DOCSUMO DATA VERSE 2023

Visualize Train Conquer

Event Date : 17^{th} January, 2023 - 23^{rd} January, 2023

A

Data Insight Report On

ArXiv Datasets for Scholarly Articles

Author

Team Non-Linear

Nirajan Bekoju

Nishant Luitel

Submission Date

23^{rd} January, 2023

Contents

List of Figures

List of Tables

Chapter 1

Introduction to Docsumo DataVerse 2023

1.1 Abstract

ArXiv is a public service repository and open-source archive for scholarly articles in the fields of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics.

1.2 Problem Description

In this data challenge, Datasets that contains abstracts and categories columns that are collected from the arXiv portal are given to the participants. There are altogether 157 subject categories. For this competition, datasets having 23 classes are provided.

The task is to build a model to predict the category given paper abstract and title.

1.3 Evaluation Metric

The evaluation metric for this competition is Mean F1-Score. The F1 score, commonly used in information retrieval, measures accuracy using the statistics precision(p) and recall(r).

Precision is the ratio of true positives (tp) to all predicted positives (tp + fp). Recall is the ratio of true positives (tp) to all actual positives (tp + fn).

The F1 score is given by:

$$\mathrm{F1} = 2\frac{\mathrm{p}\cdot\mathrm{r}}{\mathrm{p}+\mathrm{r}} \quad \text{where} \quad \mathrm{p} = \frac{\mathrm{tp}}{\mathrm{tp}+\mathrm{fp}}, \quad \mathrm{r} = \frac{\mathrm{tp}}{\mathrm{tp}+\mathrm{fn}}$$

The F1 metric weights recall and precision equally, and a good retrieval algorithm will maximize both precision and recall simultaneously. Thus, moderately good performance on both will be favored over extremely good performance on one and poor performance on the other.

1.4 Dataset Overview

Participants were given three files.

1. **train.csv** : the training set
2. **test.csv** : the test set
3. **sample_submission.csv** : a sample submission file in the correct format

1.4.1 Train Dataset Overview

The train dataset contains total four columns as follow:

1. **id** : Unique id of the article
2. **title** : Title of the article
3. **abstract** : Abstract of the article
4. **category** : Label of the article

```
sample_train_data = train_data.sample(7)
sample_train_data
```

	id	title	abstract	category
177813	2012.08209	Possible phase transition in plasma mirror modes	Mirror modes in collisionless high-temperatu...	physics
393226	2102.00193	Coupling innovation method and feasibility ana...	In order to solve the recent defect in garba...	cs
84798	1906.0761	Improving Sentiment Analysis with Multi-task L...	Sentiment analysis is directly affected by c...	cs
19018	1810.02767	Efficient Estimation of Smooth Functionals in ...	We study a problem of estimation of smooth f...	stat
101001	2103.16966	Revisiting regular sequences in light of ratio...	Regular sequences generalize the extensively...	cs
239917	2012.15227	Interaction between vortex beams and diatomic ...	The interaction between vortex beam (VB) and...	physics
42	2007.14755	Learning Transferable Push Manipulation Skills...	This paper is concerned with learning transf...	cs

Figure 1.1: Train Data Sample

Altogether, there are 8,61,236 entries in the training dataset. None of id, title and abstract column are null. However, 4 row for category column are null. The train dataset is of size 995.08 MB.

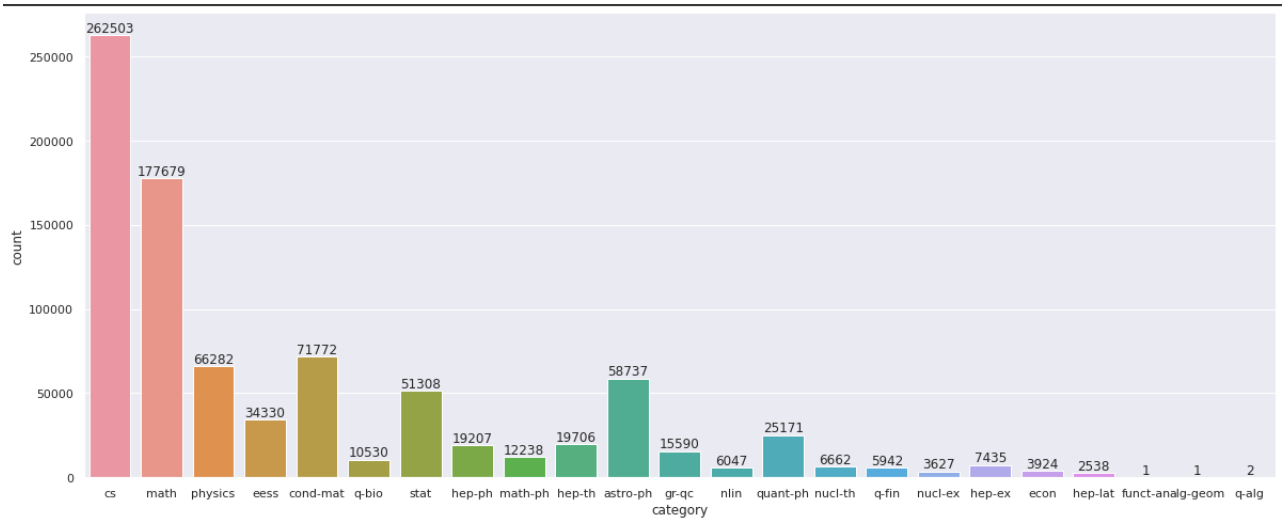

Figure 1.2: Train Data Category Countplot

The name of the categories in the train data are cs, math, physics, eess, cond-mat, q-bio, stat, hep-ph, math-ph, hep-th, astro-ph, gr-qc, nlin, quant-ph, nucl-th, q-fin, nucl-ex, hep-ex, econ, hep-lat, funct-an, alg-geom, and q-alg.

From the graph, we can observe that the count of some labels like cs, math, physics, cond-mat, astro-ph are greater than 50,000 while some labels have their count range from 2,000 to 30,000 and label "funct-an", "alg-geom" and "q-alg" have 2, 1 and 1 count respectively.

So, it is clear that the given training data is highly imbalanced and so we have to take the imbalanced into consideration. Otherwise, the model will be highly biased toward the majority class. Hence, we will have to use various sampling methods to balance the data and feed to the model for increasing the reliablility of the classifier model.

Some sample title and their category are shown below:

id	title	category
2012.08209	Possible phase transition in plasma mirror modes	physics
2102.00193	Coupling innovation method and feasibility analysis of garbage	cs
1810.02767	Efficient Estimation of Smooth Functionals in Gaussian Shift Models	stat

Table 1.1: Train Data Title Category Sample

1.4.2 Test Dataset Overview

The test dataset contains total of three columns : id, title and abstract. The size of test dataset is of 52.12 MB. There are 45,328 entries in the test dataset.

```
test_data.sample(7)
```

	id	title	abstract
15095	1910.1339	Ferromagnetic Kitaev interaction and the origi...	α-RuCl$_3$ is drawing much attention ...
5601	1903.04377	SleepNet: Automated Sleep Analysis via Dense C...	In this work, a dense recurrent convolutiona...
2416	1810.01393	Approximating the Existential Theory of the Reals	The Existential Theory of the Reals (ETR) co...
25225	2006.12492	Analog of Menchov-Trokhimchuk theorem for mono...	The aim of this work is to weaken the condit...
42520	1908.09089	Web-enabled Intelligent System for Continuous ...	A large number of sensors deployed in recent...
13220	2011.14826	Revisiting Rainbow: Promoting more Insightful ...	Since the introduction of DQN, a vast majori...
13281	2009.14295	Wall speed and shape in singlet-assisted stron...	Models with singlet fields coupling to the H...

Figure 1.3: Test Data Sample

1.4.3 Submission Format

For every author in the dataset, submission files should contain two columns: ID and Category. The file should contain a header and have the following format in csv file:

The file should contain a header and have the following format:

id	category
2107.01095	q-fin
1906.08519	cond-mat
2105.05523	math
1906.08731	cs
1703.08445	math
2106.13444	cond-mat
1812.05778	cs
1911.0796	cs
1808.07576	stat
1902.02865	cs

Figure 1.4: Submission File Format

Chapter 2

Data Insights

2.1 Text Data Analysis

Text data analysis were performed in the training datasets. Analysis like title length based on character frequency, title length based on word frequency were analysed.

2.1.1 Title length based on character frequency

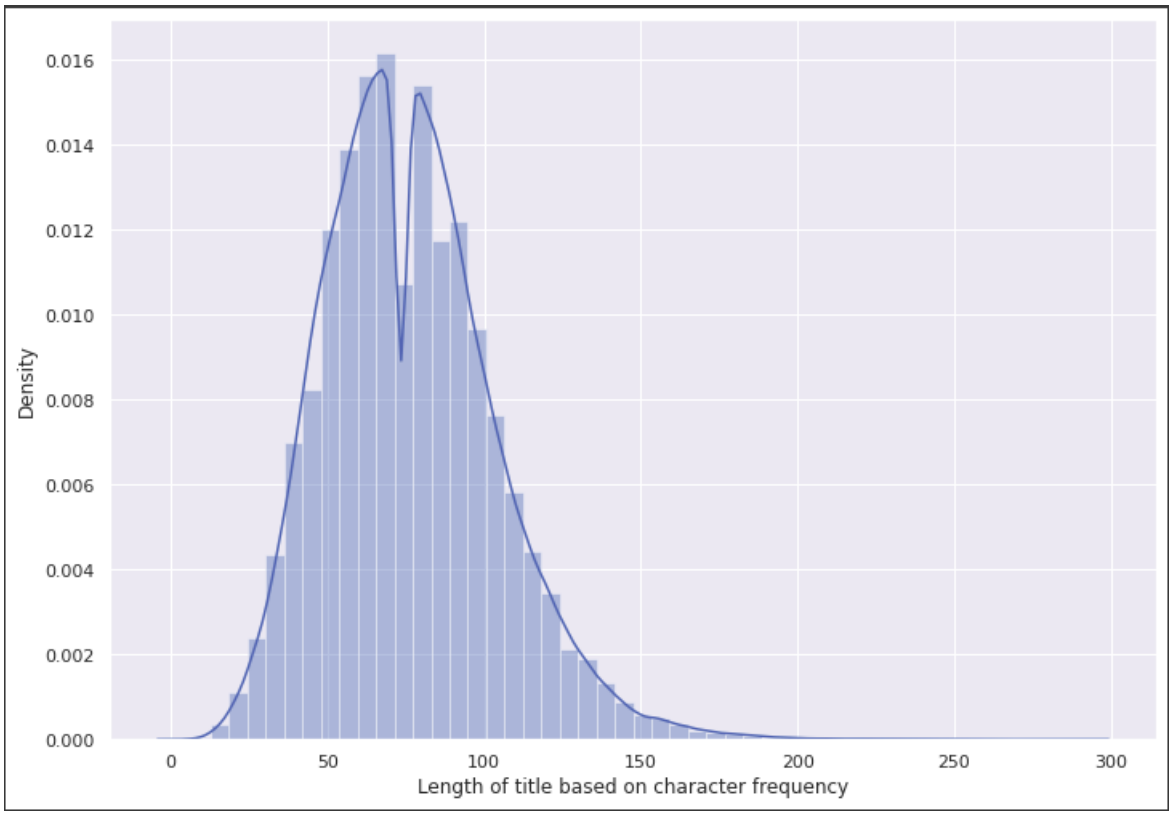

Figure 2.1: Title length based on number of characters (Bi-modal Distribution)

From the figure above, it can be deduced that the title length based on number of characters follow bimodal distribution. Its basic statistics value can be observed in following box plot.

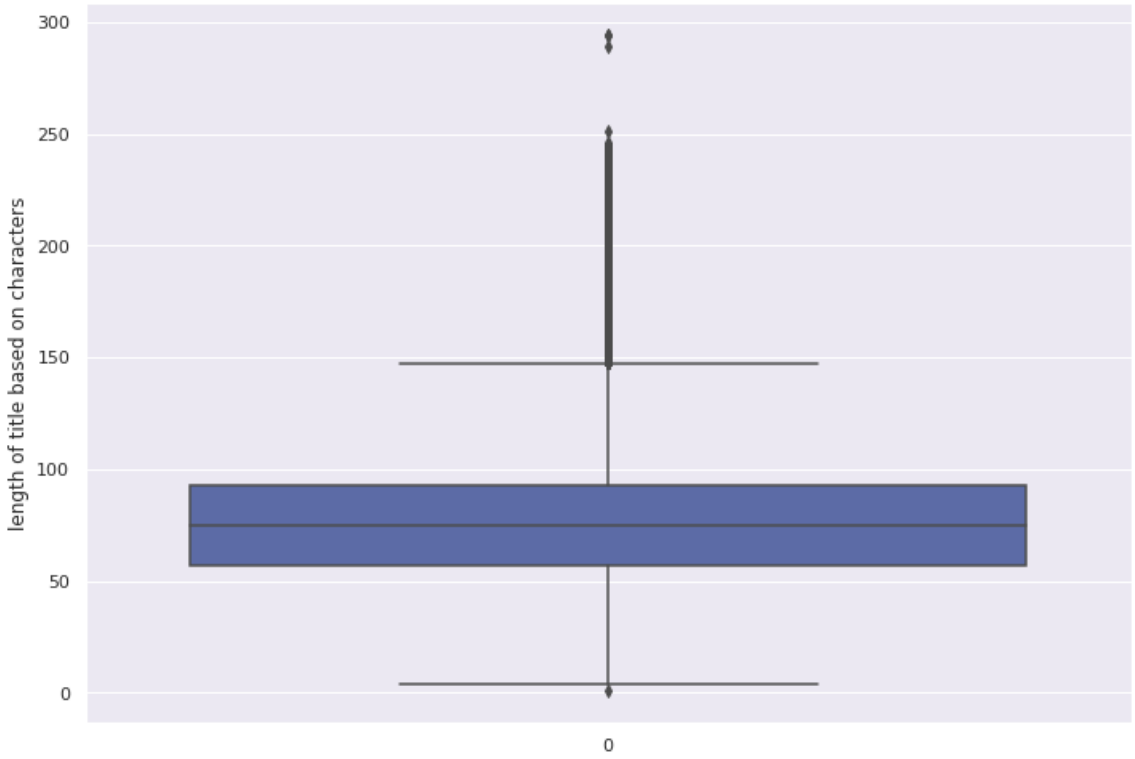

Figure 2.2: Box plot of title length based on character frequency

From the box plot, the average number of characters in a title in the training dataset is found to be 76.3678 $\pm$ 27.2314. The statistics information are shown below.

mean	76.3678
standard deviation	27.2314
minimum	1
first quartile	57
median	75
third quartile	93
maximum	294

Table 2.1: Title length statistics based on character frequency

2.1.2 Title length based on word frequency

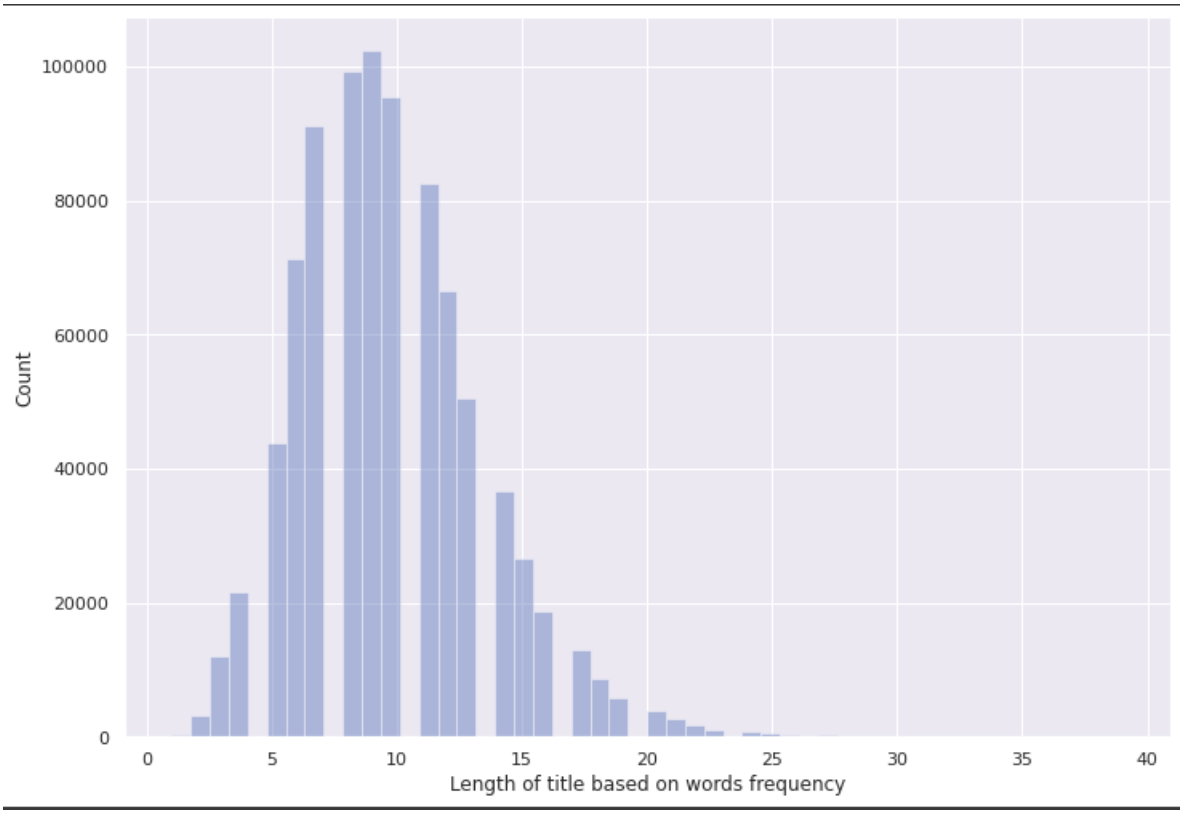

Figure 2.3: Title length based on word frequency (Right Skewed Normal Distribution)

From the above distribution plot, it can be deduced that the title length based on word frequency follow right skewed normal distribution.

From the box plot Fig : 2.4, the average number of words in a title in the training dataset is found to be 9.7681 $\pm$ 3.6153. The statistics information are shown below.

mean	9.7681
standard deviation	3.6153
minimum	1
first quartile	7
median	9
third quartile	12
maximum	39

Table 2.2: Title length statistics based on word frequency

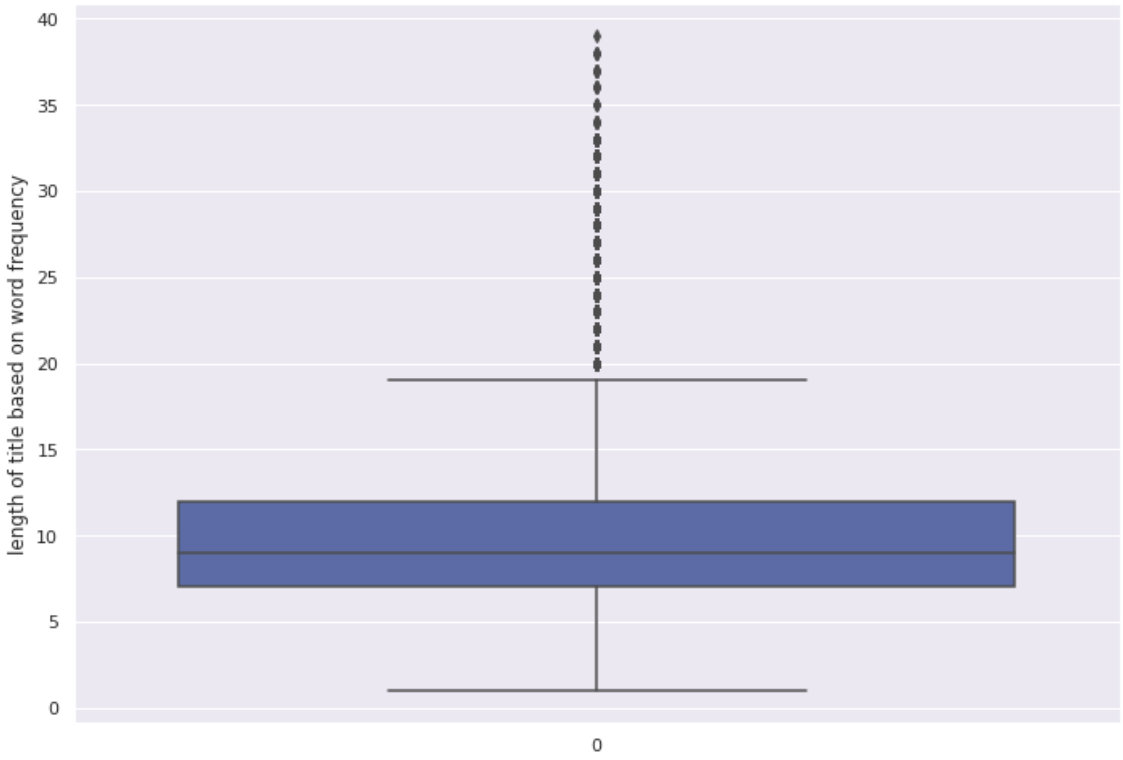

Figure 2.4: Box plot of title length based on word frequency

2.2 Word Cloud

A word cloud (also known as a tag cloud) is a visual representation of words. Cloud creators are used to highlight popular words and phrases based on frequency and relevance. They provide quick and simple visual insights that can lead to more in-depth analyses.

For this report, word cloud is created using the title column of the training data removing all stopwords. From the word cloud Fig 2.5, we can see the clear dominance of the majority class : cs, math, physics, cond-mat, stat, astro-ph, and quant-ph.

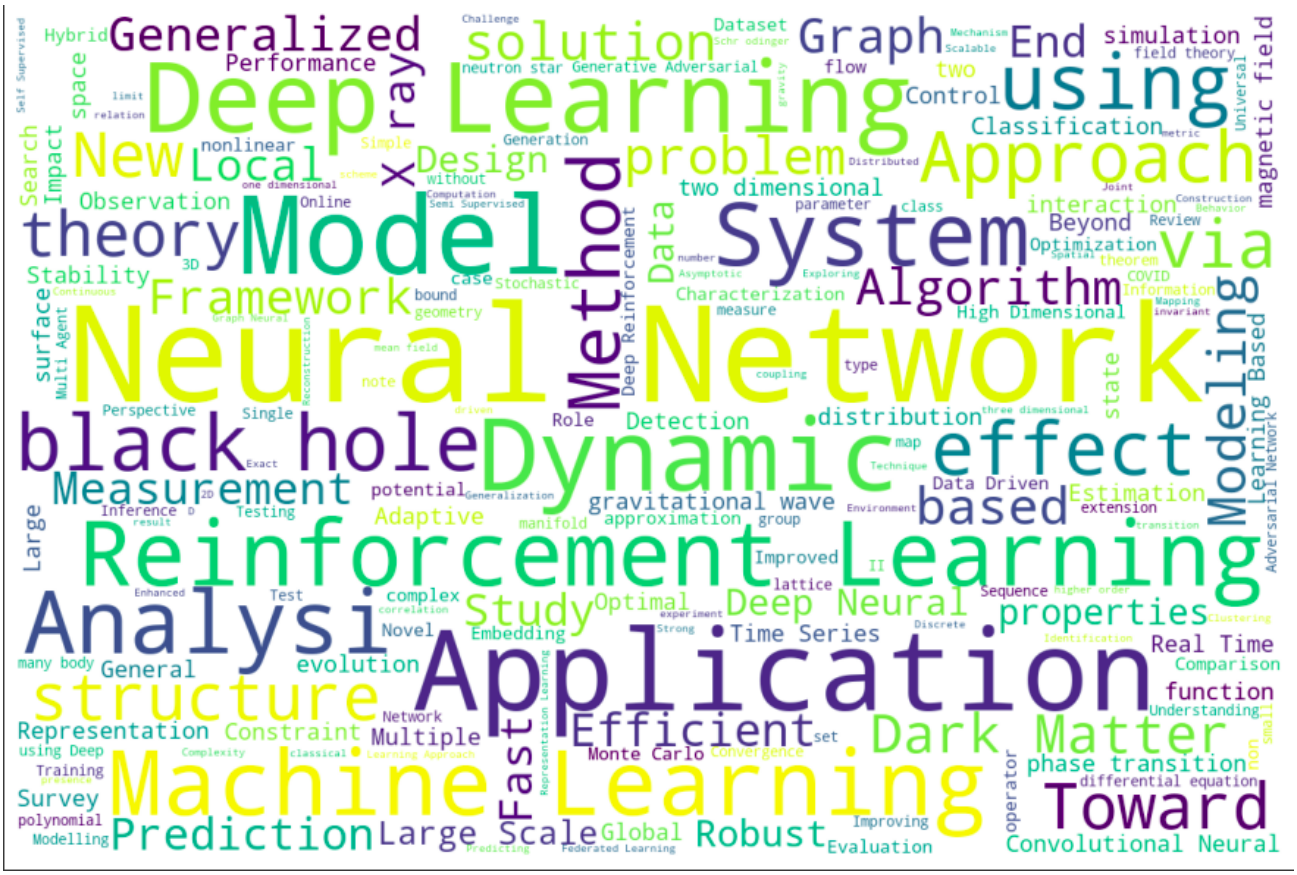

Figure 2.5: Title Word Cloud

Let's view the word cloud on specific topics.

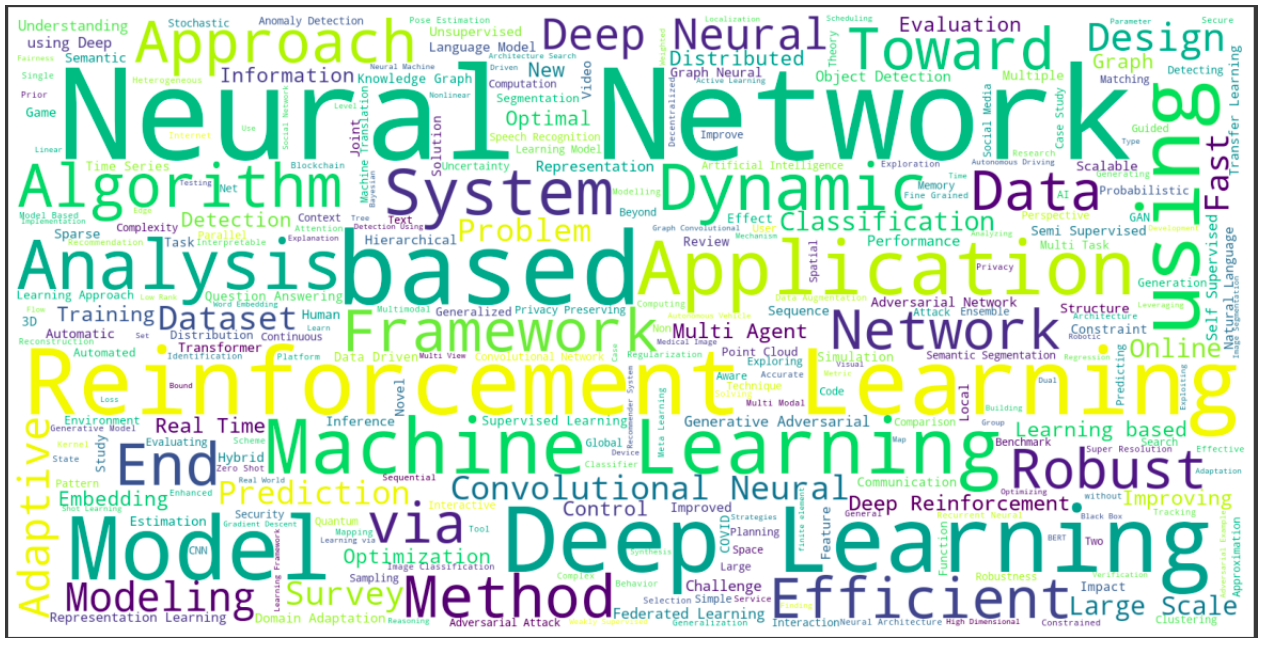

Figure 2.6: CS Word Cloud

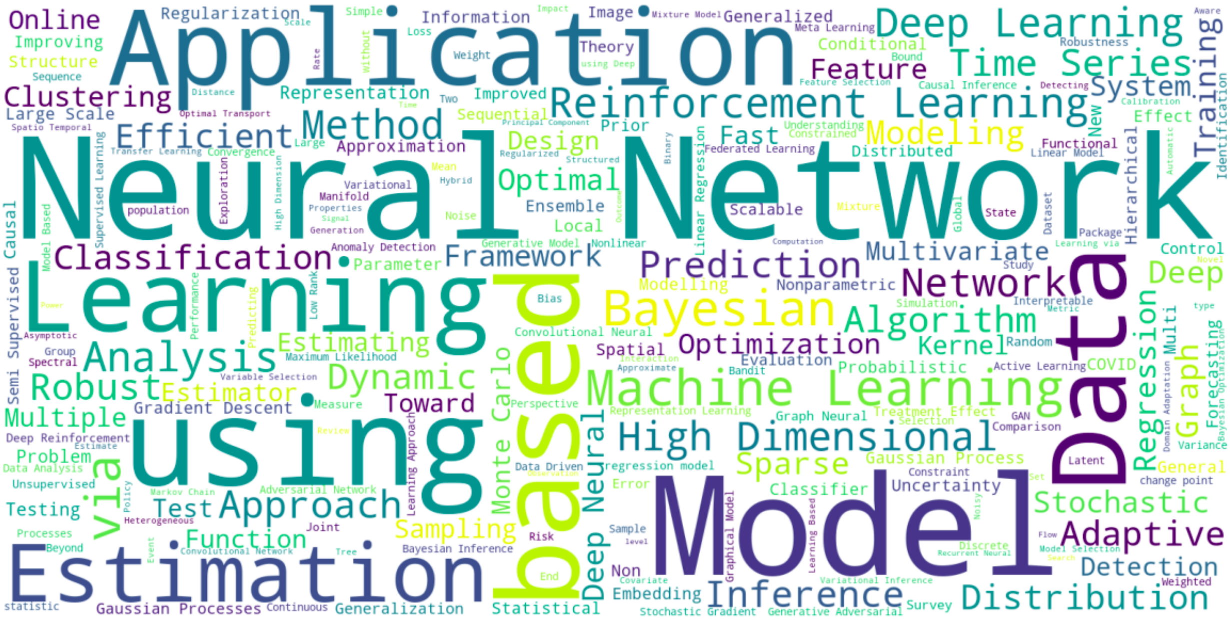

Figure 2.7: Stat Word Cloud

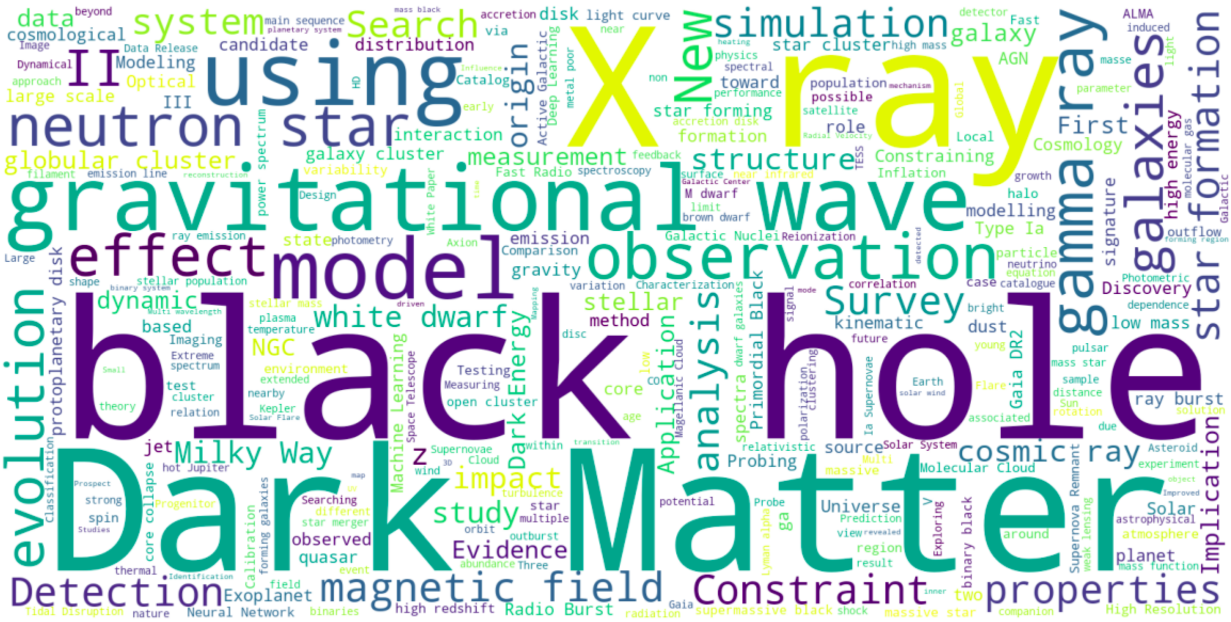

Figure 2.8: Astro-ph Word Cloud

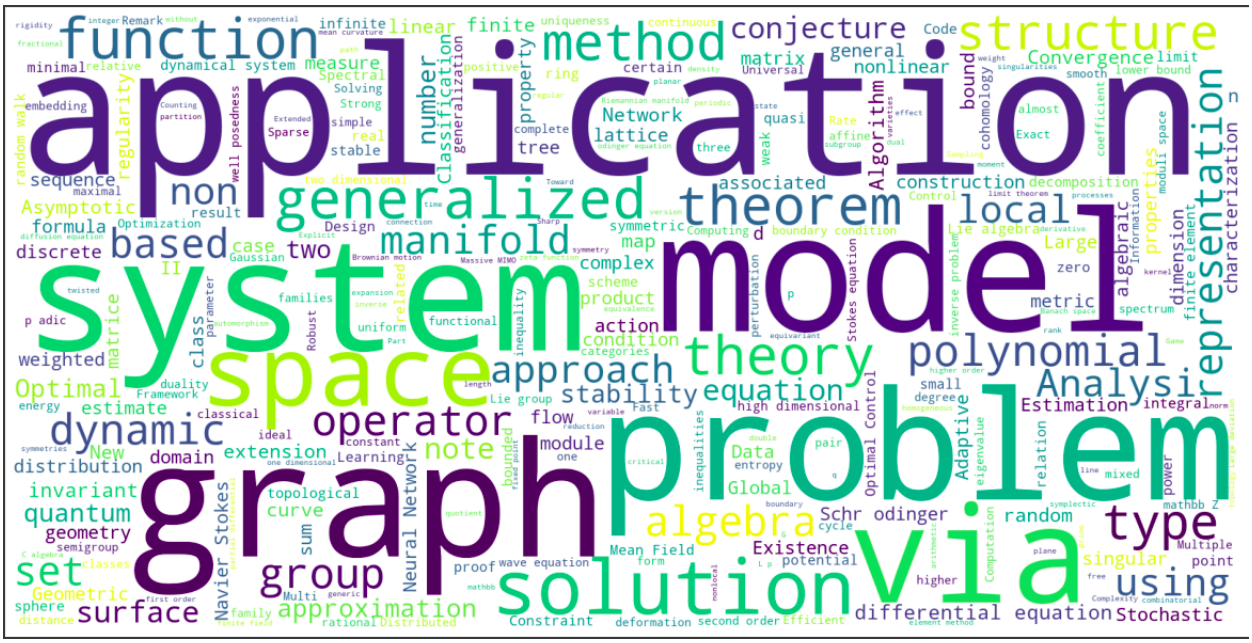

Figure 2.9: Math Word Cloud

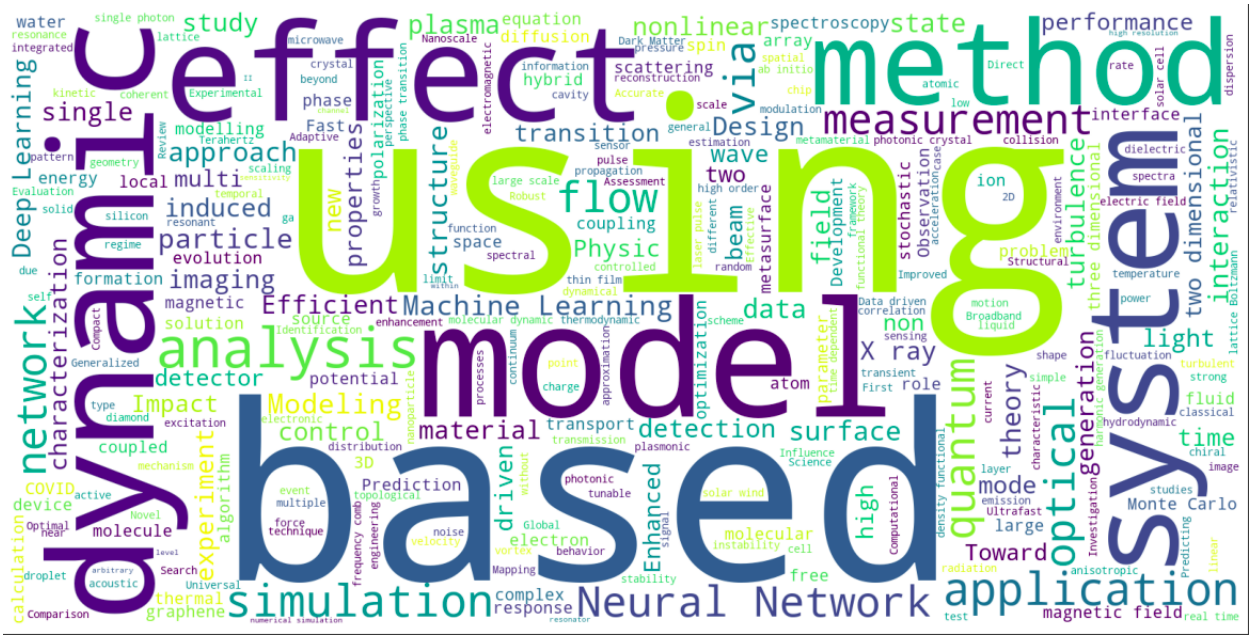

Figure 2.10: Physics Word Cloud

2.2.1 Inference from label based word cloud

From cs and stat word cloud, we can see lots of similar word in both labels like model, neural network, machine learning, network, data, distribution, etc. When looking on to the abstract of these labels, just by judging through the corpus, some of them were even hard for human to classify between them. Similar was the case with astro-ph and physics which can be clearly visualized in the word cloud.

2.2.2 Minority class word cloud

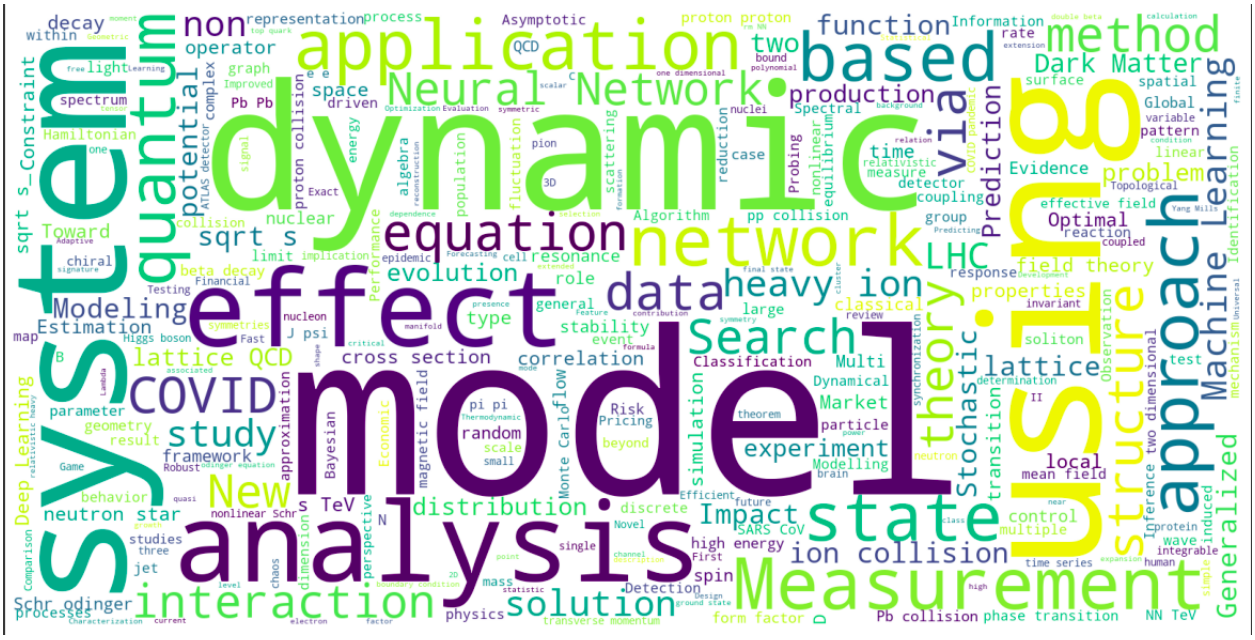

Figure 2.11: minority class Word Cloud

Minority class includes labels with number of data less than 15000 in training datasets. The classes are q-bio, hep-ex, math-ph, nucl-th, nlin, q-fin, econ, nucl-ex, hep-lat, q-alg, funct-an, and alg-geom.

2.3 N-gram Exploration

For the exploration of the most frequency unigram, stopwords were removed from the corpus created from the whole training corpus. Then, we got the count of all unique word from the corpus in a hash map. Using the hash map, we drew this bar graph by taking the top 40 most frequent words. The result is as follow :

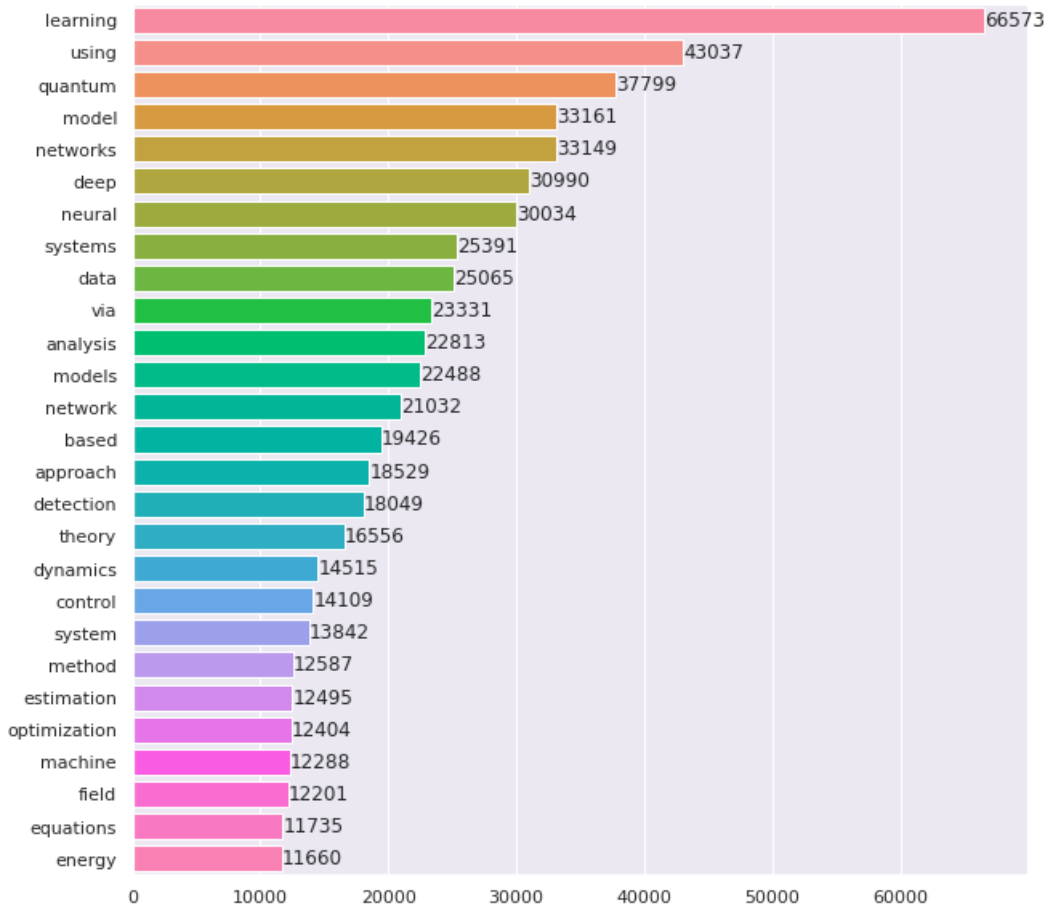

Figure 2.12: Top 40 most frequent words

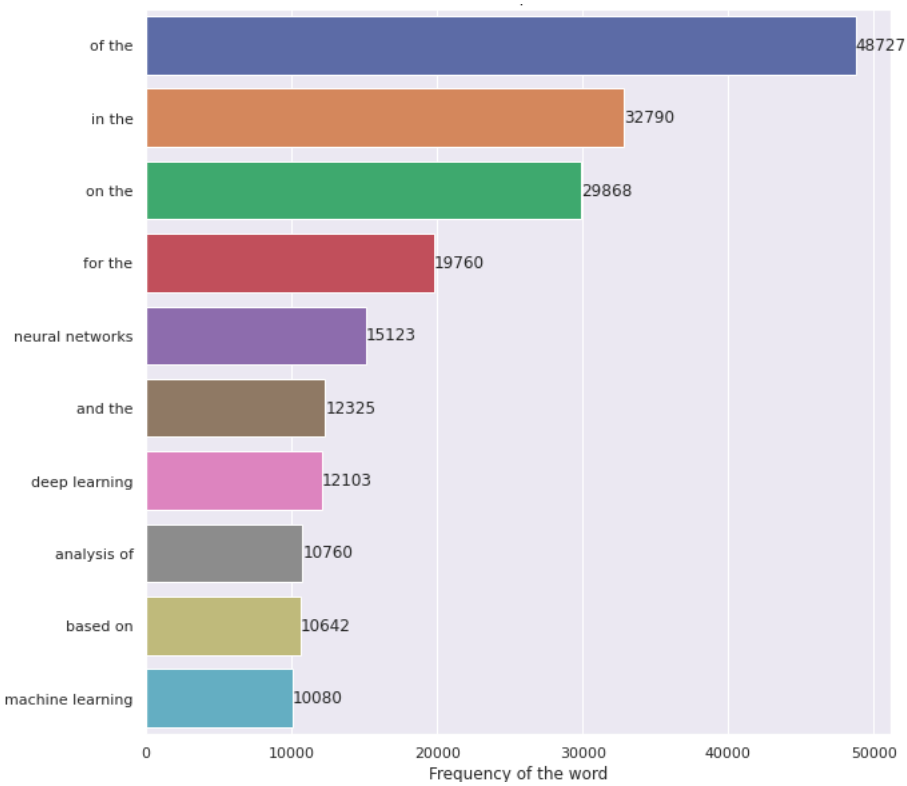

Figure 2.13: Most frequent bi-grams

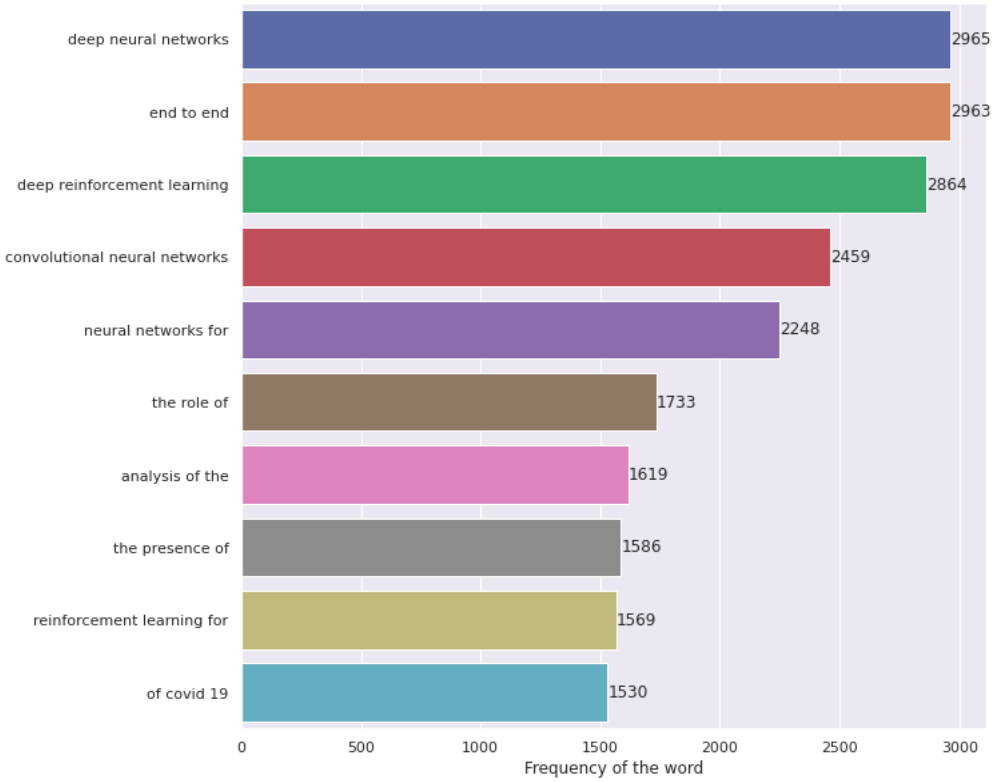

Figure 2.14: Most frequent tri-grams

2.4 Named Entity Recognition

Named-entity recognition (NER) (also known as (named) entity identification, entity chunking, and entity extraction) is a subtask of information extraction that seeks to locate and classify named entities mentioned in unstructured text into pre-defined categories such as person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc. Following table shows the label of NER along with its description.

TYPE	DESCRIPTION
PERSON	People, including fictional.
NORP	Nationalities or religious or political groups.
FAC	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, institutions, etc.
GPE	Countries, cities, states.
LOC	Non-GPE locations, mountain ranges, bodies of water.
PRODUCT	Objects, vehicles, foods, etc. (Not services.)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART	Titles of books, songs, etc.
LAW	Named documents made into laws.
LANGUAGE	Any named language.
DATE	Absolute or relative dates or periods.
TIME	Times smaller than a day.
PERCENT	Percentage, including "%".
MONEY	Monetary values, including unit.
QUANTITY	Measurements, as of weight or distance.
ORDINAL	"first", "second", etc.
CARDINAL	Numerals that do not fall under another type.

Figure 2.15: NER Label Description

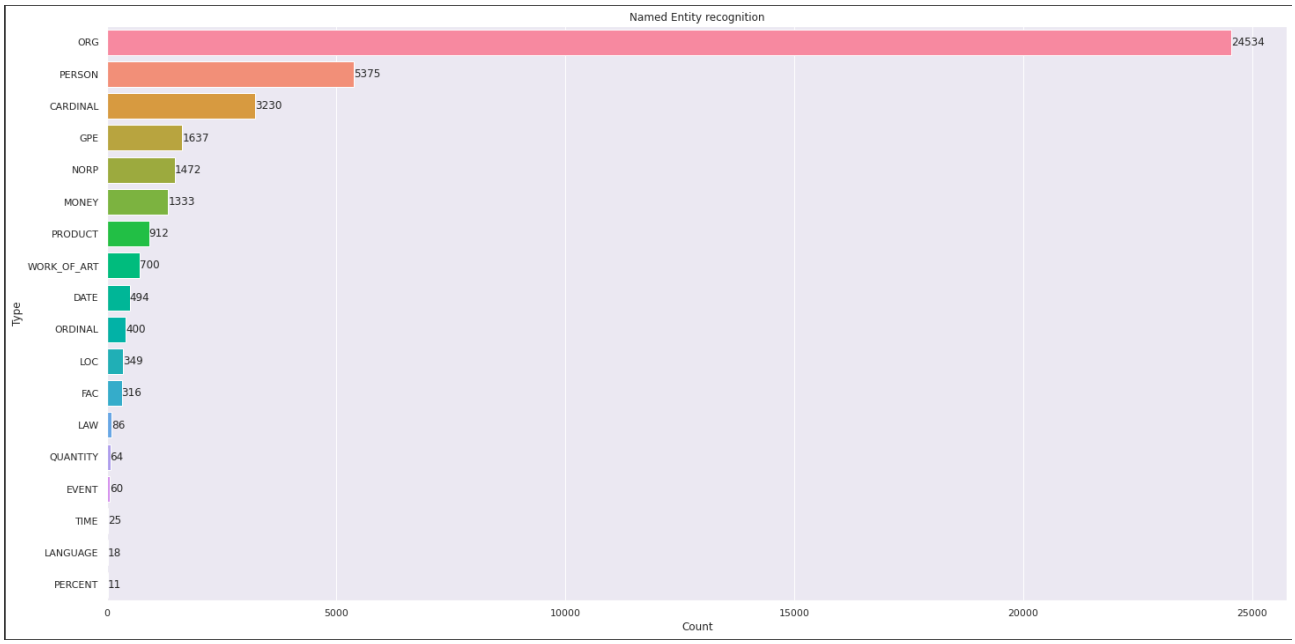

Figure 2.16: NER Countplot

2.4.1 NER Label Exploration

For the generation of following plot based on specific NER label, 50,000 sample were taken from the title column of the training dataset.

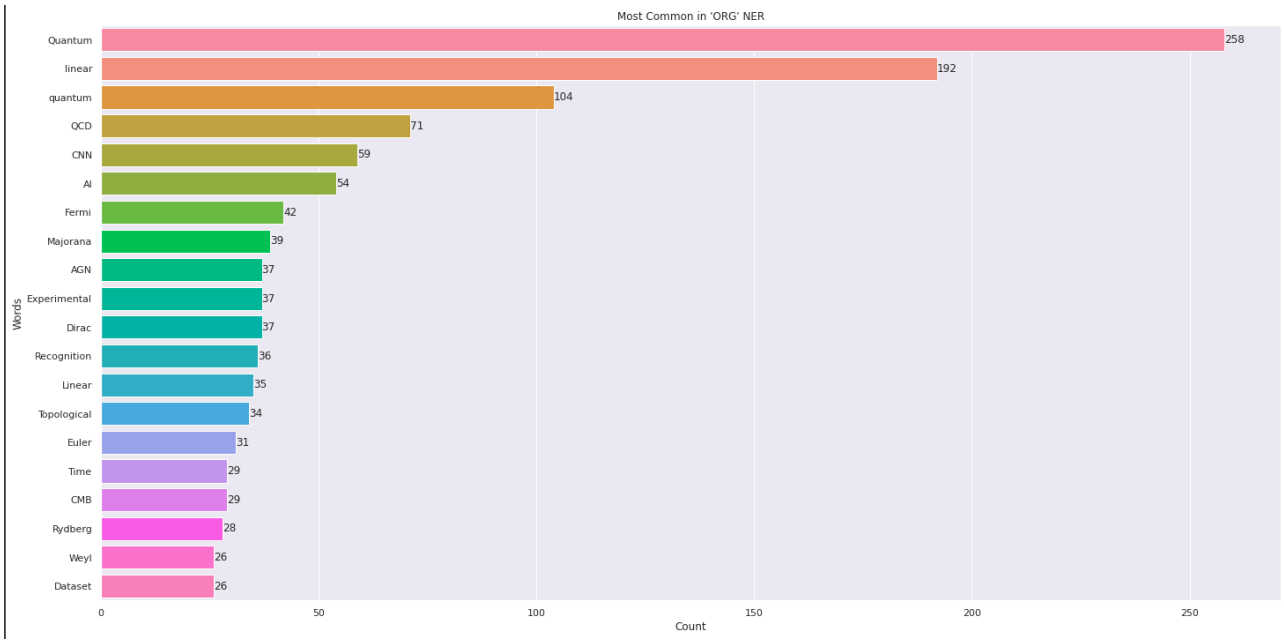

Figure 2.17: Most common word in ORG NER

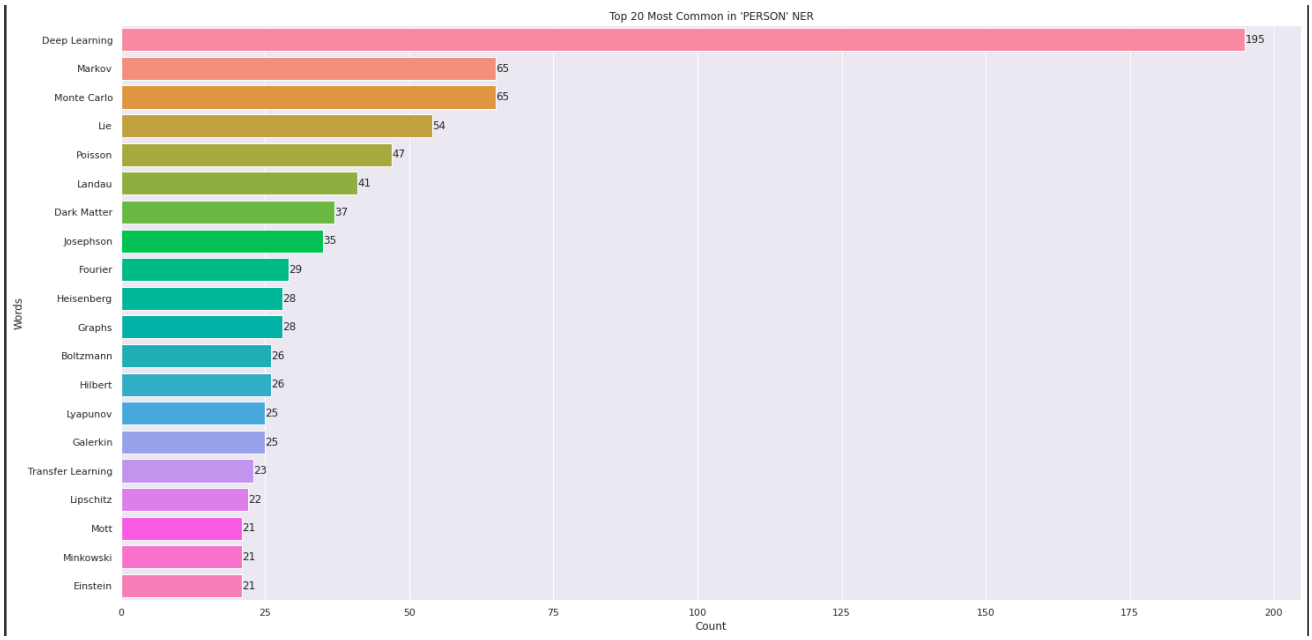

Figure 2.18: Most common word in PERSON NER

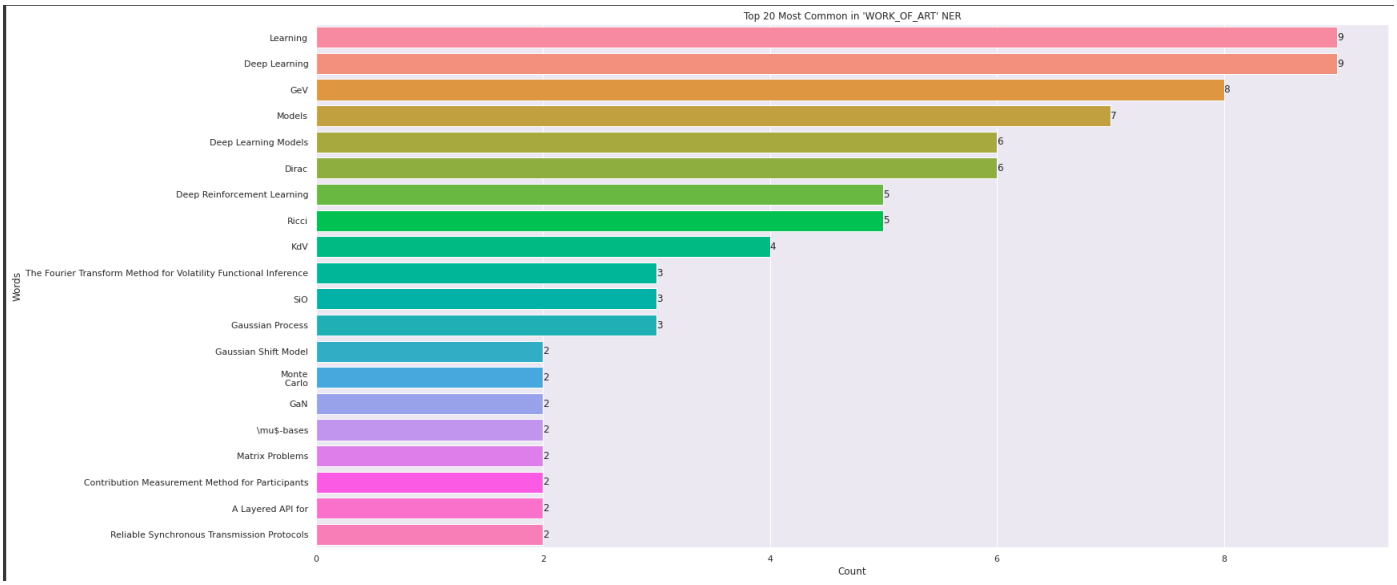

Figure 2.19: Most common words in WORK OF ART NER

2.4.2 Inference from NER

From fig 2.16, it can be deduced that ORG highly dominate the title of the articles following PERSON, CARDINAL, etc.

2.5 Part of Speech Tagging

Parts of speech (POS) tagging is a method that assigns part of speech labels to words in a sentence.

NNP	Proper Noun, Singular
NN	Noun, Singular
IN	Preposition or subordinating conjunction
JJ	Adjective
NNS	Noun, plural
DT	Determiner
CC	Coordinating conjunction
VBG	Verb, gerund or present participle
:	Mid-sentence punctuation (: ; ... – -)
$	Currency Sign

Table 2.3: POS Label and Description

2.5.1 Label based POS exploration

For the POS exploration, 1,00,000 sample of data were taken from the training data whose distribution is shown in fig 2.20.

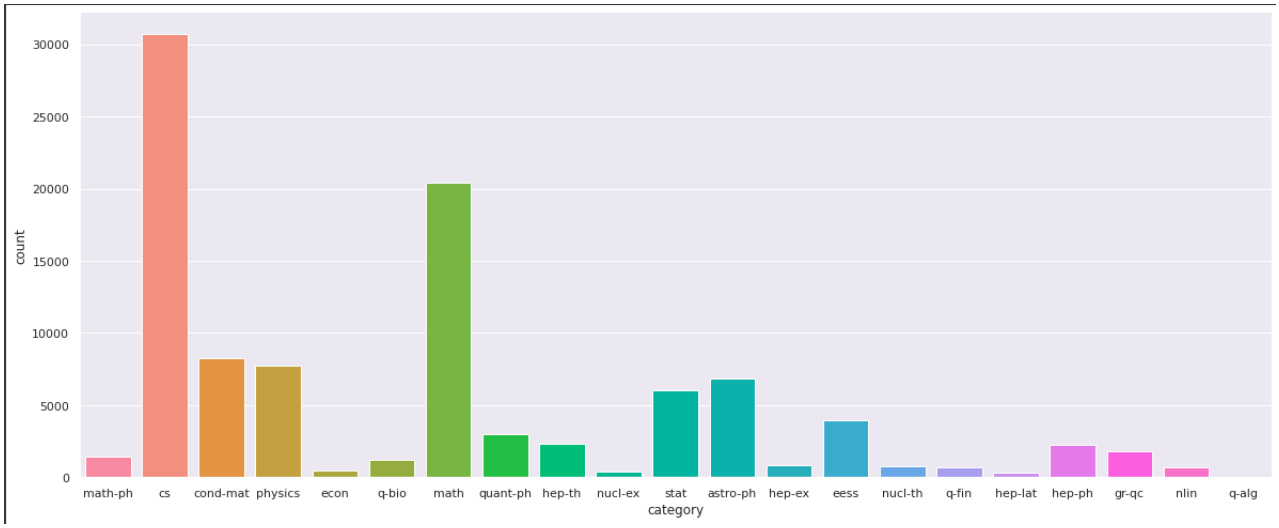

Figure 2.20: 1,00,000 sample data for POS

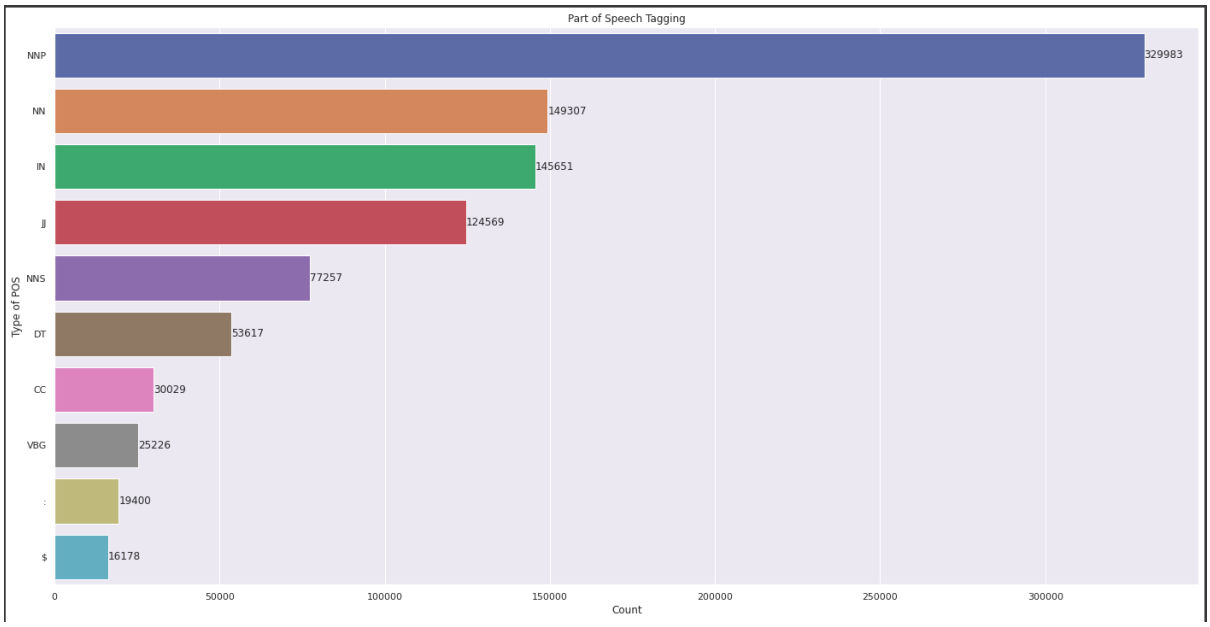

Figure 2.21: POS Countplot for the 1,00,000 sample data

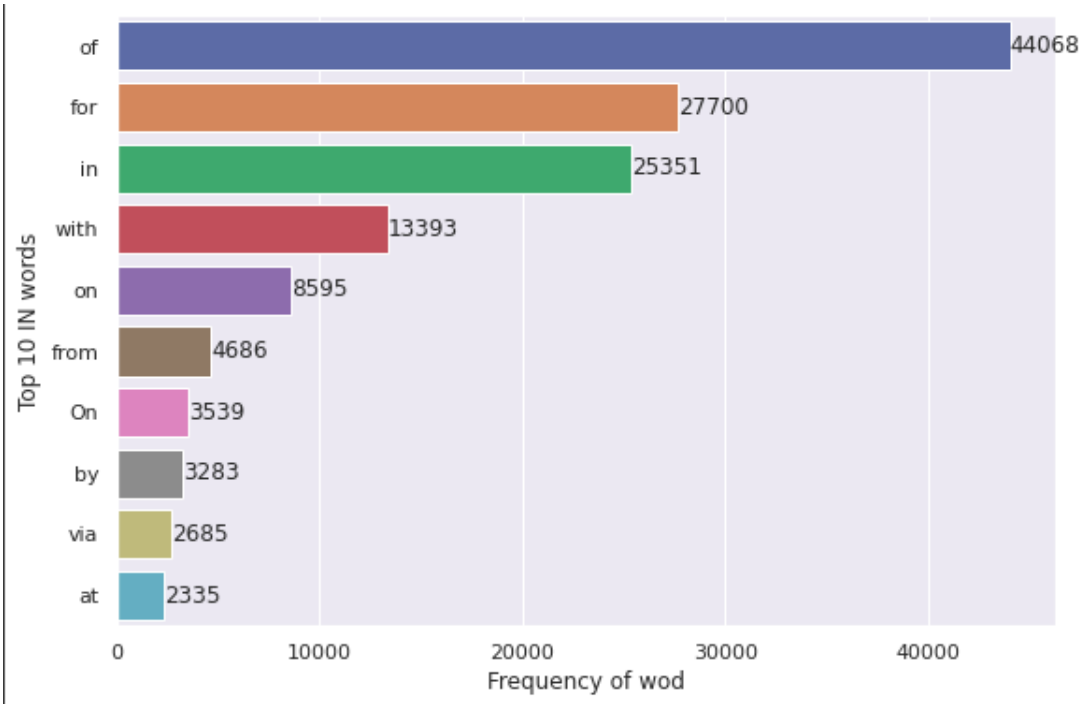

Figure 2.22: Top 10 preprosition in title

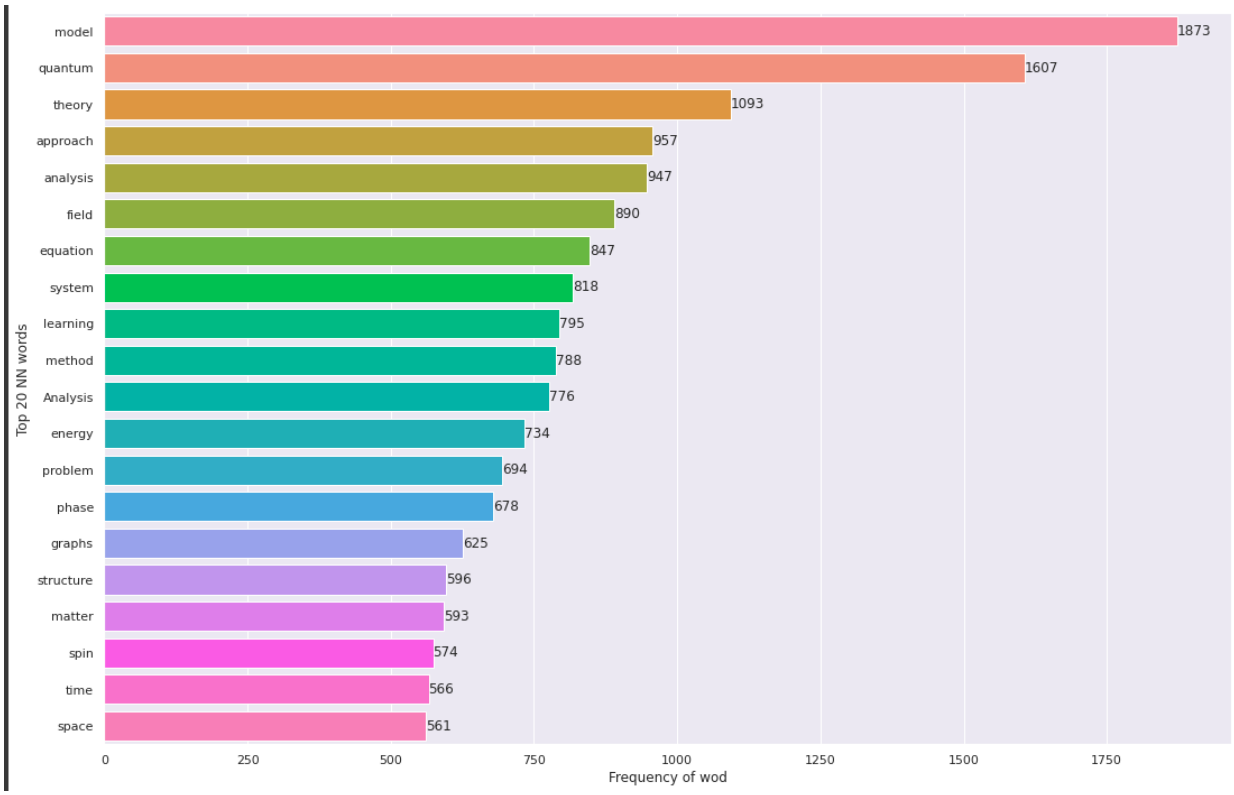

Figure 2.23: Top 20 NN in title

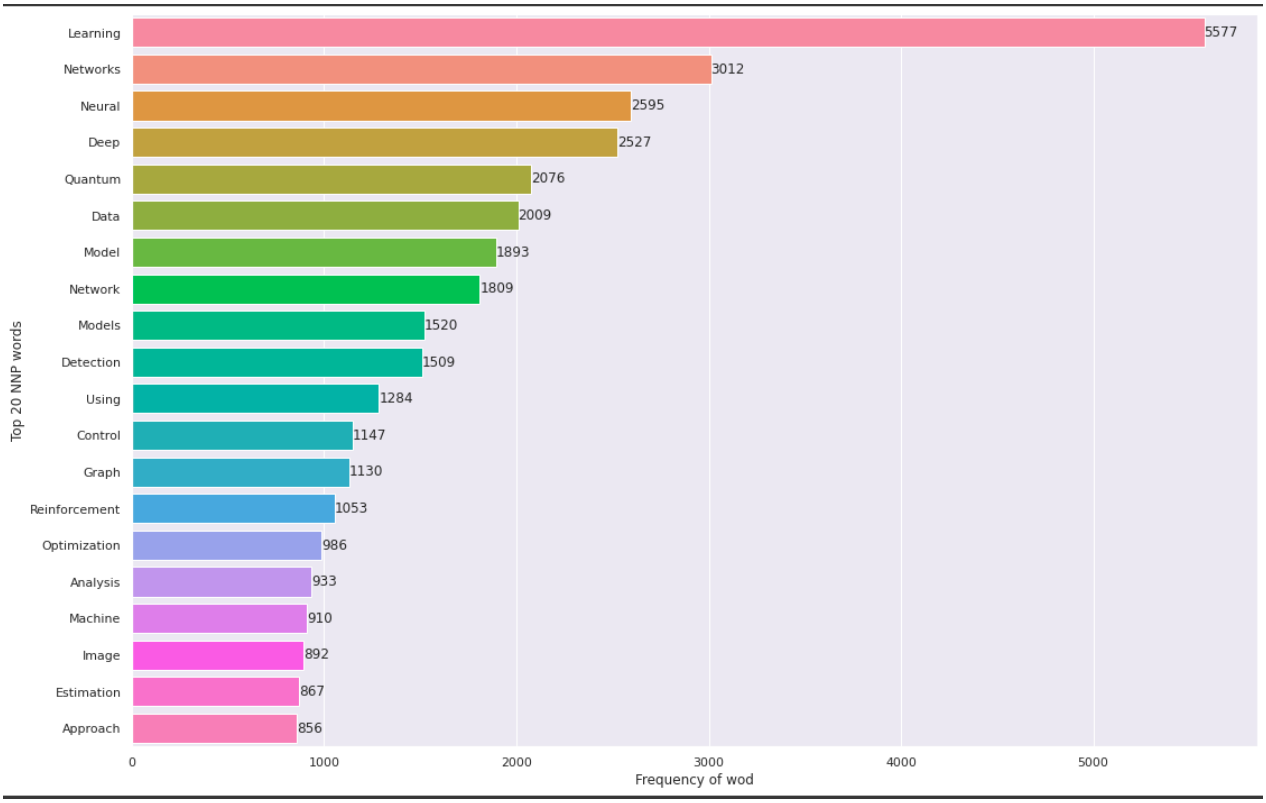

Figure 2.24: Top 20 NNP in title

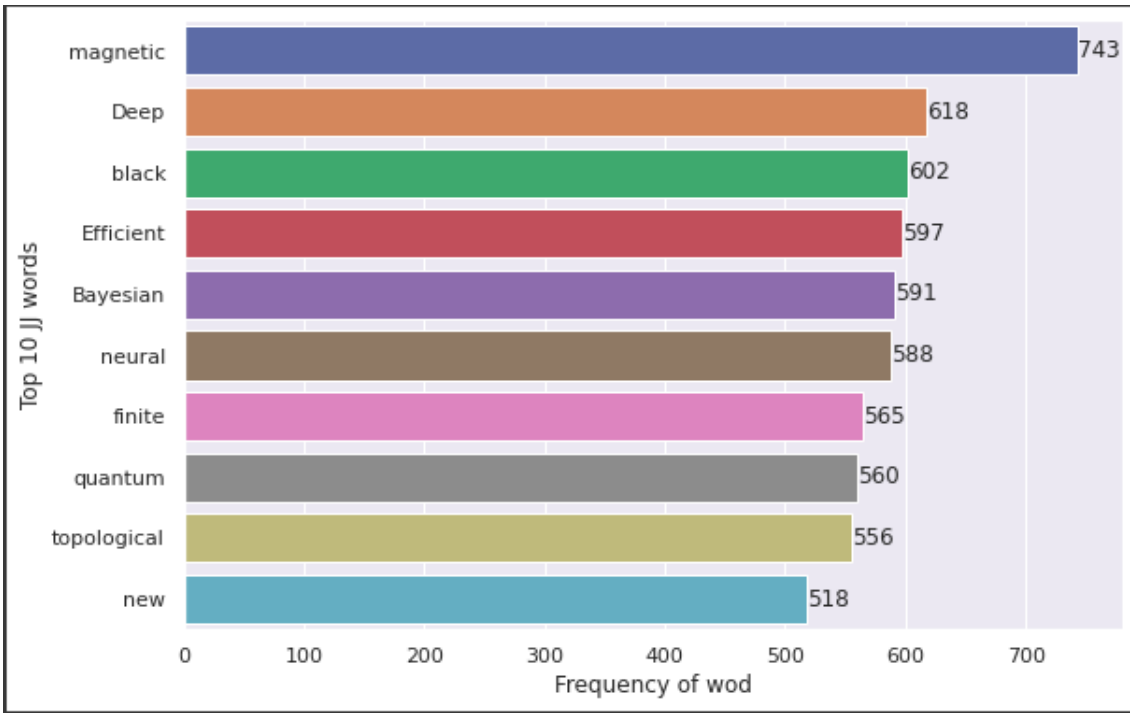

Figure 2.25: Top 10 JJ in title

2.5.2 Inference from POS exploration

From fig 2.21, proper singular noun, prepositions and adjectives were found to dominate the title of the training data which are normally the case. In NNP count plot, most words were found to be from majority class as in fig 2.20.