Subreddit Classification:

r/Physics and r/chemistry

Sophia Scarano

Primary Goal

1. Predict which subreddit a post belongs to based on its content

Overview

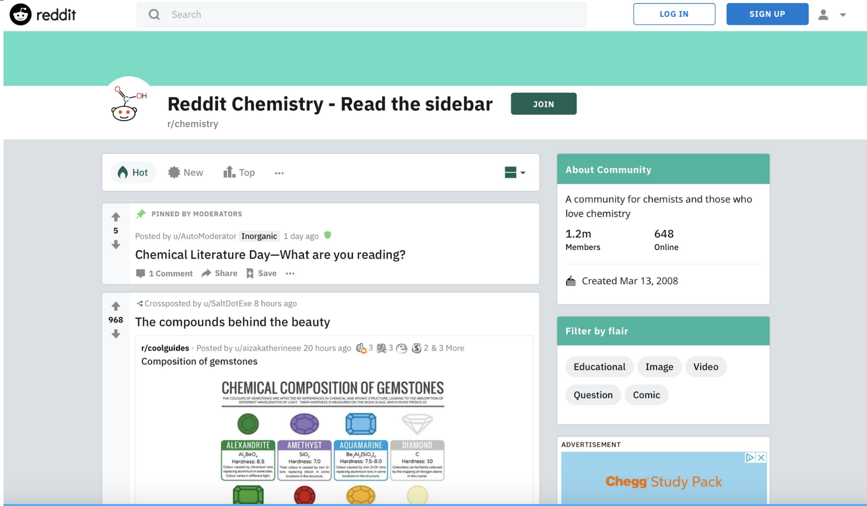
1. Get subreddit data
2. Data cleaning and preprocessing
3. Baseline Accuracy?
4. Variable selection
    a. Which features to use?
    b. How to simplify data (lemmatization, stemming)
5. Explore:
    a. Vectorizers
    b. Classifiers
6. Compare results

r/Physics r/Chemistry

- **r/Physics** and **r/chemistry**: two subreddits where redditors can ask questions pertaining to, or discuss topics of, physics or chemistry topics.

- Both of these fields have specific jargon, potentially allowing for easier discrete classification

- This allows us to build an appropriate model


reddit
Search
LOG IN
SIGN UP
JOIN
Hot
New
Top
Share
Save
Educational
Image
Video
Question
Comic
ADVERTISEMENT
Chegg Study Pack

Reddit Chemistry - Read the sidebar

r/chemistry

PINNED BY MODERATORS

5 Posted by u/AutoModerator Inorganic 1 day ago

Chemical Literature Day—What are you reading?

1 Comment

968 Crossposted by u/SaltDotExe 8 hours ago

The compounds behind the beauty

r/coolguides · Posted by u/aizakatherineee 20 hours ago 3 3 2 & 3 More

Composition of gemstones

CHEMICAL COMPOSITION OF GEMSTONES
THE COLOURS OF GEMSTONES ARE AFFECTED BY DIFFERENCES IN CHEMICAL AND ATOMIC STRUCTURE, LEADING TO THE ABSORPTION OF DIFFERENT WAVELENGTHS OF LIGHT. THEIR HARDNESS IS MEASURED ON THE MOHS SCALE, WHICH RUNS FROM 0-10.
ALEXANDRITE
Al_2BeO_4
Hardness: 8.5
Colour caused by chromium ions replacing aluminium in some sites. Colour varies in different light.
AMETHYST
SiO_2
Hardness: 7.0
Their colour is caused by iron 3+ ions replacing silicon in some locations in the structure.
AQUAMARINE
$Be_3Al_2(SiO_3)_6$
Hardness: 7.5-8.0
Colour caused by iron 2+/3+ ions replacing aluminium ions in some locations in the structure.
DIAMOND
C
Hardness: 10
Colourless; can be faintly coloured by the trapping of nitrogen atoms in the crystal.

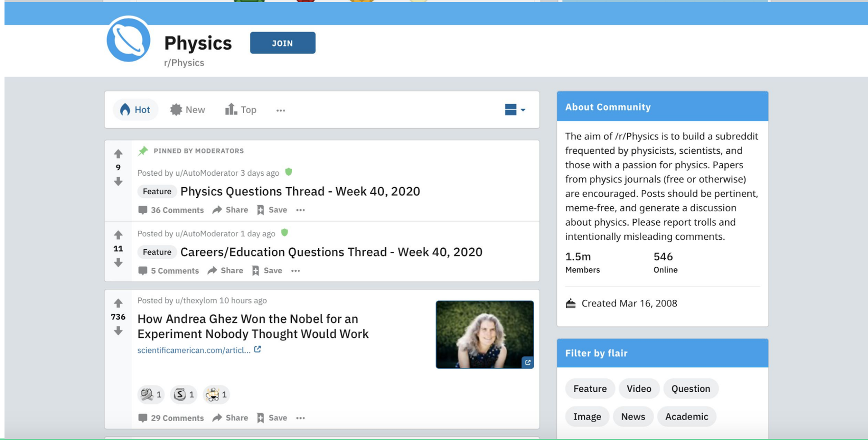
About Community

A community for chemists and those who love chemistry

1.2m Members 648 Online

Created Mar 13, 2008

Filter by flair



Physics

r/Physics

PINNED BY MODERATORS

9 Posted by u/AutoModerator 3 days ago

Feature Physics Questions Thread - Week 40, 2020

36 Comments Share Save

11 Posted by u/AutoModerator 1 day ago

Feature Careers/Education Questions Thread - Week 40, 2020

5 Comments Share Save

736 Posted by u/thexylom 10 hours ago

How Andrea Ghez Won the Nobel for an Experiment Nobody Thought Would Work

scientificamerican.com/articl...

1 1 1

29 Comments Share Save

About Community

The aim of /r/Physics is to build a subreddit frequented by physicists, scientists, and those with a passion for physics. Papers from physics journals (free or otherwise) are encouraged. Posts should be pertinent, meme-free, and generate a discussion about physics. Please report trolls and intentionally misleading comments.

1.5m Members 546 Online

Created Mar 16, 2008

Filter by flair

Factors to prioritize

1. Post text content
2. Post title

Data Cleaning and Preprocessing

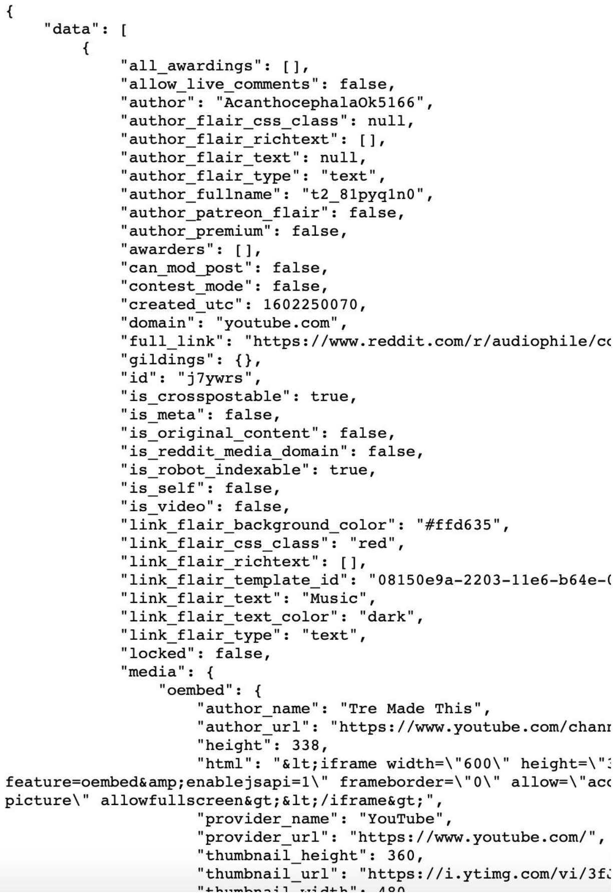
Methodology

Build a Set of Reddit Comments

- **Reddit API?**
  - Strict limitations on amount of data we can access (can only pull 100 of the most recent comments)

- **Pushshift API**
  - Open-source alternative to Reddit's API
  - Can specify date and time we start our pull from
  - Data returned as list of dictionaries

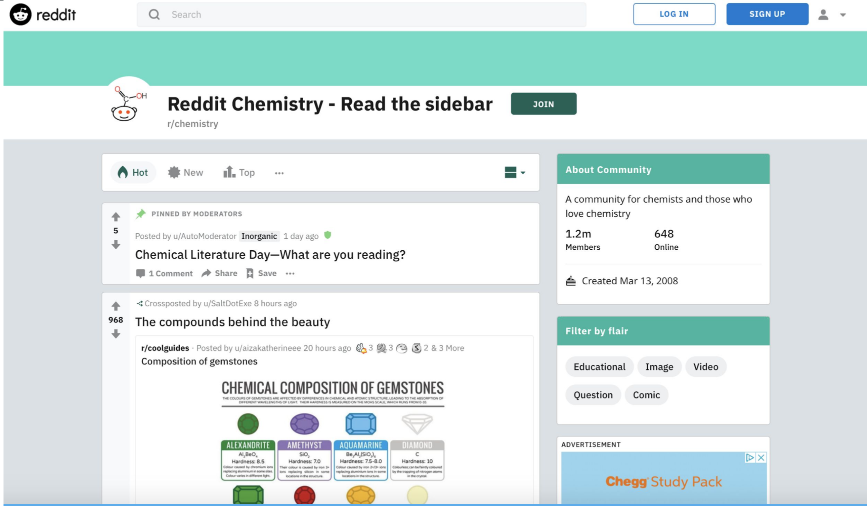
- **My data**
  - All comments for each subreddit in the past two years



```
{
    "data": [
        {
            "all_awardings": [],
            "allow_live_comments": false,
            "author": "AcanthocephalaOk5166",
            "author_flair_css_class": null,
            "author_flair_richtext": [],
            "author_flair_text": null,
            "author_flair_type": "text",
            "author_fullname": "t2_81pyq1n0",
            "author_patreon_flair": false,
            "author_premium": false,
            "awarders": [],
            "can_mod_post": false,
            "contest_mode": false,
            "created_utc": 1602250070,
            "domain": "youtube.com",
            "full_link": "https://www.reddit.com/r/audiophile/c
            "gildings": {},
            "id": "j7ywrs",
            "is_crosspostable": true,
            "is_meta": false,
            "is_original_content": false,
            "is_reddit_media_domain": false,
            "is_robot_indexable": true,
            "is_self": false,
            "is_video": false,
            "link_flair_background_color": "#ffd635",
            "link_flair_css_class": "red",
            "link_flair_richtext": [],
            "link_flair_template_id": "08150e9a-2203-11e6-b64e-
            "link_flair_text": "Music",
            "link_flair_text_color": "dark",
            "link_flair_type": "text",
            "locked": false,
            "media": {
                "oembed": {
                    "author_name": "Tre Made This",
                    "author_url": "https://www.youtube.com/chan
                    "height": 338,
                    "html": "&lt;iframe width=\"600\" height=\"
feature=oembed&amp;enablejsapi=1\" frameborder=\"0\" allow=\"ac
picture\" allowfullscreen&gt;&lt;/iframe&gt;",
                    "provider_name": "YouTube",
                    "provider_url": "https://www.youtube.com/",
                    "thumbnail_height": 360,
                    "thumbnail_url": "https://i.ytimg.com/vi/3f
                    "thumbnail_width": 480
```

Data Cleaning

- **Only keep rows:**
  - Where ‘selftext’ and ‘title’ are present, and have at least 4 words
- **Remove**:
  - Special characters
  - Any word that is an overly common word, or ‘stop word’ in the english language
    - Reduces noise
  - Non-letter characters
- **Expand contractions**
  - Didn’t end up working
- **Tokenizing**
  - Easier for processing
- **Lemmatizing**
  - Keep roots of words
- **Stemming**
  - Crude version of lemmatizing



Reddit Chemistry - Read the sidebar

r/chemistry

PINNED BY MODERATORS

5

Posted by u/AutoModerator Inorganic 1 day ago

Chemical Literature Day—What are you reading?

968

Crossposted by u/SaltDotExe 8 hours ago

The compounds behind the beauty

r/coolguides · Posted by u/aizakatherineee 20 hours ago 3 3 2 & 3 More

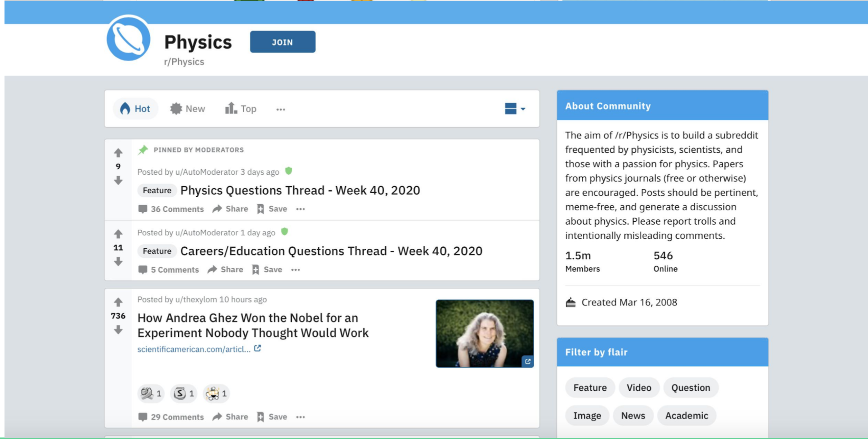
Composition of gemstones

About Community

A community for chemists and those who love chemistry

1.2m	648
Members	Online

Created Mar 13, 2008



Physics

r/Physics

PINNED BY MODERATORS

9

Posted by u/AutoModerator 3 days ago

Feature Physics Questions Thread - Week 40, 2020

11

Posted by u/AutoModerator 1 day ago

Feature Careers/Education Questions Thread - Week 40, 2020

736

Posted by u/thexylom 10 hours ago

How Andrea Ghez Won the Nobel for an Experiment Nobody Thought Would Work

scientificamerican.com/articl...

1 1 1

About Community

The aim of /r/Physics is to build a subreddit frequented by physicists, scientists, and those with a passion for physics. Papers from physics journals (free or otherwise) are encouraged. Posts should be pertinent, meme-free, and generate a discussion about physics. Please report trolls and intentionally misleading comments.

1.5m	546
Members	Online

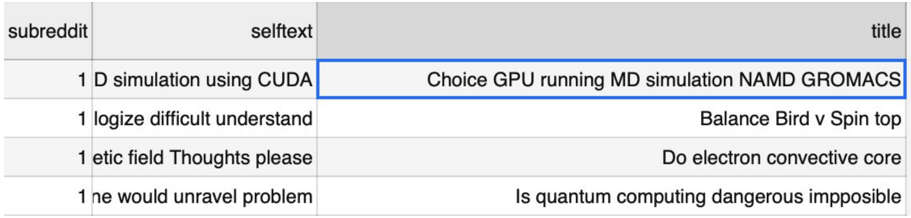
Created Mar 16, 2008

Data Preprocessing

- **Created CSV files:**
  - Lemmatized data (removed stop words)
  - Stemmed data (removed stop words)

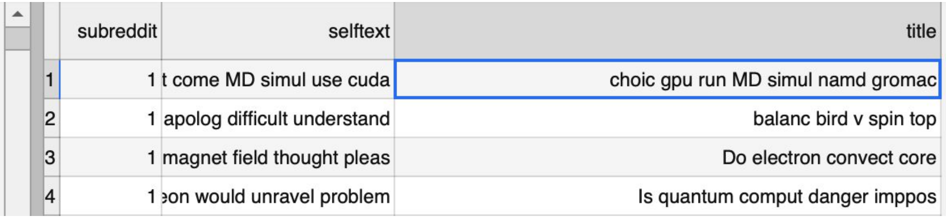
- **Kept columns**:
  - Subreddit
  - Selftext
  - Title

LEMMATIZATION



subreddit	selftext	title
1	D simulation using CUDA	Choice GPU running MD simulation NAMD GROMACS
1	logize difficult understand	Balance Bird v Spin top
1	etic field Thoughts please	Do electron convective core
1	ne would unravel problem	Is quantum computing dangerous imppossible

STEMMING



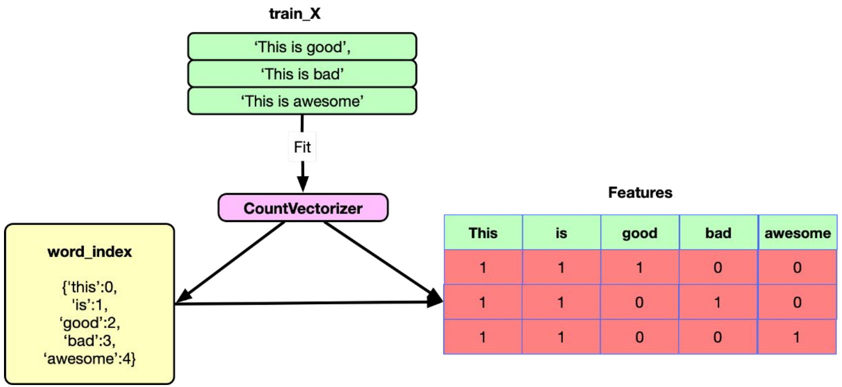
	subreddit	selftext	title
1	1	t come MD simul use cuda	choic gpu run MD simul namd gromac
2	1	apolog difficult understand	balanc bird v spin top
3	1	magnet field thought pleas	Do electron convect core
4	1	eon would unravel problem	Is quantum comput danger imppos

Natural Language Processing

1. Naive Bayes
2. Linear Regression

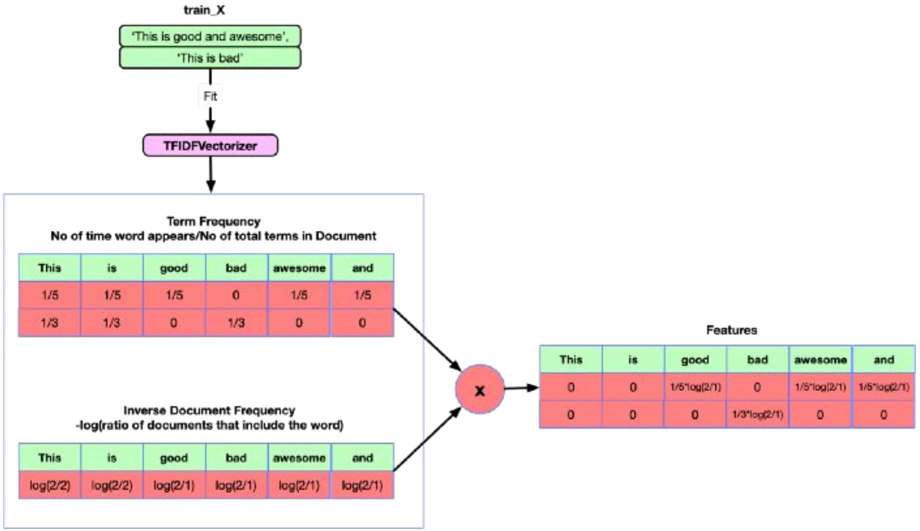
Vectorizers

COUNTVECTORIZER



This	is	good	bad	awesome
1	1	1	0	0
1	1	0	1	0
1	1	0	0	1

TFIDFVECTORIZER



Term Frequency
No of time word appears/No of total terms in Document

This	is	good	bad	awesome	and
1/5	1/5	1/5	0	1/5	1/5
1/3	1/3	0	1/3	0	0

Inverse Document Frequency
-log(ratio of documents that include the word)

This	is	good	bad	awesome	and
log(2/2)	log(2/2)	log(2/1)	log(2/1)	log(2/1)	log(2/1)

Features

This	is	good	bad	awesome	and
0	0	1/5*log(2/1)	0	1/5*log(2/1)	1/5*log(2/1)
0	0	0	1/3*log(2/1)	0	0

Vectorizers

- **Bag of Words:**
  - Use a vectorizer to split comment into words
  - Convert each comment into a vector of word frequencies
- **CountVectorizer**
  - Creates pure frequency vector
- **TfidfVectorizer**
  - Normalizes frequencies
  - Up-weighting rare words
    - Good for jargon identification

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

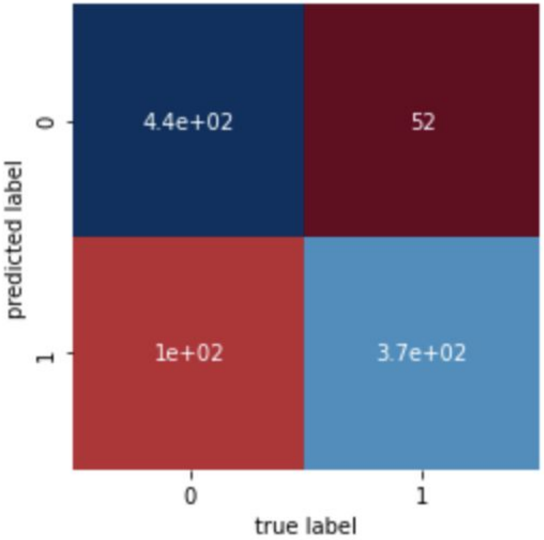
df_x = number of documents containing x

N = total number of documents

Modeling with Naive Bayes

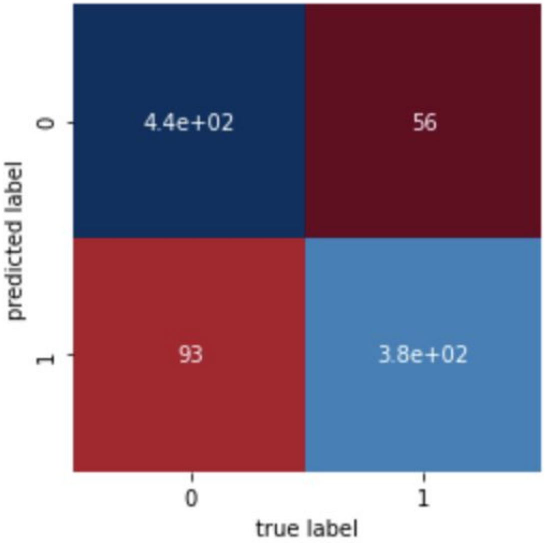
LEMMATIZED

Accuracy score: 0.84
Precision score: 0.88
Recall score: 0.78

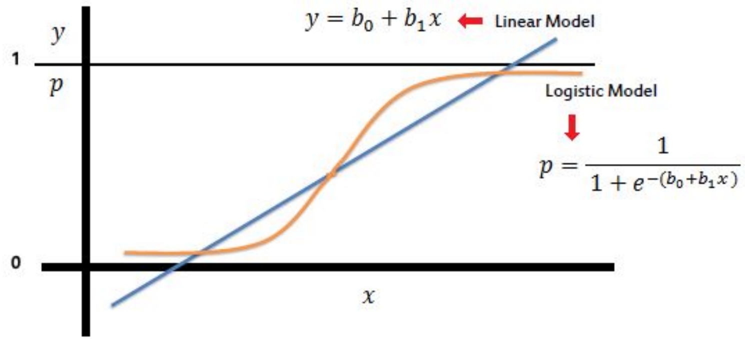


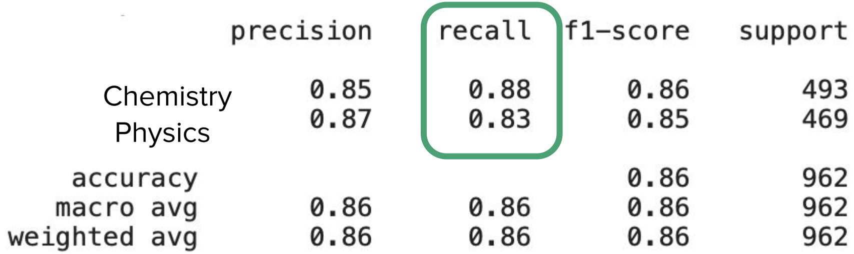
STEMMED

Accuracy score: 0.85
Precision score: 0.87
Recall score: 0.80

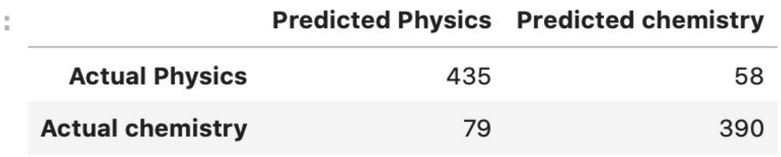


Logistic Regression: CountVectorizer





```
               precision    recall  f1-score   support

   Chemistry        0.85      0.88      0.86       493
     Physics        0.87      0.83      0.85       469

    accuracy                            0.86       962
   macro avg        0.86      0.86      0.86       962
weighted avg        0.86      0.86      0.86       962
```



	Predicted Physics	Predicted chemistry
Actual Physics	435	58
Actual chemistry	79	390

- **Logistic regression: classic technique**
  - Here with Stemmed data
  - Highly interpretable
- Can look at words or n-grams the model associates most to a subreddit

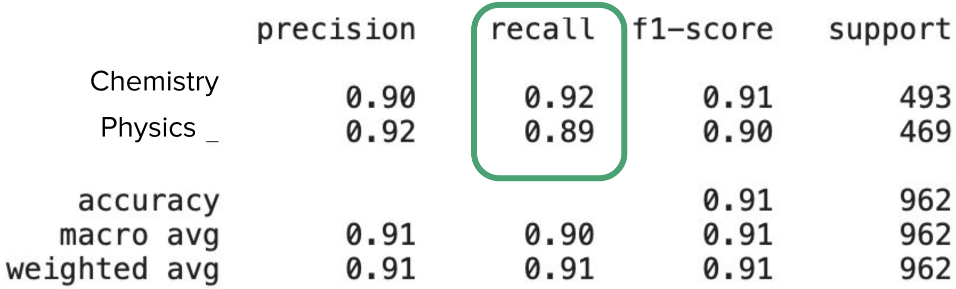
Outperformed Naive Bayes (recall score .80)

Logistic Regression: TfidfVectorizer

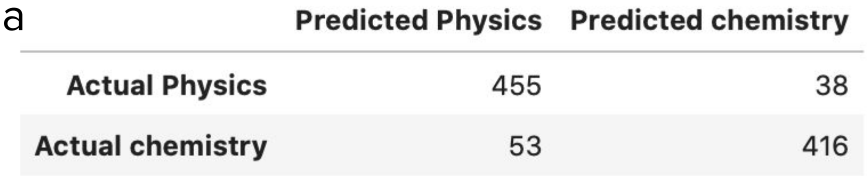
- **TfidfVectorizer:**
    - Evaluates how relevant a word is to a collection of documents

Best performance so far

- Uses Tfidf Vectorization with normalization
- Ridge regulation with a strength of alpha = 1
- Excludes stop words
- Includes all words (n-grams that appear in at least one comment)



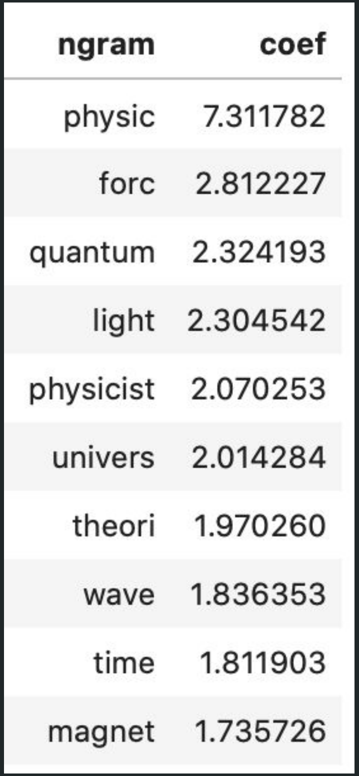
```
              precision    recall  f1-score   support

   Chemistry       0.90      0.92      0.91       493
   Physics _       0.92      0.89      0.90       469

    accuracy                           0.91       962
   macro avg       0.91      0.90      0.91       962
weighted avg       0.91      0.91      0.91       962
```



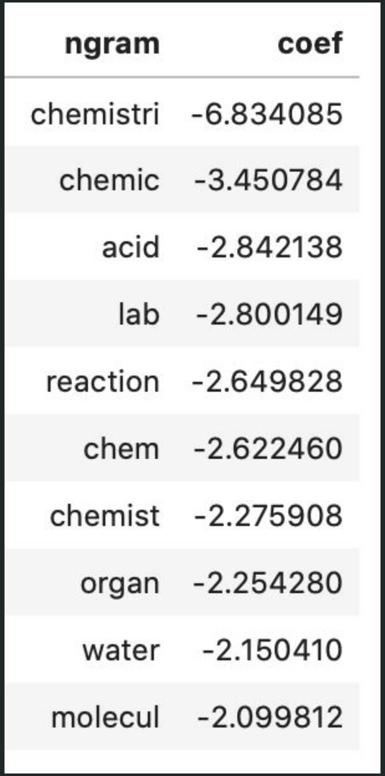
	Predicted Physics	Predicted chemistry
Actual Physics	455	38
Actual chemistry	53	416

Comparisons

- Naive Bayes:
  - Lemmatizing: .78
  - Stemming: .80
- Logistic Regression:
  - CountVectorizer:
    - Chemistry: .88
    - Physics: .83
  - Tfidf:
    - Chemistry: .92
    - Physics: .89

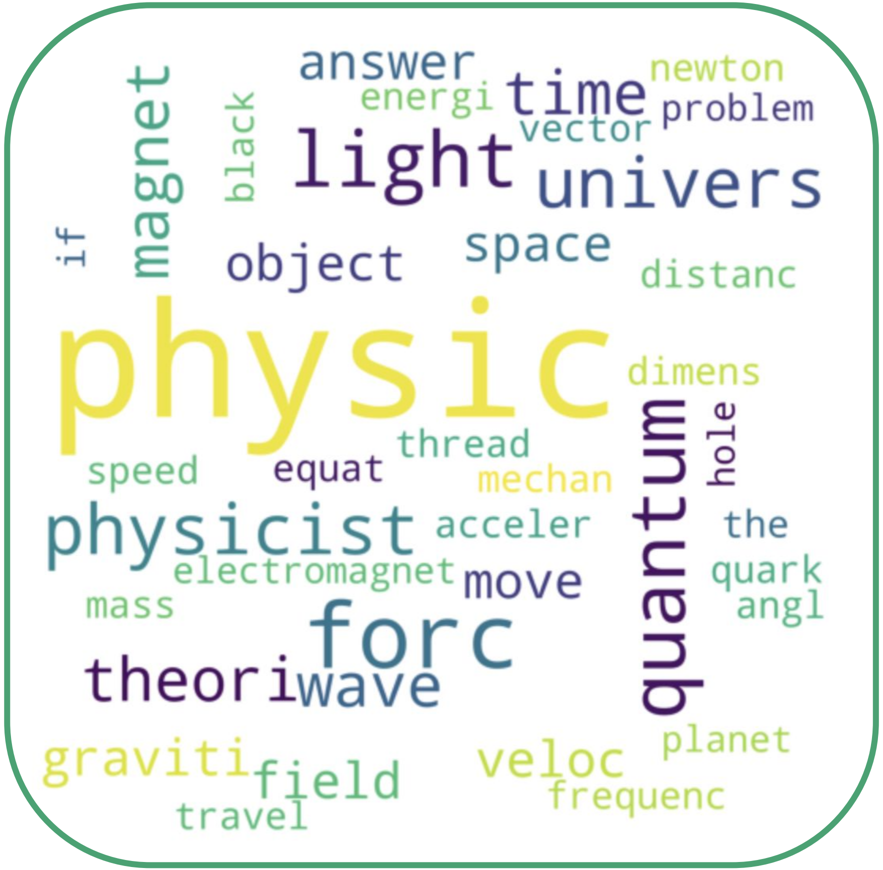


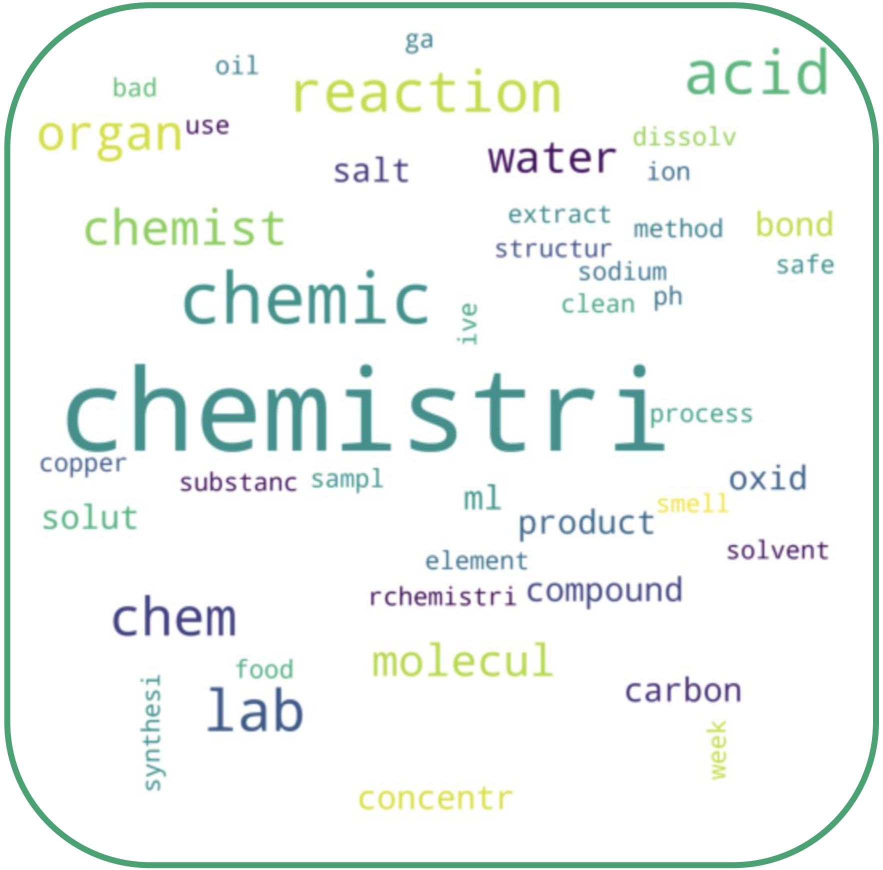
ngram	coef
physic	7.311782
forc	2.812227
quantum	2.324193
light	2.304542
physicist	2.070253
univers	2.014284
theori	1.970260
wave	1.836353
time	1.811903
magnet	1.735726



ngram	coef
chemistri	-6.834085
chemic	-3.450784
acid	-2.842138
lab	-2.800149
reaction	-2.649828
chem	-2.622460
chemist	-2.275908
organ	-2.254280
water	-2.150410
molecul	-2.099812

Common Word Roots





Future Directions

1. Try with just using ‘title’ as a feature

2. More classification models:
   a. Random Forest?

3. Get more data

Sources

- https://www.mygreatlearning.com/blog/multinomial-naive-bayes-explained/
- https://towardsdatascience.com/naive-bayes-document-classification-in-python-e33ff50f937e
- https://monkeylearn.com/blog/what-is-tf-idf/
-