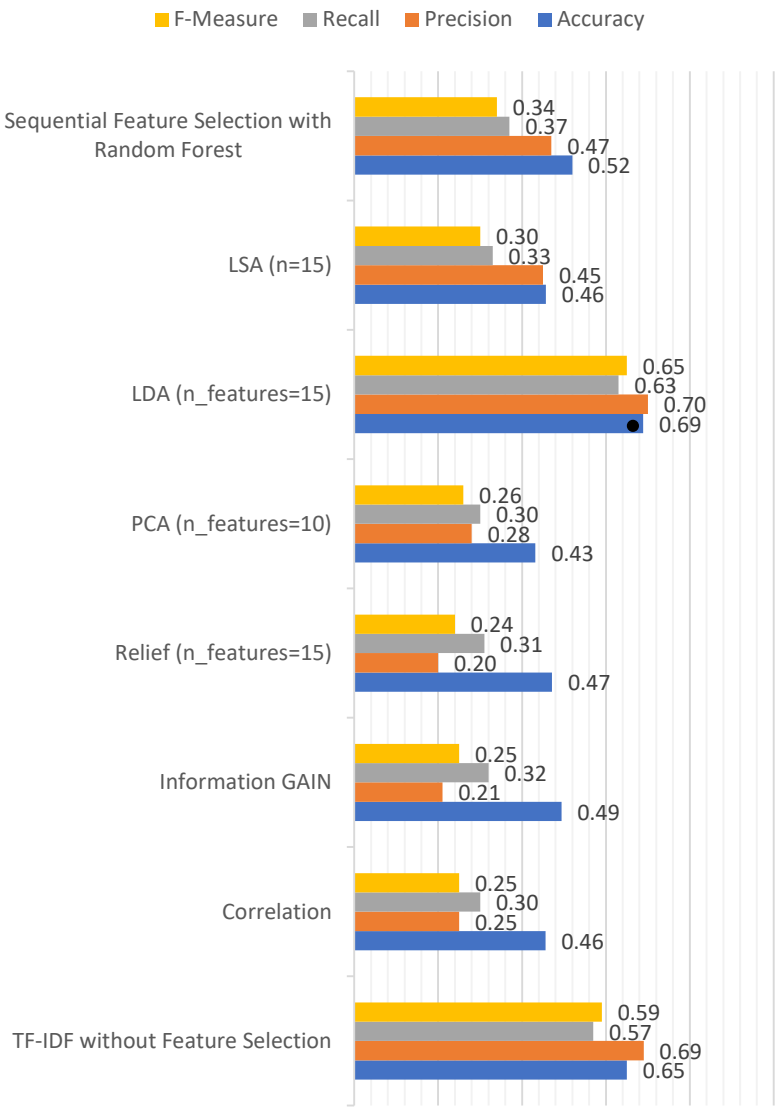


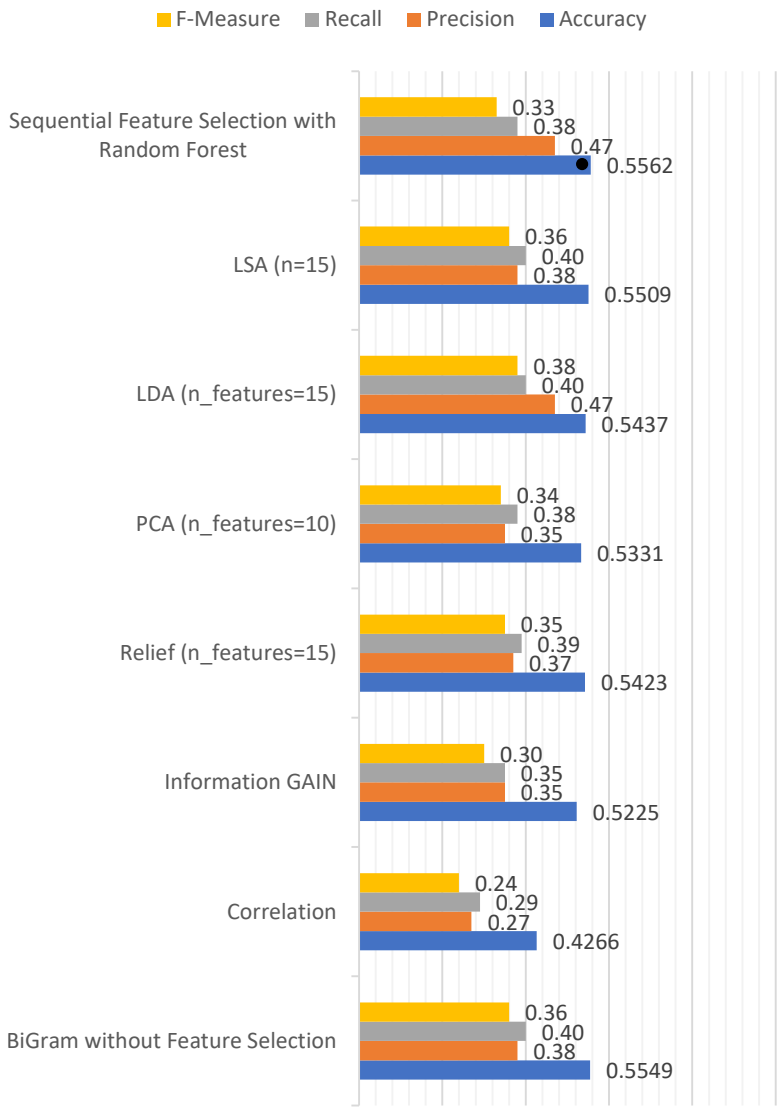
Count Vectorizer vs. Bi Gram on Random Forest





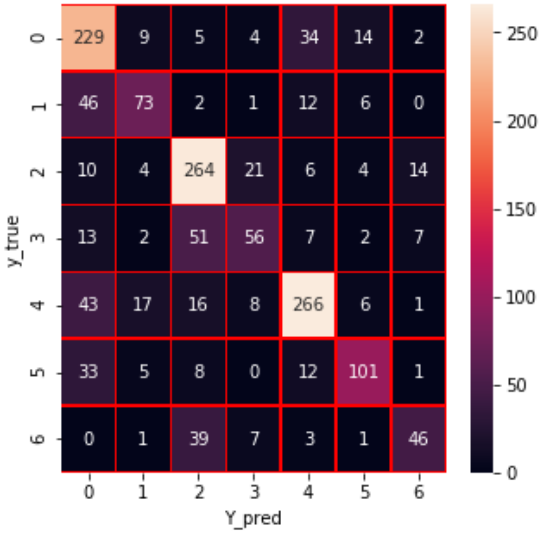
Count Vectorizer vs. Bi Gram on Logistic Regression



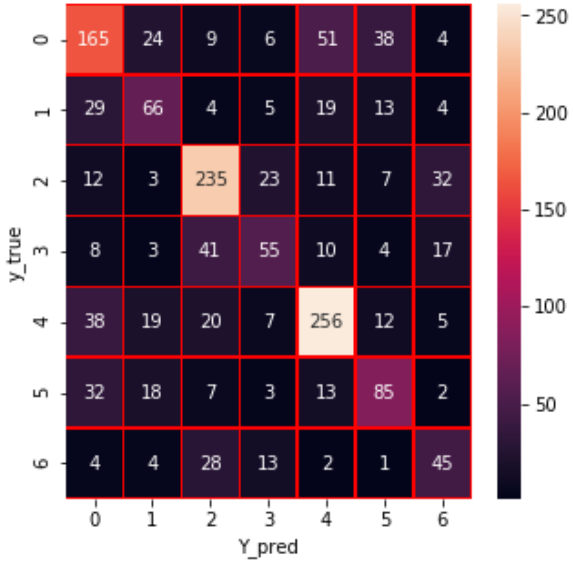


Confusion Matrix (LDA)

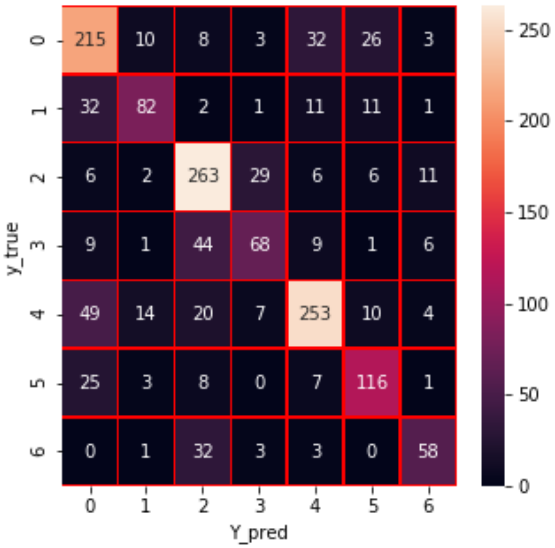
KNN



Decision Tree

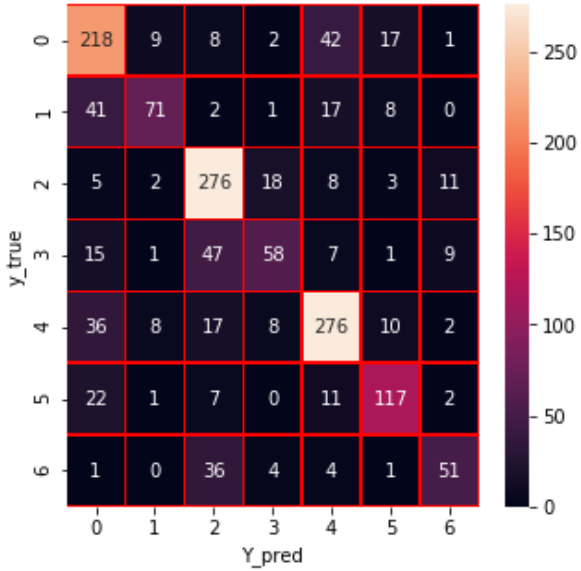


Naive Bayes

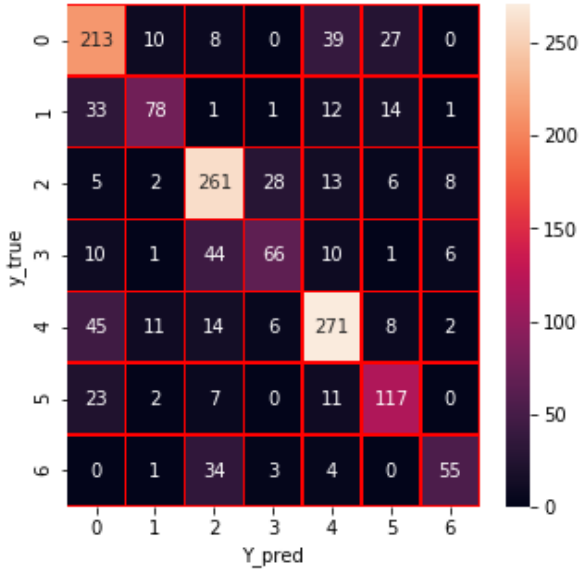


Confusion Matrix (LDA)

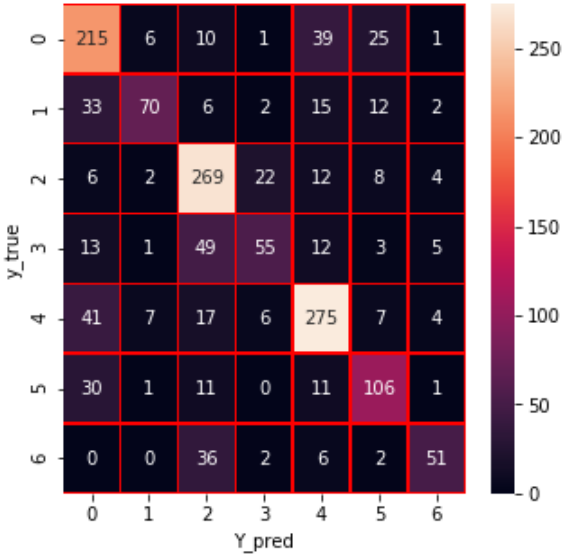
Random Forest



SVM



Logistic Regression



Summary

- As we seen on the charts and matrixes, the most successful feature selection algorithm is TF-IDF without future selection in general.
- But, working with the large dimension brings us different kinds of problems.
  - As the dimensions increase, time complexity's impact will be huge on learning models (Kind of curse of dimensionalty).
  - Irrevelant and outliers data.
  - Visualization.

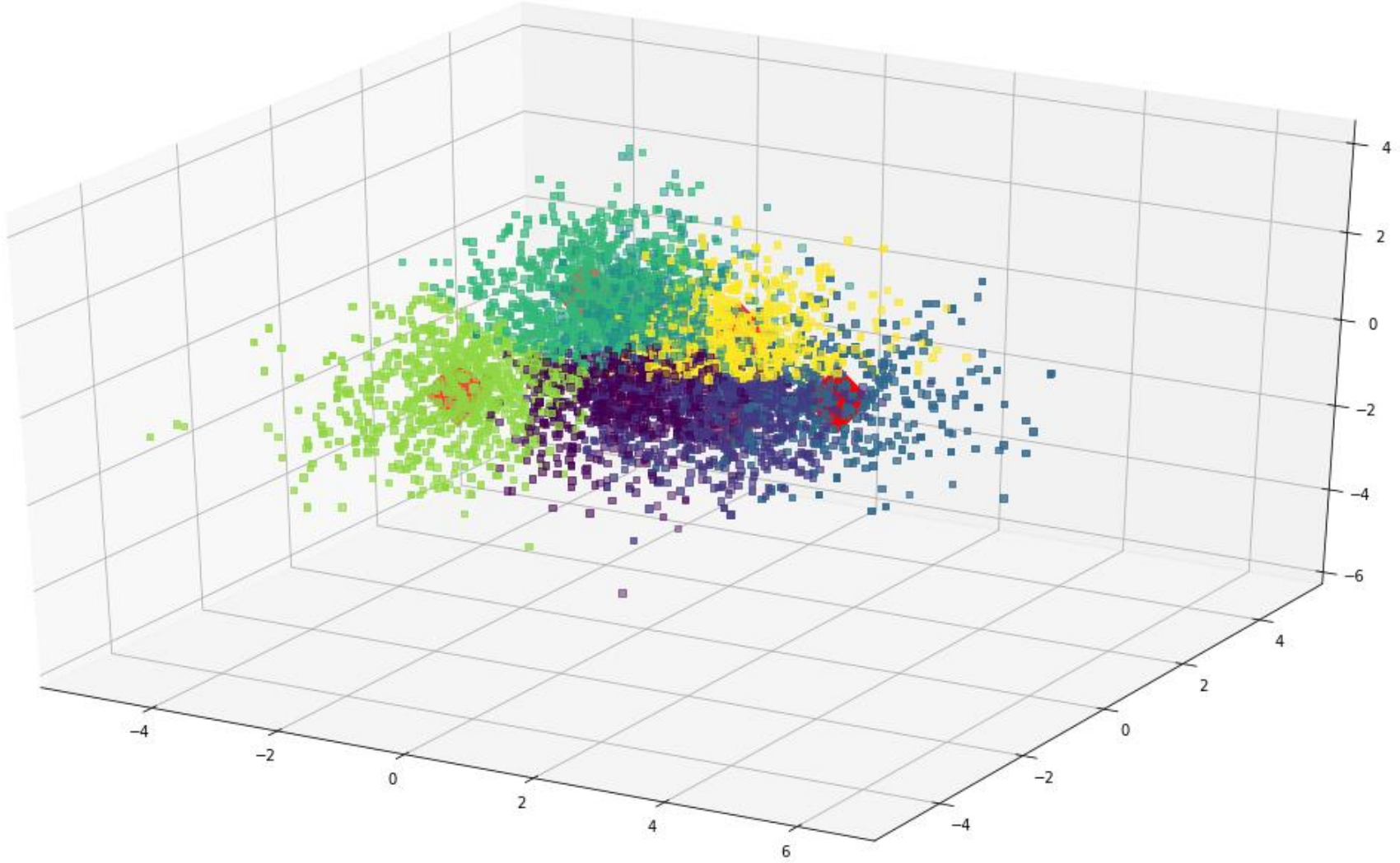
Summary

- So that; we decided to continue with the LDA algorithm.

- While we working on LDA, our features number decreased to six (thanks to data mining gods).

- We got nearly same results with the features of six instead of thousands. This fact shows us the power of LDA algorithm very strongly.

Extra(K-Means, Hiearархical C.)



n_clusters=7 - n_components=3

Dendrogram