Semantic Content Recommendation System Using AWS Sagemaker

This project tries to develop a semantic content recommendation system using AWS Sage maker . Data has been trained and tested on a 20 newsgroups dateset and comparision of two supervised algorithms namely Multinomial Naive Bayes Classifier and Support Vector Machine(SVM) has been done on the basis of it.

AUTHORS

Eeshan Garg (11918)

AFFILIATIONS

Cluster Innovation Centre,
University Of Delhi, 110007, India
Mentor - Dr. Nirmal Yadav

INTRODUCTION

Information retrieval is the science of searching for information in a document, searching for documents themselves, or searching for metadata that describe data. This project combines the techniques of naive bayes and support vector machine for information retrieval. This approach uses bayes formulae for text document preprocessing and then uses support vector machine for classification

DATASET

List of the 20 newsgroups, partitioned respective to their subject matter is as follows:

- alt.atheism
- comp.graphics
- comp.os.ms-windows.mis
- comp.sys.ibm.pc.hardwar
- comp.sys.mac.hardware
- comp.windows.x
- misc.forsale
- rec.autos
- rec.motorcycles
- rec.sport.baseball
- rec.sport.hockey
- sci.crypt
- sci.electronics
- sci.med
- sci.space
- soc.religion.christian
- talk.politics.guns
- talk.politics.mideast
- talk.politics.misc
- talk.politics.misc

```
From: lerxst@wam.umd.edu (where's my thing)
Subject: WHAT car is this!?
Nntp-Posting-Host: rac3.wam.umd.edu
Organization: University of Maryland, College Park
Lines: 15

 I was wondering if anyone out there could enlighten me on this car I saw
the other day. It was a 2-door sports car, looked to be from the late 60s/
early 70s. It was called a Bricklin. The doors were really small. In addition,
the front bumper was separate from the rest of the body. This is
all I know. If anyone can tellme a model name, engine specs, years
of production, where this car is made, history, or whatever info you
have on this funky looking car, please e-mail.

Thanks,
- IL
   ---- brought to you by your neighborhood Lerxst ----
```

Example of a Newsgroup Document

DATA PREPROCESSING

Stripping the headers, footers and quotations

```
"A fair number of brave souls who upgraded their SI clock oscillator have\nshared their experiences for this poll. Pl
ease send a brief message detailing\nyour experiences with the procedure. Top speed attained, CPU rated speed,\nadd o
n cards and adapters, heat sinks, hour of usage per day, floppy disk\nfunctionality with 800 and 1.4 m floppies are e
specially requested.\n\nI will be summarizing in the next two days, so please add to the network\nknowledge base if y
ou have done the clock upgrade and haven't answered this\npoll. Thanks."
```

Data frame after Tokenization

	going	something	computer	system	might	please	reply-to:	using	never	can't	...	paint	data.	argic	plastic	gordon	assumed	impact	back.
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0
1	0.0	10.0	0.0	0.0	6.0	3.0	0.0	1.0	0.0	6.0	...	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0
2	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
19992	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
19993	0.0	2.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
19994	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
19995	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
19996	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

19997 rows × 2000 columns

MODELS

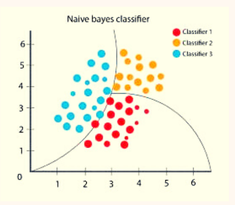
Multinomial Naive Bayes

This algorithm is a probabilistic learning method that is mostly used in Natural Language Processing. It predicts the tag of a text such as a piece of email or newspaper article. It calculates the probability of each tag for a given sample and then gives the tag with the highest probability as output.

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

using Bayesian probability terminology, the above equation can be written as

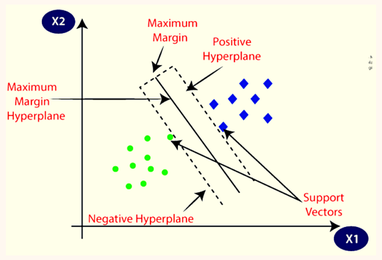
$$\text{Posterior} = \frac{\text{prior x likelihood}}{\text{evidence}}$$



Naive Bayes Classifier

Support Vector Machine

The objective of this algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points.
To separate the two classes of data points, there are many possible hyper planes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes.



SVM

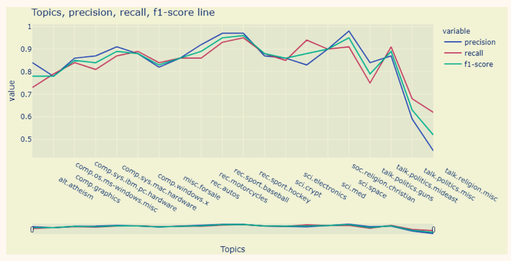
RESULTS

The following classification report and line graph show performance of each classifier Multinomial Naive Bayes and SVM, in regards to precision, recall scores and F1-Score.

Multinomial Naive Bayes

	precision	recall	f1-score	support
alt.atheism	0.86	0.71	0.78	284
comp.graphics	0.80	0.80	0.80	251
comp.os.ms-windows.misc	0.82	0.84	0.83	241
comp.sys.ibm.pc.hardware	0.86	0.81	0.84	254
comp.sys.mac.hardware	0.92	0.88	0.90	247
comp.windows.x	0.90	0.90	0.90	239
misc.forsale	0.84	0.84	0.84	262
rec.autos	0.85	0.85	0.85	267
rec.motorcycles	0.93	0.89	0.91	296
rec.sport.baseball	0.98	0.95	0.97	256
rec.sport.hockey	0.97	0.95	0.96	235
sci.crypt	0.91	0.92	0.92	230
sci.electronics	0.86	0.83	0.85	252
sci.med	0.87	0.90	0.88	249
sci.space	0.88	0.92	0.90	234
soc.religion.christian	0.99	0.93	0.96	270
talk.politics.guns	0.83	0.70	0.76	294
talk.politics.mideast	0.86	0.92	0.89	262
talk.politics.misc	0.56	0.68	0.61	216
talk.religion.misc	0.45	0.67	0.54	159
avg / total	0.86	0.85	0.85	5000

Classification Report

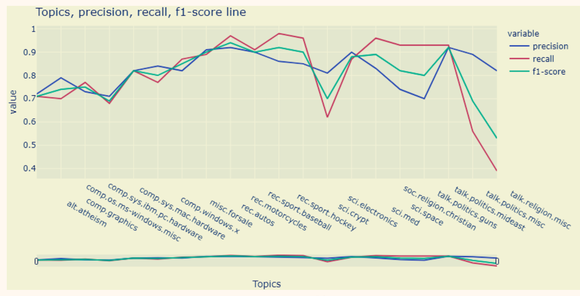


Line Chart

Support Vector Mchine

	precision	recall	f1-score	support
alt.atheism	0.72	0.71	0.71	319
comp.graphics	0.79	0.70	0.74	389
comp.os.ms-windows.misc	0.73	0.77	0.75	394
comp.sys.ibm.pc.hardware	0.71	0.68	0.69	392
comp.sys.mac.hardware	0.82	0.82	0.82	385
comp.windows.x	0.84	0.77	0.80	395
misc.forsale	0.82	0.87	0.85	390
rec.autos	0.91	0.89	0.90	396
rec.motorcycles	0.92	0.97	0.94	398
rec.sport.baseball	0.90	0.91	0.90	397
rec.sport.hockey	0.86	0.98	0.92	399
sci.crypt	0.85	0.96	0.90	396
sci.electronics	0.81	0.62	0.70	393
sci.med	0.90	0.87	0.88	396
sci.space	0.83	0.96	0.89	394
soc.religion.christian	0.74	0.93	0.82	398
talk.politics.guns	0.70	0.93	0.80	364
talk.politics.mideast	0.92	0.93	0.92	376
talk.politics.misc	0.89	0.56	0.69	310
talk.religion.misc	0.82	0.39	0.53	251
avg / total	0.83	0.82	0.82	7532

Classification Report



Line Chart

CONCLUSION

From this study, we found that Naive Bayes does very well in text classification. It should be noted that SVM does achieve very good scores as well. In short, given a text classification problem, any of the two could be used without a big compromise on the classification accuracy. But if we want to do it in less training time, then it would be advised to use Naïve Bayes instead of SVM.

FUTURE DIRECTIONS

It would be fascinating to explore what makes every one of the classifiers proceed as they do, may it be from the scores every classifier accomplishes in characterizing the text, or the assets it consumes; time, and computer memory during training and testing. Research can be done on the effect of component extraction and portrayal on the performance of every classifier. For example Multinomial Naive Bayes in this study performs far superior to SVM on account of the BOW approach.

REFERENCES

https://aws.amazon.com/getting-started/hands-on/semantic-content-recommendation-system-amazon-sagemaker/
https://medium.com/@siyao_sui/nlp-with-the-20-newsgroups-dataset-ab35cd0ea902
https://github.com/tanishq9/Text-Classification-20-Newsgroups
http://archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups