**Notes on project**

Problem-

1.Class imbalance

When we encounter such problems, we are bound to have difficulties solving them with standard algorithms. Conventional algorithms are often biased towards the majority class, not taking the data distribution into consideration. In the worst case, minority classes are treated as outliers and ignored. For some cases, such as fraud detection or cancer prediction, we would need to carefully configure our model or artificially balance the dataset, for example by [under sampling or oversampling](https://en.wikipedia.org/wiki/Oversampling_and_undersampling_in_data_analysis) each class.

It is desirable to have a classifier that gives high prediction accuracy over the majority class, while maintaining reasonable accuracy for the minority classes. Therefore, we will leave it as it is.

Need to read about oversampling/under sampling

<https://en.wikipedia.org/wiki/Oversampling_and_undersampling_in_data_analysis>

A Nice text analysis using multi class with class imbalance

<https://towardsdatascience.com/multi-class-text-classification-with-scikit-learn-12f1e60e0a9f>

2. Multi label classification- a text belongs to multiple categories, no mutually exclusive class

This article is telling BOG approach for guten burg dataset that her did not work very good but also refreed to a thesis which did not remove names from the features to identify gernes.

If multil label then how to measure accuracy?

3.Also interesting to see which doesn’t fall in any of the gerne

4. Using nltk?

5. Can title of the book and author give gerne?

6, Can we predict author by gerne?

7. Sentiment analysis, need to read paper on it

8. sentimensts+gerne? To guide the classification?

9. Named entity recoganisation, names appearing again and again and adding to list

**Notes-**

**Just and overview of the whole process**

<https://towardsdatascience.com/machine-learning-general-process-8f1b510bd8af>

Nice work on handling lareg dataset, memory handling, garbage collecting etc

<https://www.kaggle.com/frankherfert/tips-tricks-for-working-with-large-datasets>

<https://towardsdatascience.com/bag-of-words-in-genre-identification-on-the-project-gutenberg-dataset-a98ee146d53f>

she is taking about the feature selection using BOG and XGBOost and compared methods with accuracy for the same.

<https://medium.com/@emilyjia/genre-identification-and-the-compositional-effect-of-genre-in-literature-a4a97a16e681>

A thesis link she shared-

<https://www.aclweb.org/anthology/C18-1167/?source=post_page--------------------------->

If it’s a multi label classification a simple comment classification implementation, uses on vs all etc techniques

<https://towardsdatascience.com/journey-to-the-center-of-multi-label-classification-384c40229bff>

A Production ready text analysis sample

<https://towardsdatascience.com/a-production-ready-multi-class-text-classifier-96490408757>

<https://medium.com/@robert.salgado/multiclass-text-classification-from-start-to-finish-f616a8642538>

One more model evaluation

<https://towardsdatascience.com/multi-class-text-classification-model-comparison-and-selection-5eb066197568>

there is a nice article on topic modelling, how to get the topic out of text

<https://towardsdatascience.com/basic-nlp-on-the-texts-of-harry-potter-topic-modeling-with-latent-dirichlet-allocation-f3c00f77b0f5>

also on sentiment analysis

<https://towardsdatascience.com/basic-nlp-on-the-texts-of-harry-potter-sentiment-analysis-1b474b13651d>

one on feature selection methods with code examples

<https://medium.com/analytics-vidhya/feature-selection-methods-with-code-examples-a78439477cd4>

some tips and tricks on multi class classification

<https://medium.com/@b.terryjack/tips-and-tricks-for-multi-class-classification-c184ae1c8ffc>

Excellent article with visualation includinf LDA

<https://www.machinelearningplus.com/nlp/topic-modeling-visualization-how-to-present-results-lda-models/>

Rough steps

**EDA/**

**Pre Processing-**

Take care of the file size, garbade collector etc, there is a link above for that

Head, shape, describe, info, isnull.sum

Check total data count

Word count of each gerne

Word count distribution per gerne plot

Check sample files

Check distribution

Check null values

Remove punctuation

Remove html tags

Remove non alphabetic characters

Lemmatization/stemming

Studying or Studied = study

Remove stop words, since only English novel hence easy

Articles, common words etc

Label Encoding of Encoding Target classes

May be see classes balance

Most frequent classes

World cloud for each class/all documents

Feature Engineering

Bag of words

TF-IDF

Here can use GRID Search CV for parameter tuning, n grams, min/max\_df

Use feature selection, use chi-square, mutual infor to find the terms in each class that are most co related.

PCA

To make the classifier design production ready, we can create a pipeline of all these processes discussed above thus making it easier to move to other systems

Train test validation split

k-fold cross

train model

plot confusion matrix

measure time

check prediction

run another models

do the same

choose the best model

simple article on predict horror movie author from given text

<https://towardsdatascience.com/a-machine-learning-approach-to-author-identification-of-horror-novels-from-text-snippets-3f1ef5dba634>

Nice implementation of PCA

<https://stackabuse.com/implementing-pca-in-python-with-scikit-learn/>

<https://wilkelab.org/classes/SDS348/2016_spring/worksheets/class9.html>

Nice multi dimensional visualtisation on wine dataset

<https://medium.com/swlh/effective-visualization-of-multi-dimensional-data-a-hands-on-approach-b48f36a56ee8>