Science or Science Fiction?

Bibor Szabo
DSI Student
General Assembly

Problem Statement

1. Can we build a predictive model to classify a reddit submission into science or science fiction categories with a higher than 70% accuracy?

2. What are the parameters of the most accurate model?

r/ askscience

Is it a myth or a fact that dogs can "sniff" cancer?

Biology

Ive heard of it a long time ago, that dogs are able to detect/sniff/smell cancer but never knew whether that is true or if so where it originated from. Does anybody know? Im personally no expert with animals and biology but I doubt that dogs have the ability to do that.

r/ scifi

is having a "one trillion planet empire" realistic?

when I was researching to make my universe feel huge and almost realistic size. and most sci-fi universes felt off, with all the light speed and fast radio comms, it just didn't make the universe feel huge at all, rather like it was all in one system.

Most Frequent Words

r/ askscience

Way	Water
‘Ve	Understand
Time	Space
Say	Question
Possible	Make

r/ scifi

‘Ve	Time
Think	Story
Star	Space
Series	Sci
Really	Read

Data

r/ askscience

Random sample

~ 24_000

r/ scifi

All

~ 24_000

Data Cleaning

Missing Data:

[deleted]
[removed]
NaN

Cleaning Text:

Url
Emojis
Non-word characters (/, \, (), [], etc.)

Models

Transformers

- Word Net Lemmatizer
- Porter Stemmer

- Count-Vectorizer
- Tfidf-Vectorizer

Estimators

- Logistic Regression

- Bernoulli Naive Bayes

- Decision Tree Classifier
- Ada Boost Classifier
- Gradient Boosting Classifier

Best Model

Tfidf-Vectorizer - Logistic Regression

Max_features= 10_000 words

Stop words = Not removed

Unigrams

Regularized Logistic Regression (C = 1.85 equivalent to α = 0.54)

How it Works?

- Creates a list of:
  - the most frequent subreddit_specific
  - 10_000 words
- Compares:
  - Words in post ----------> Words in list
- Calculates:
  - Probability ----------> Post in subreddit
- Classifies:
  - Based on greater probability

Evaluation of Best Model

Show me a post and I tell you

if it is

science or science fiction.

Conclusion

Accuracy

50% ⇨ **94.5%**

References

https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1

https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn/

https://towardsdatascience.com/getting-your-text-data-ready-for-your-natural-language-processing-journey-744d52912867

https://towardsdatascience.com/a-complete-exploratory-data-analysis-and-visualization-for-text-data-29fb1b96fb6a

https://www.nltk.org/book/ch01.html