Scientific Text Topic Classifier

Presented by
Artem Ramus

Background, by the data publisher

Context
In India, every year lacs of students sit for competitive examinations like JEE Advanced, JEE Mains, NEET, etc. These exams are said to be the gateway to get admission into India's premier Institutes such as IITs, NITs, AIIMS, etc. Keeping in mind that the competition is tough as lacs of students appear for these examinations, there has been an enormous development in Ed Tech Industry in India, fortuning the dreams of lacs of aspirants via providing online as well as offline coaching, mentoring, etc. This particular data set consists of questions/doubts raised by students preparing for such examinations.

Content
The data set contains Students-questions.csv file in version 1 as of now. Inside the CSV file, we have two columns:
1. eng: The full question or description of the questions
2. Subject: Which subject does the question belong to. It has 4 classes, physics, chemistry, biology, and mathematics.
So, it's basically an NLP problem where we have the question description and we need to find out which subject does this question belongs to.

Introduction, by the data publisher

A note from a publisher of the data set
The goal is to classify the given texts into 4 subjects i.e, maths, physics, chemistry, and biology.

Challenges
Data cleaning (contains random special characters, symbols, expressions which might contain class dependent information. Also carries latex formulas, mathematical functions, etc.), class imbalance. Customized NLP techniques (lemmatization, stemming, stop word removal must be carried out carefully to distinguish between classes.). Overfitting.

A link to the data set
https://www.kaggle.com/mrutyunjaybiswal/iitjee-neet-aims-students-questions-data

Bag of Words modeling

Count vectorization and TF-IDF

Data cleaning

	subject	text	engineered_text
0	Biology	An anti-forest measure is\nA. Afforestation\nB...	forest measure afforestation selective grazing...
1	Chemistry	Among the following organic acids, the acid pr...	among organic acid acid present rancid butter ...
2	Maths	If the area of two similar triangles are equal...	area similar triangle equal equilateral isosce...
3	Biology	In recent year, there has been a growing\nconc...	recent year growing concern gradually increasi...
4	Physics	Which of the following statement\nregarding tr...	regarding transformer transformer make use far...

- Lemmatizing with nltk.stem.WordNetLemmatizer
- Stopwords removal with NLTK and custom words, common to all the scientific texts.

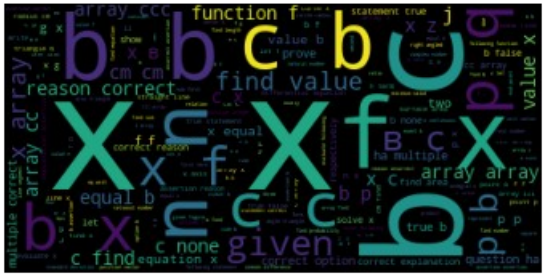
Exploratory Data Analysis – word clouds with default pipeline

After leaving only alphabetical tokens and lemmatization

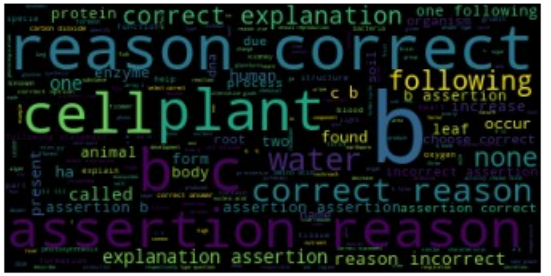
Common scientific words found in all the subjects, like ‘correct’ and ‘reason’

One-character symbols found in all the subjects.

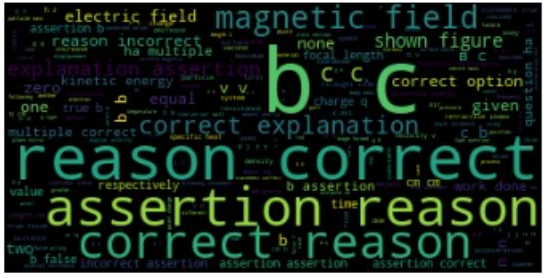
Math



Biology



Physics



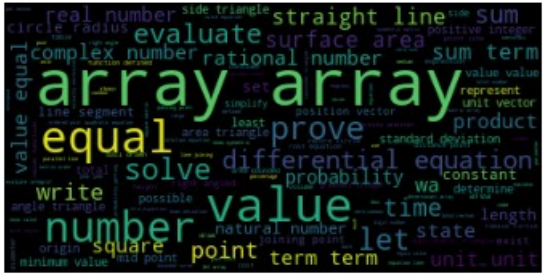
Chemistry

Feature Engineering – word clouds with customized pipeline

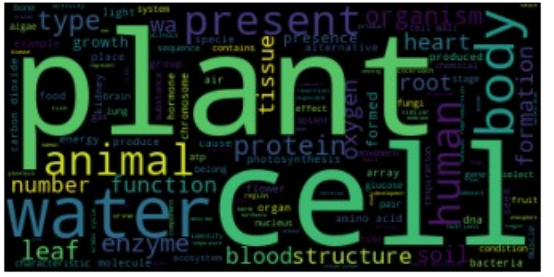
After removing short words and common scientific words

Words distributions became more meaningful and representative for the subject

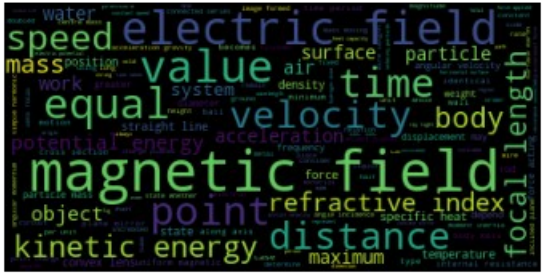
Math



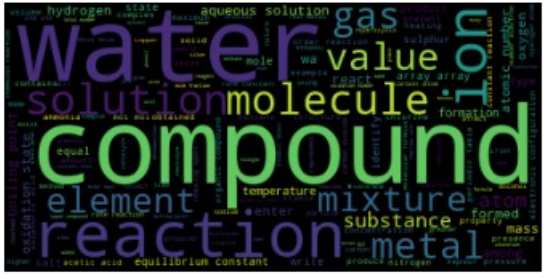
Biology



Physics



Chemistry

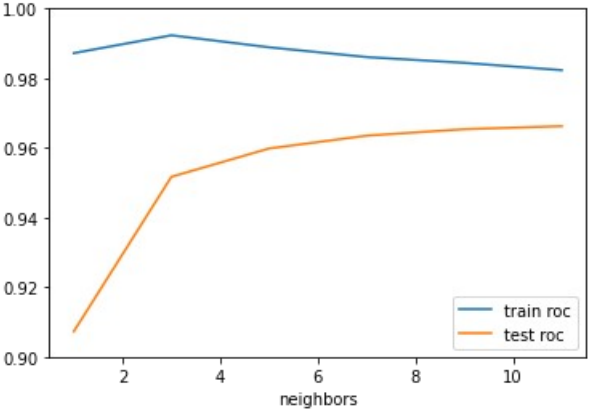


Results – overfitting mitigation

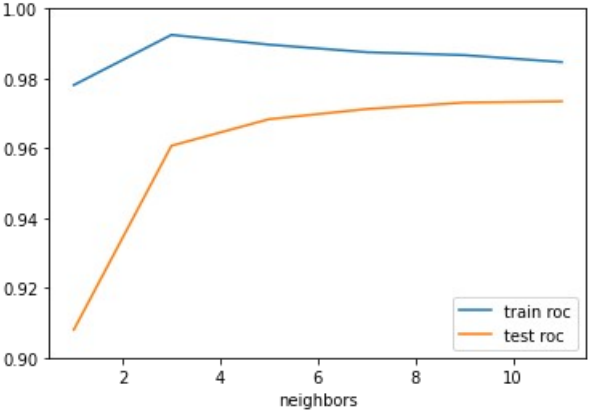
Modeling after default text preparation gave overfitting. After removal of short words and common scientific words, validity of the solution improved.

TF-IDF vectorization with 1-grams vs. 1-grams & 2-grams testes. 1-grams' scores are slightly better.

Default text preparation



After removal of short words and common scientific words

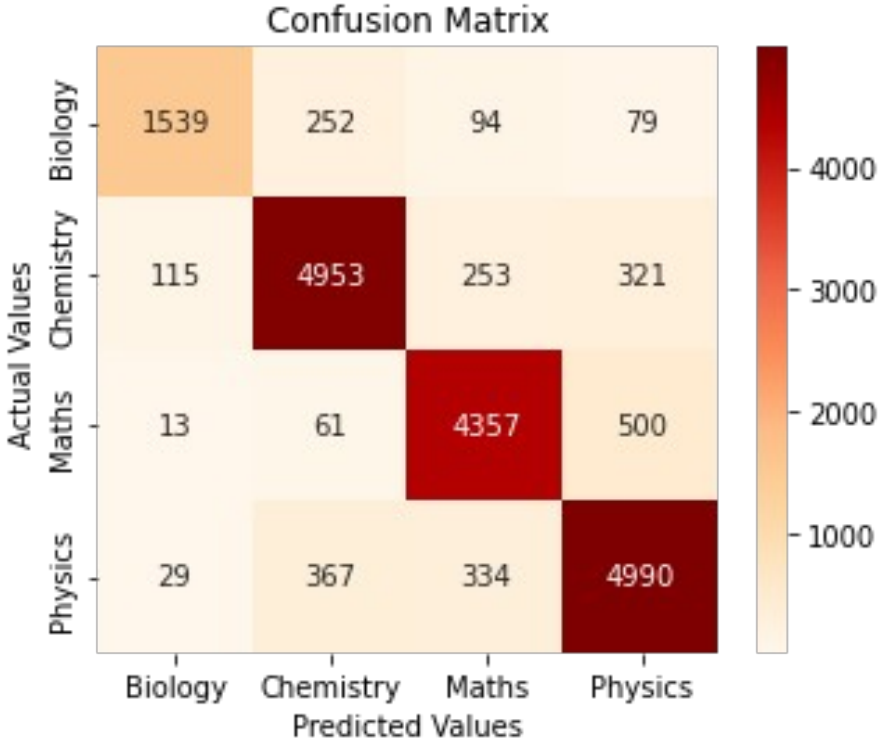


Results – confusion matrix

With 7 neighbors, one-vs-rest weighted ROC-AUC train score is 0.9875 and test - 0.9721.

Accuracy per class

Biology	1539/1964	0.78
Chemistry	4953/5642	0.88
Math	4357/4931	0.88
Physics	4990/5720	0.87



Neural Networks modeling

Transformers BertForSequenceClassification

Neural Networks modeling

Transformers BertForSequenceClassification

Neural Networks modeling

- BertTokenizer
- BERT pre-trained Model
- AdamW Optimizer, learning rate 1e-5, epsilon 1e-8
- 36 batches
- 2 epochs

Validation

Epoch 1
Training loss : 0.2170
Validation loss : 0.1573
F1 Score (Weighted) : 0.9465

Epoch 2
Training loss : 0.1204
Validation loss : 0.1416
F1 Score (Weighted) : 0.9527

Epoch 3
Training loss : 0.0867
Validation loss : 0.1516
F1 Score (Weighted) : 0.9542

Results

After second epoch

Accuracy per class

Biology	1767/1964	0.89
Chemistry	5353/5642	0.94
Math	4849/4931	0.98
Physics	5426/5720	0.94

The end

Thank you for your attention!