



CREATING A SUBREDDIT RELEVANCE TOOL



r/science vs r/philosophy

GROUP 5 - PROJECT 3
JANET - JIE YONG - MARK - WAFIR

Who are we?

Science Subreddit
Moderators

Janet



Mark



Philosophy Subreddit
Moderators

Jie Yong



Wafir



Who are you?



28.4 million



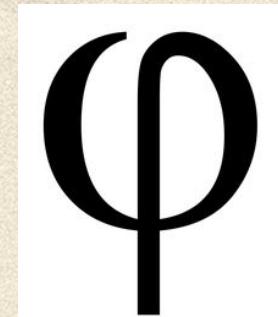
φ

16.9 million

The Problem



Science



Philosophy



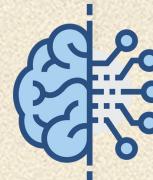
The Problem.

According to the [nytimes](#), for roughly 98 percent of the last 2,500 years of Western intellectual history, philosophy was considered the mother of all knowledge.

Today, science, not philosophy, has taken up the mantle as the world's de-facto source of truth, with some no longer sure what philosophy is or is good for anymore.

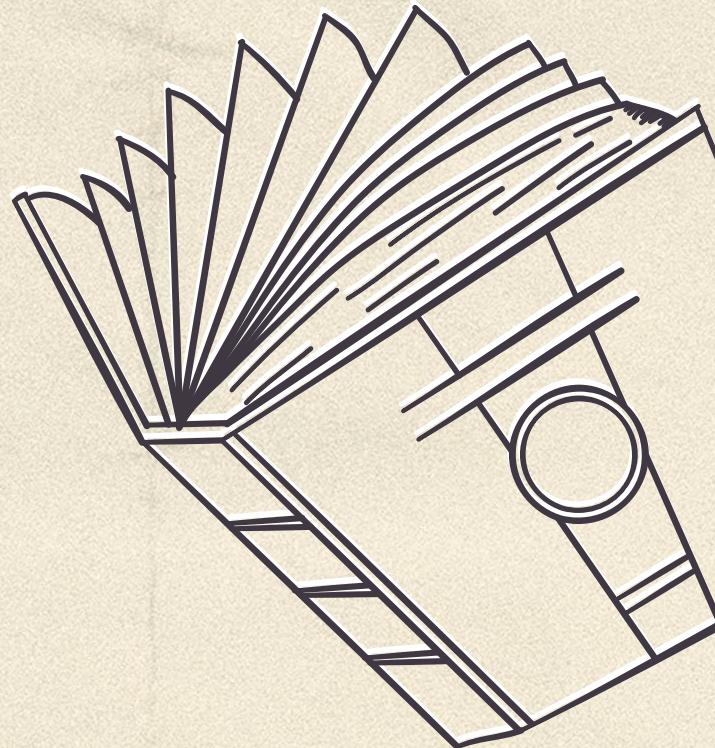
This begs the question: In which subreddit, would your thought/article/post be best placed in today's context?

Science or **Philosophy**?

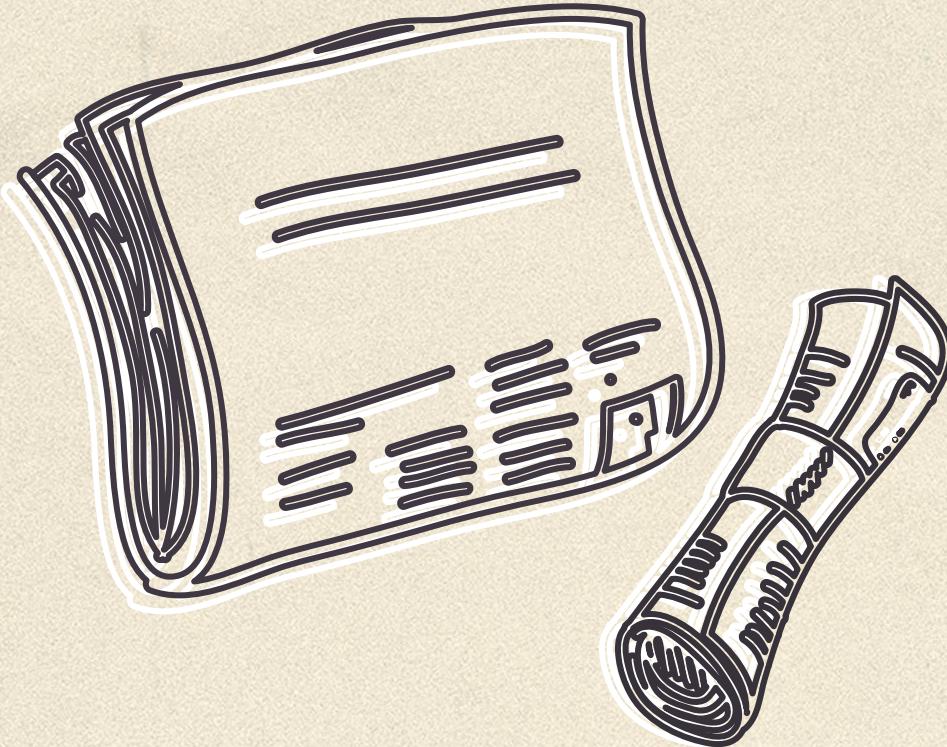


Agenda

- 01** Introduction (*Jie Yong*)
- 02** Data Extraction (*Jie Yong*)
- 03** Exploratory Data Analysis (*Wafir*)
- 04** Modelling (*Mark*)
- 05** Conclusions (*Janet*)
- 06** Next Steps (*Janet*)



Data Extraction



Pushshift API

API used to download from reddit

25,000

Posts downloaded from Science and Philosophy Subreddits each

**4 Oct 22,
12:00am**

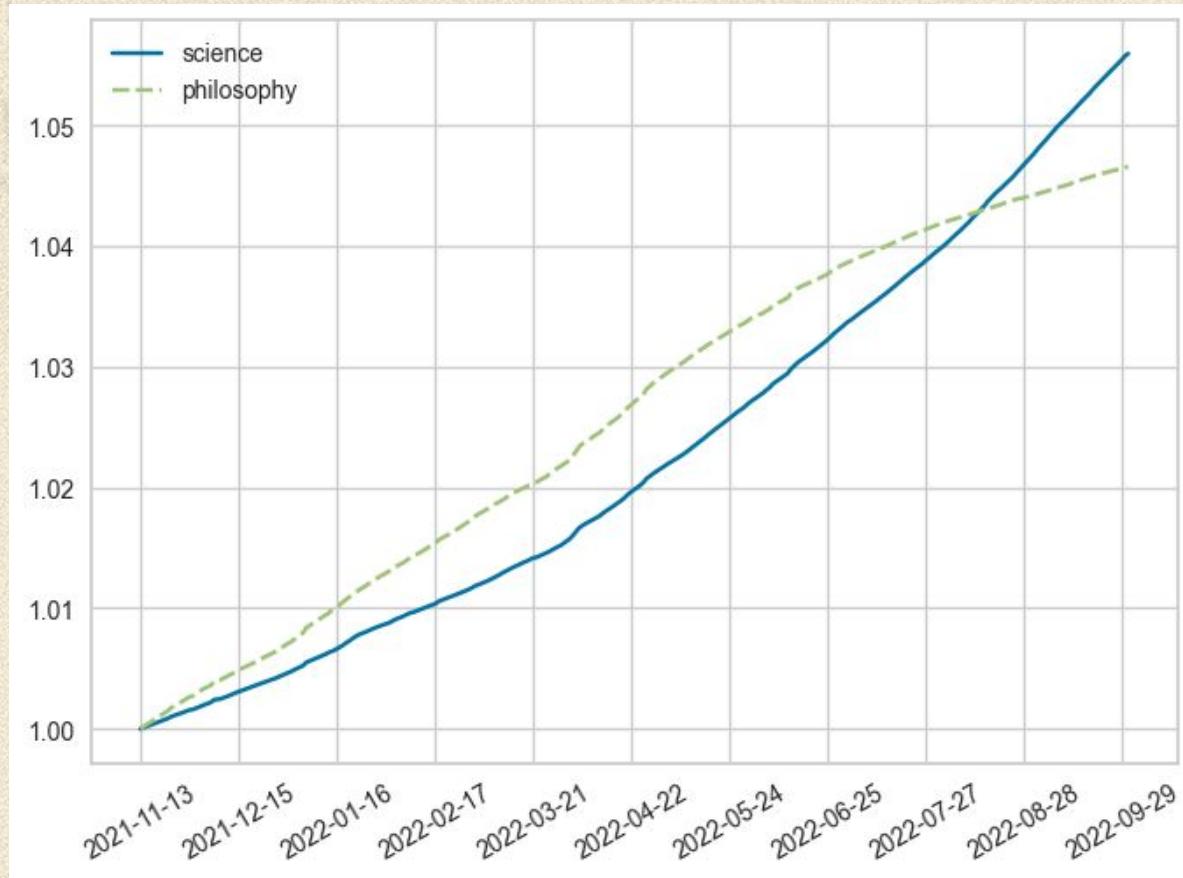
Datetime scrapping was done

03

Exploratory Data Analysis



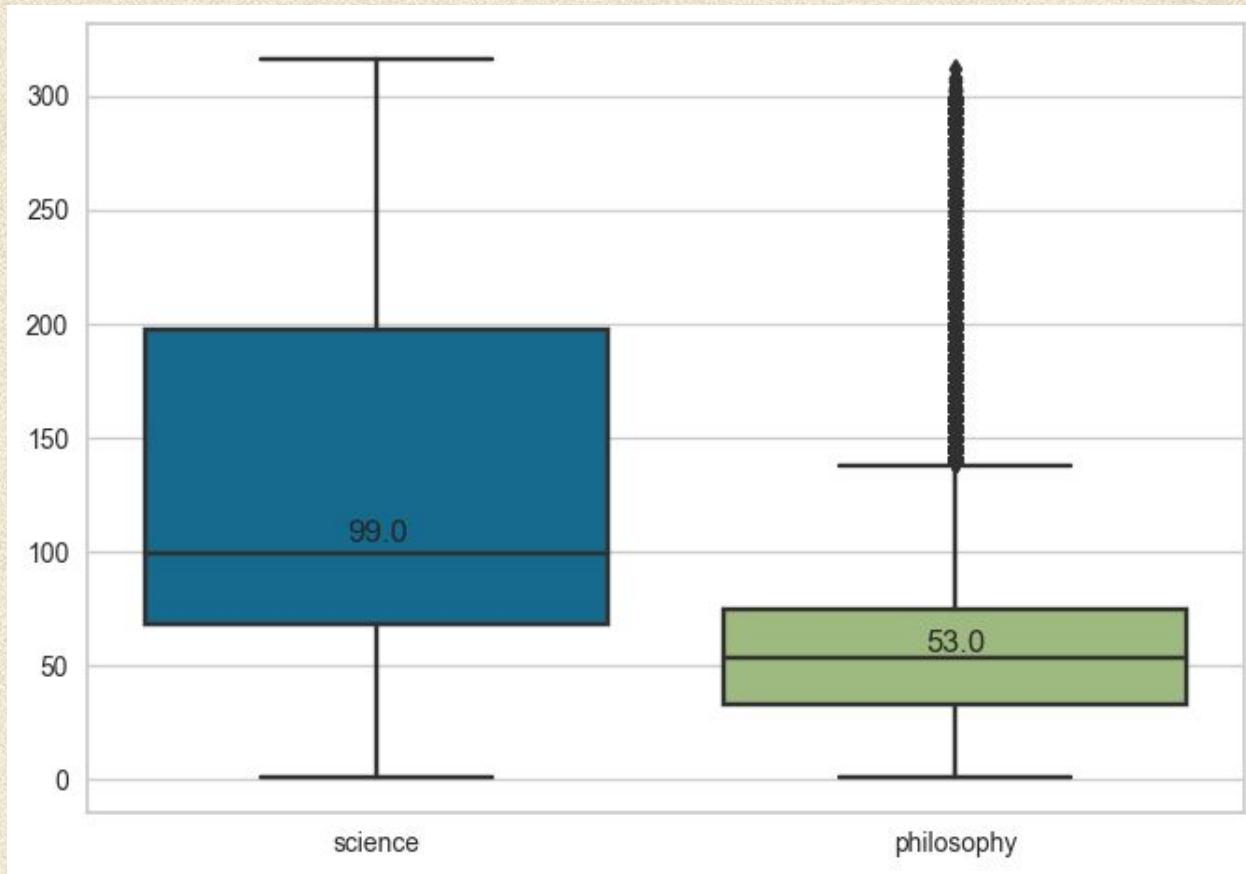
r/Science vs r/Philosophy



Member Count:

r/Science - 28.4m
r/Philosophy - 16.9m

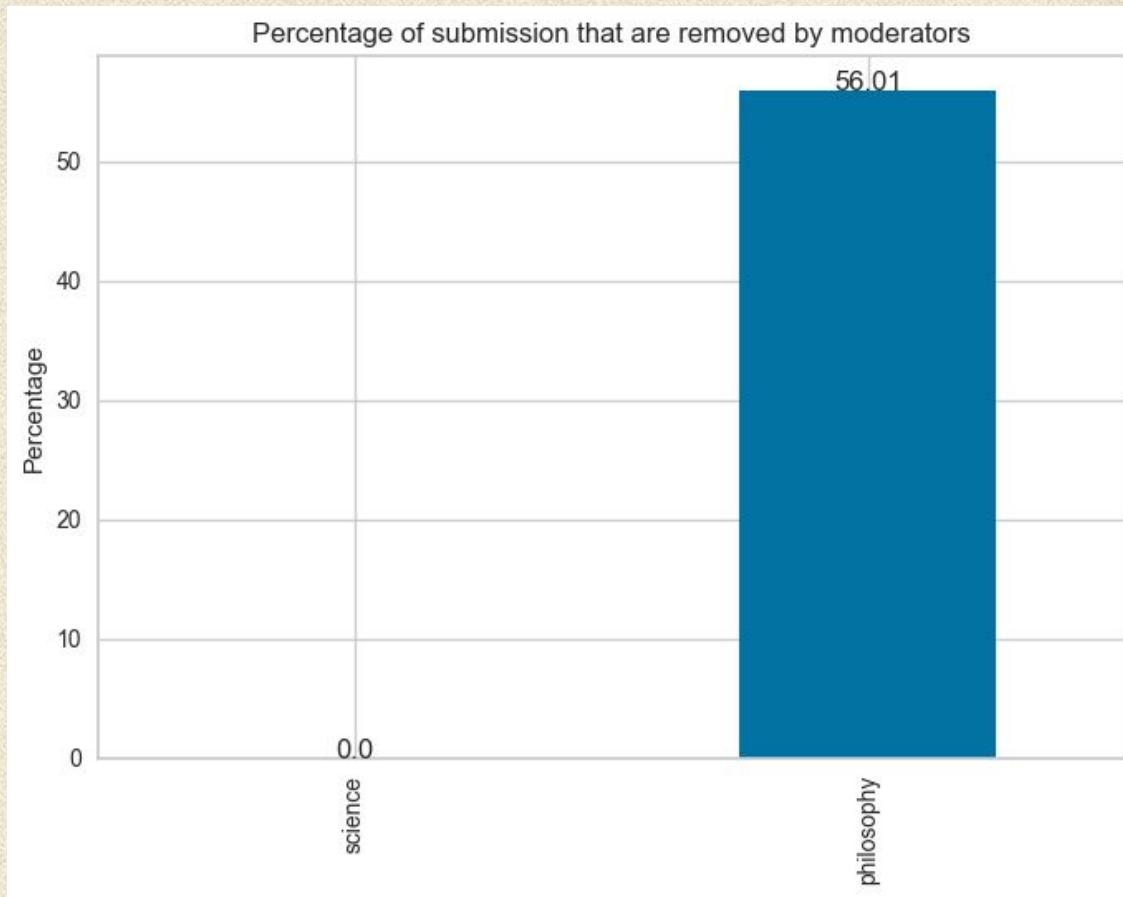
Submission Length



Number of Comments

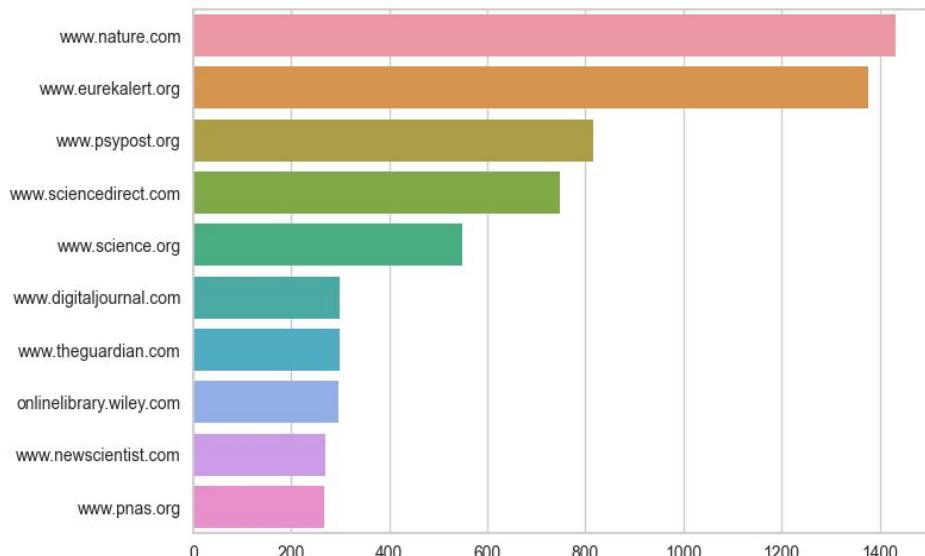


Subreddit Moderation

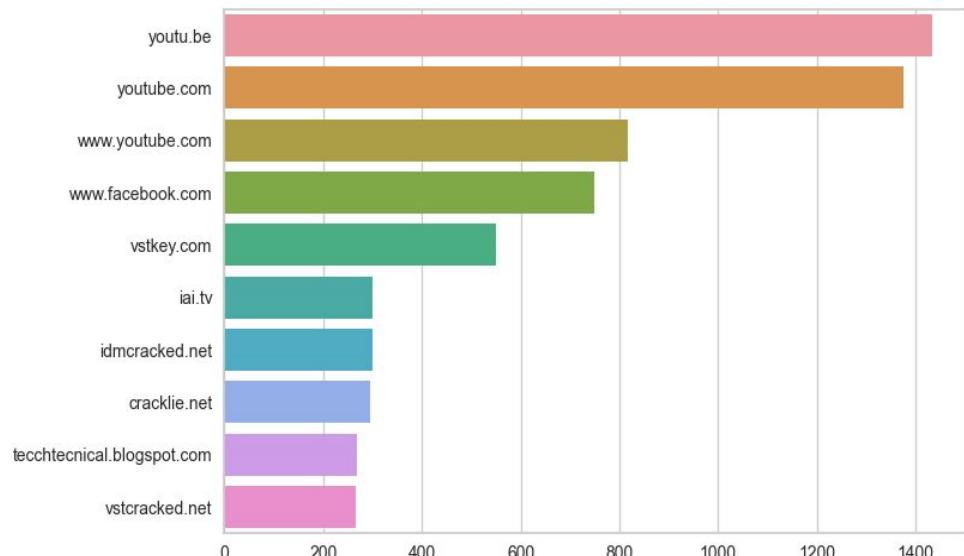


Top 10 Domain Shared

r/Science

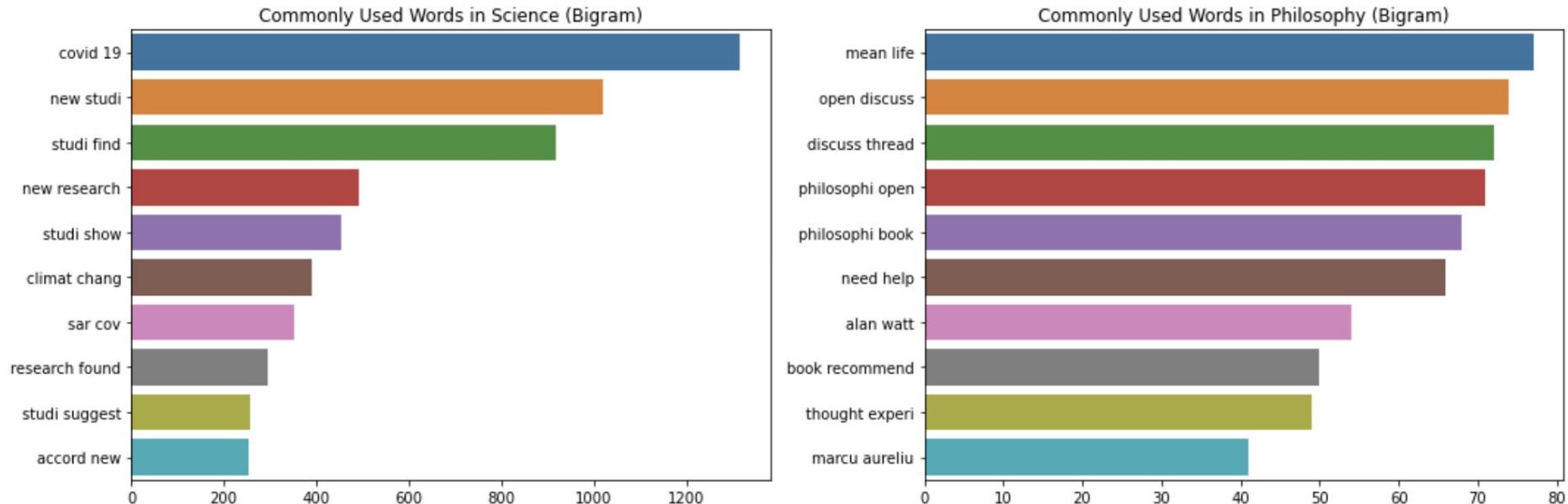


r/Philosophy



Top Frequent n-grams

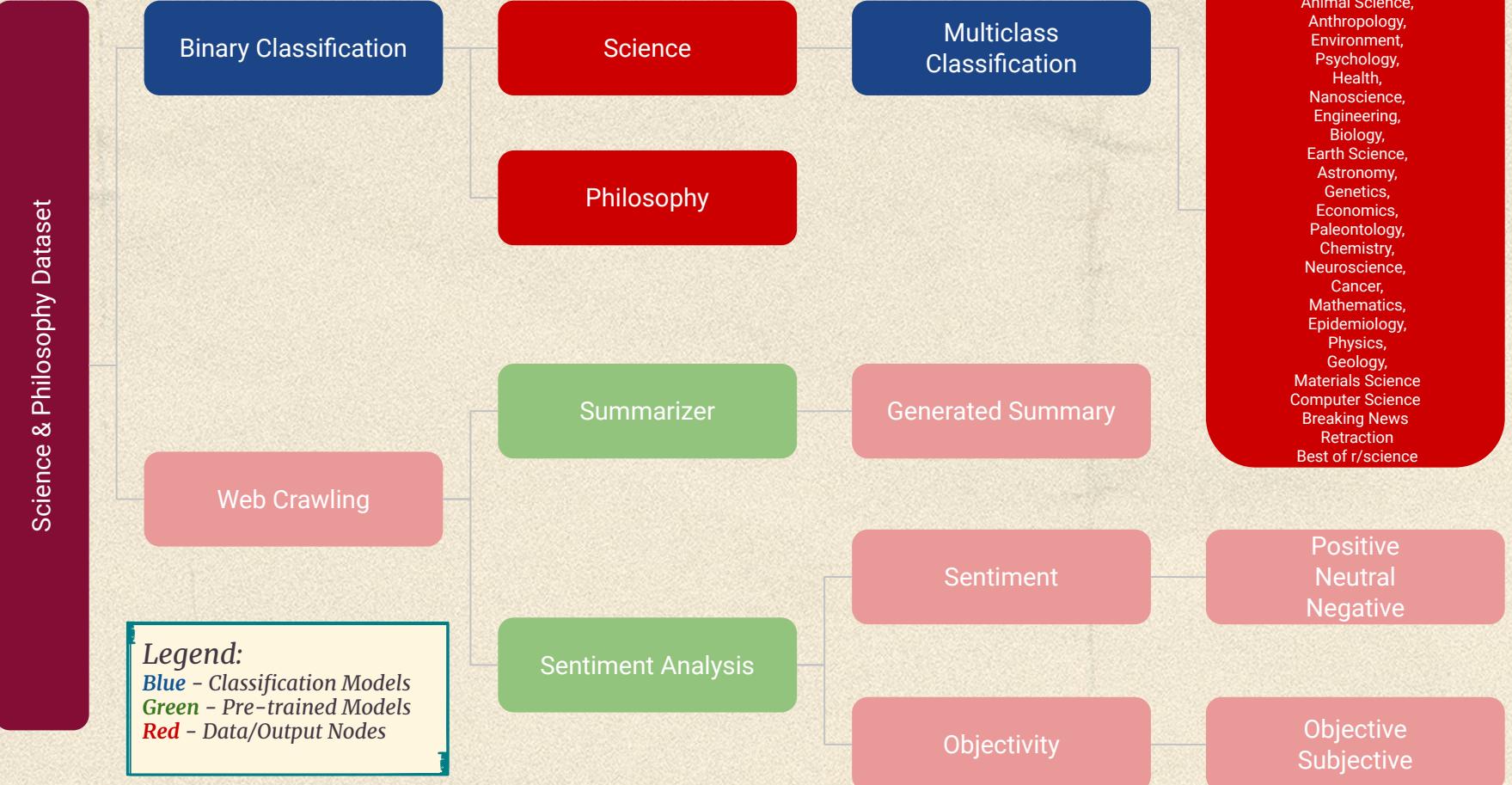
Top 10 Topics



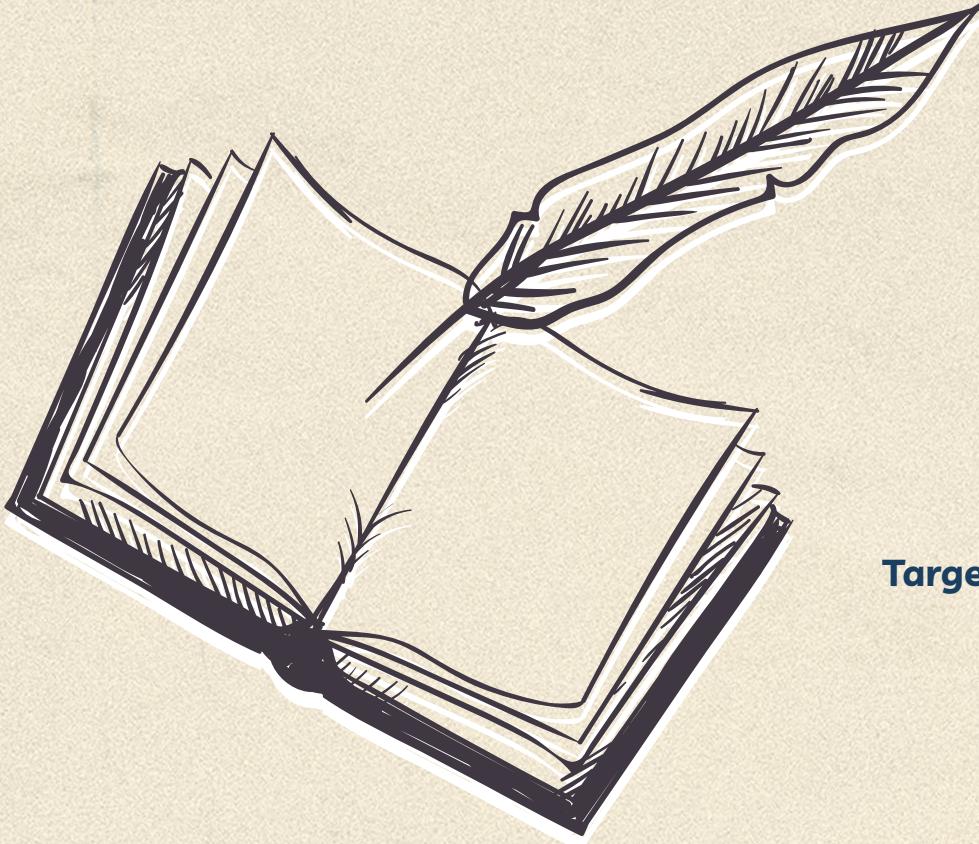
04 part 1

Modelling





Modeling Data Selection



Target

Title

Title of post

Vectors for classification

Selftext

Content of post

Url

Shared url in post

Subreddit

Science or Philosophy subreddit

Flair

Subcategory within subreddit

Binary Classification – Science v Philosophy

PyCaret Top Model Results

Model	Preprocessing	CV Accuracy
MultinomialNB	Manual (all submissions)	0.9550
MultinomialNB	Manual (remove deleted submissions)	0.9408
MultinomialNB	CountVectorizer (all submissions)	0.9653 *
MultinomialNB	CountVectorizer (remove deleted submissions)	0.9441
MultinomialNB	TfidfVectorizer (all submissions)	0.9640
MultinomialNB	TfidfVectorizer (remove deleted submissions)	0.9406
Random Forest	CountVectorizer (all submissions)	

*Top performing models
for PyCaret Binary
Classification*

```
{'cvec_binary': True,  
 'cvec_max_df': 1.0,  
 'cvec_max_features': None,  
 'cvec_min_df': 1,  
 'cvec_ngram_range': (1, 2),  
 'cvec_stop_words': 'english',  
 'cvec_token_pattern': '\\w+'}
```

Multibinomial Naive Bayes
With Count Vectorization

Relevant Metrics:

Accuracy - 0.965327

Precision - 0.979745

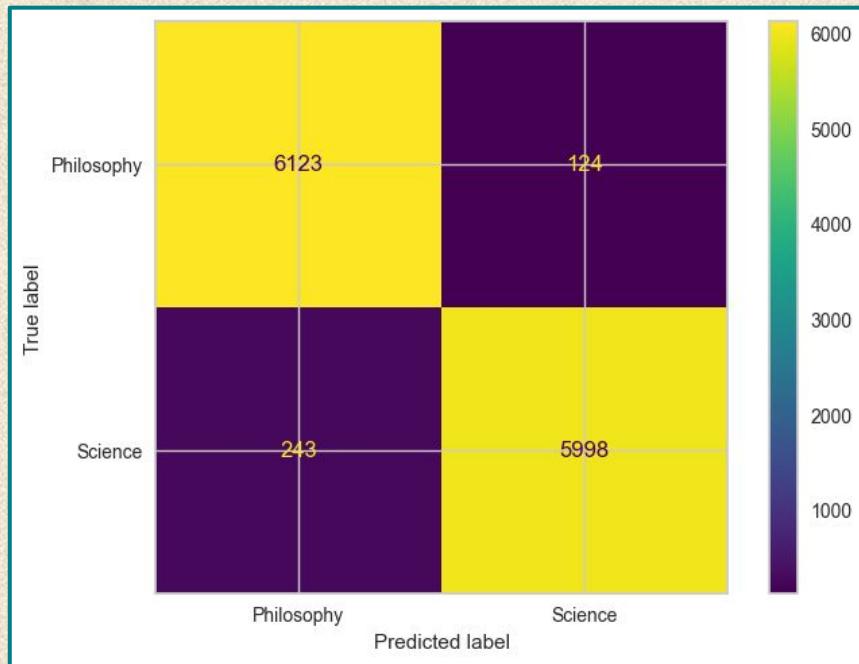
Recall - 0.961064

F1 Score - 0.970315



Binary Classification – Science v Philosophy

PyCaret Top Model Confusion Matrix



Multibinomial Naive Bayes

With Count Vectorization

Relevant Metrics:

Accuracy - 0.965327

Precision - 0.979745

Recall - 0.961064

F1 Score - 0.970315



Multiclass Classification – Science

PyCaret Top Model Results

Ridge Classification Model
With TF-IDF Vectorization



Model	Accuracy	Recall	Precision	F1
Tfidf PyCaret	0.5252*	0.3795	0.5151	0.5089*
Countvec PyCaret	0.4858	0.3476	0.4841	0.4770
facebook/bart-large-mnli	0.2400	0.2400	0.2650	0.2444
valhalla/distilbart-mnli-12-1	0.3000	0.3000	0.5429	0.3487
valhalla/distilbart-mnli-12-3	0.3400	0.3400	0.4547	0.3585
typeform/distilbert-base-uncased-mnli	0.3200	0.3200	0.4313	0.3019
Narsil/deberta-large-mnli-zero-cls	0.4400	0.4400	0.5163	0.4599

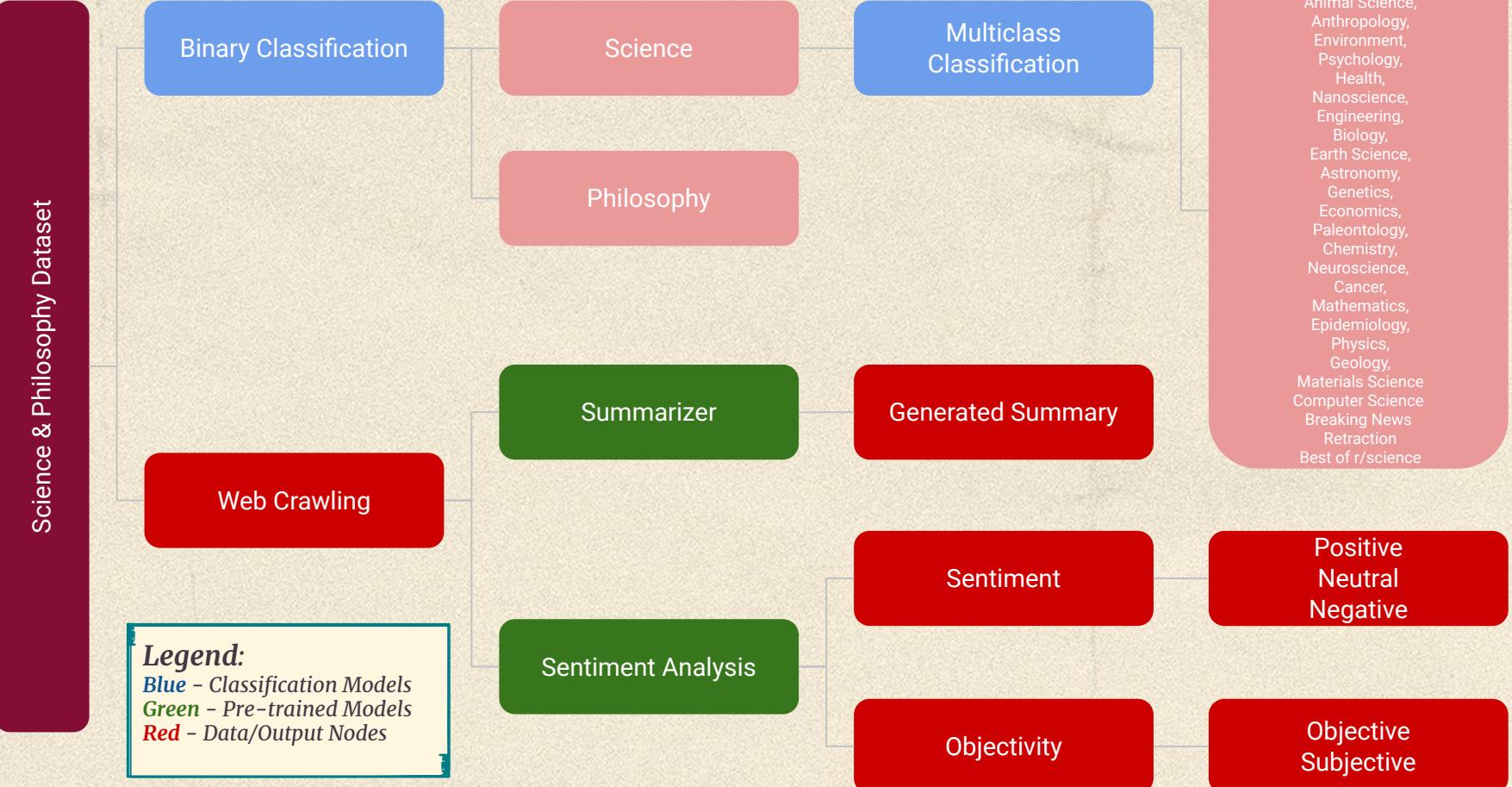
Top performing models for PyCaret Multiclass Classification

RidgeClassifier Confusion Matrix

04 part 2

Web Crawling





99% post only have content + link

Posted by u/Meatrition 2 hours ago

Health Time-restricted eating with or without low-carbohydrate diet reduces visceral fat and improves metabolic syndrome: A randomized trial
cell.com/cell-r...

10 Comments

675 Posted by u/Additional-Two-7312 9 hours ago

Astronomy Our moon has been slowly drifting away from Earth over the past 2.5 billion years, research finds
pnas.org/doi/fu...

140 Comments

14 hours ago

Article [PDF] Mackie and the Meaning of Moral Terms
jhaponline.org/jhap/a...

1 Comment

2 days ago

Blog Informed Consent and the Joe Rogan Experience
prindleinstitute.org/2022/0...

1 Comment



Summarizer

Hugging Face
sshleifer/distilbart-cnn-12-6

Reddit Post

Reddit post by u/Additional-Two-7312 9 hours ago. Upvotes: 95, Downvotes: 0. Title: Insects today are causing unprecedented levels of damage to plants, even as insect numbers decline, according to new research. Source: eurekalert.org/news-r... (link). Subreddit: Environment. Post options: 10 Comments, Award, Share, Save, ...

95 Insects today are causing unprecedented levels of damage to plants, even as insect numbers decline, according to new research

eurekalert.org/news-r... (link)

Environment

10 Comments Award Share Save ...

Article Link

<https://www.eurekalert.org/news-releases/967297>

Summary

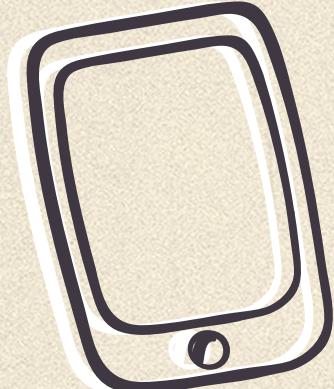
First-of-its-kind study compares insect damage of modern-era plants with that of fossilized leaves from as far back as the Late Cretaceous period, nearly 67 million years ago. Scientists say a warming climate, urbanization and introduction of invasive species likely have had a major impact.



Sentiment Analysis

spacytextblob

Reddit Post



Posted by u/Additional-Two-7312 9 hours ago

95 Insects today are causing unprecedented levels of damage to plants, even as insect numbers decline, according to new research

eurekalert.org/news-r... ↗

Environment

10 Comments Award Share Save ...



<https://www.eurekalert.org/news-releases/967297>

Article Link

Outputs



A black and white line drawing of an open book lying flat. A single feather quill pen is positioned diagonally across the pages, with its tip pointing towards the top right. The background is a light beige color.

05

Summary

Summary



Statistics

r/Science

28.4 million members

2000 posts & comments / day

100 words / post

r/Philosophy

16.9 million members

200 posts & comments / day

50 words / post

Substantial spam posts



Goals

Guide our users to:

1. Identify which Subreddit to posts at
2. Write a neutral & objective posts
3. Write a summary for the external links



Modelling

Binary Classification

Multinomial Naive Bayes
Accuracy: 0.965

Multiclass Classification

Ridge Classifier
Accuracy: 0.525

Sentiment Analysis

spacytextblob

Summarizer

HuggingFace





The Solution: An App for That

Figure out
where and
what to put
in a post

M	T	W	T	F	S	S
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

This South part
of the world, containing
almost the third part of
the Globe, is yet unknowne certa-
nly sea-coasts excepted which
rather shewe, there is a land,
then diverse either Land
people, or Commodities.



Post Title	Insects today are causing unprecedented levels of damage to plants, even as insect numbers declin
Post Content	
URL	https://www.eurekalert.org/news-releases/967297
Inspect	





Post Title

You are not obliged to vote, and maybe shouldn't. An argument from analogy.

Post Content

(This section is currently empty.)

//

URL

https://wonderandaporia.substack.com/p/should-you-vote-an-argument-from?r=1l11lq&utm_campaign=Substack%20Email%20Link&utm_medium=Email

Inspect



06

Next Steps



Philosophy - Multiclass Classification

Epistemology	Knowledge & Truth
Metaphysics	Reality & Being
Logic	Argumentation & Reason
Axiology	Aesthetics & Ethics
Political	State & Government



Additional Features

Word Correction

Suggestion to user to potentially improve comments & upvotes

Popularity Prediction

Using regression analysis to predict the number of comments & upvotes of the posts

Auto Post

From streamlit, auto-post to reddit after user is satisfied with the results

Spam Detection

Detect non-related posts



Expand to other fact based Subreddits

r/Economics

r/Astronomy

r/History

r/Geography

r/Health





THANK YOU

Try out our beta!



<https://tinyurl.com/subreddit-rel>

