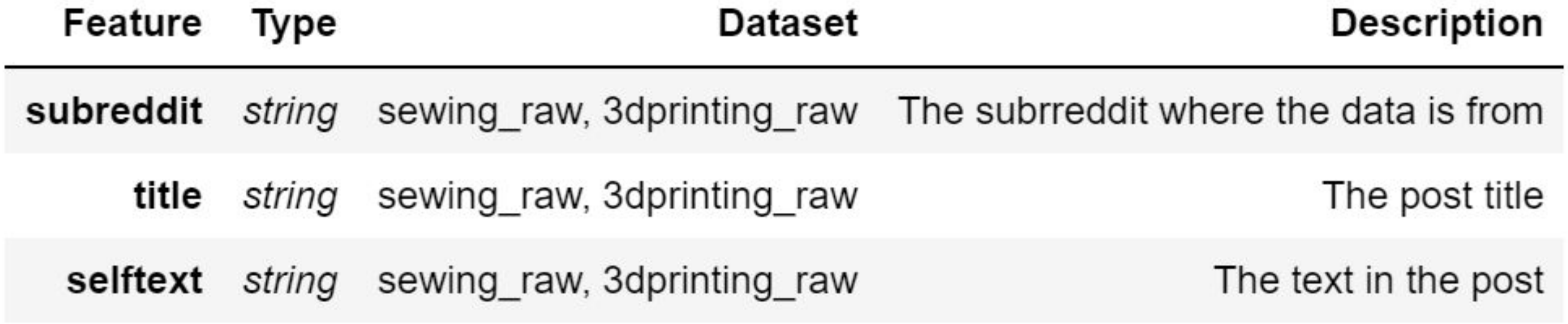
Project 3:
Reddit Post Classifier

Purpose

The purpose of this project is to create a classifier to judge which subreddit a given post is from, the sewing subreddit or the 3Dprinting subreddit.

Data Collecting

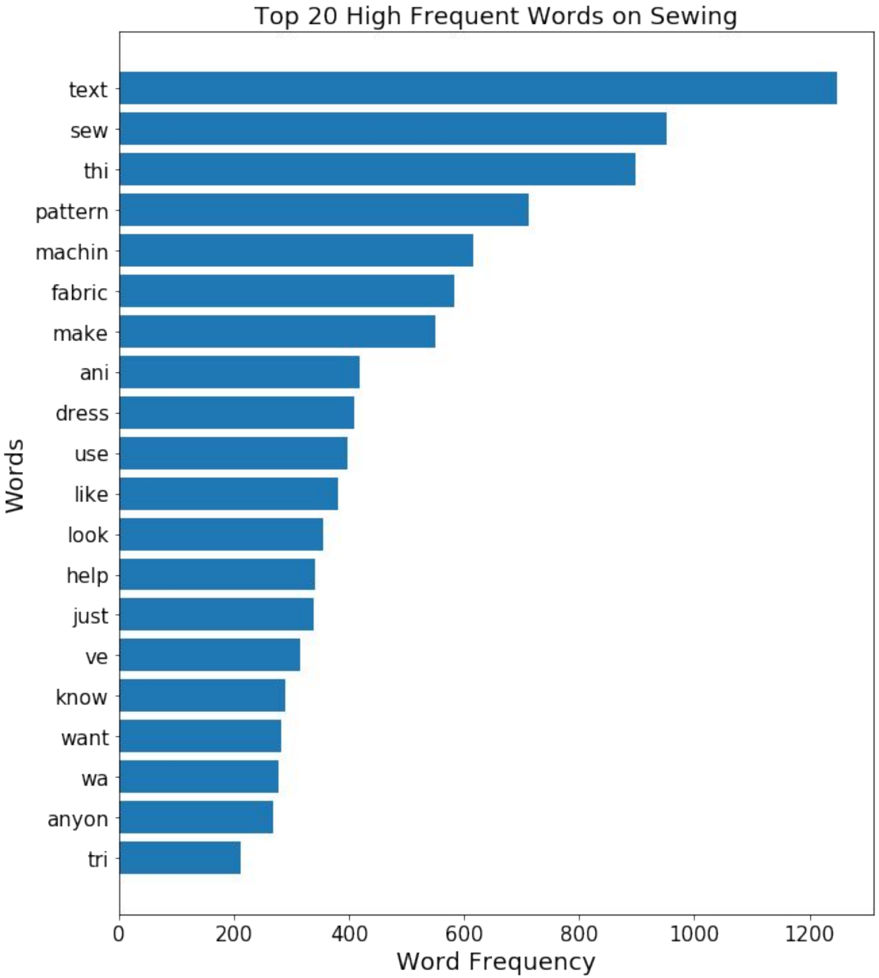
- Data were collected using pushshift's API
- 2000 data were collected from each subreddit, the sewing and the 3Dprinting



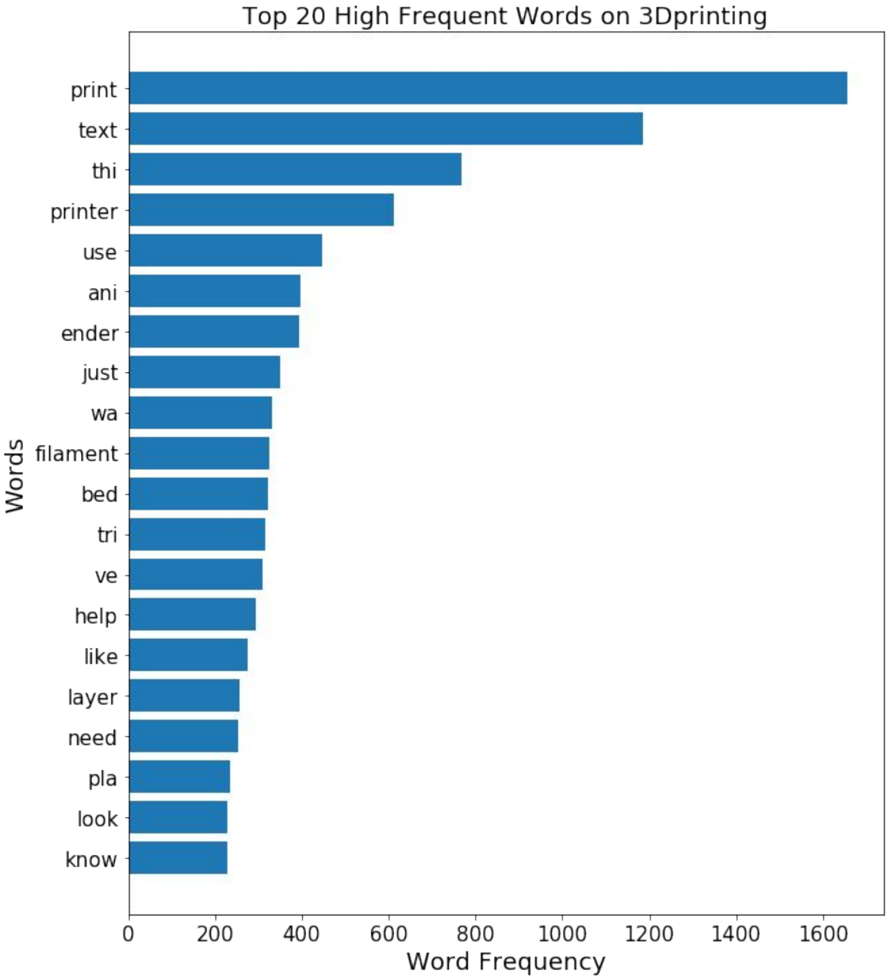
Feature	Type	Dataset	Description
subreddit	*string*	sewing_raw, 3dprinting_raw	The subrreddit where the data is from
title	*string*	sewing_raw, 3dprinting_raw	The post title
selftext	*string*	sewing_raw, 3dprinting_raw	The text in the post

Data Cleaning

- Drop the duplications
- Fill nan with a unique meaningless text ‘None text’
- Remove HTML tags (using BeautifulSoup)
- Remove URL
- Remove emojis
- Remove not words
- Stemming (using PorterStemmer)
- Combine the selftext and the title



Top 20 High Frequent Words on Sewing



Top 20 High Frequent Words on 3Dprinting

Modeling

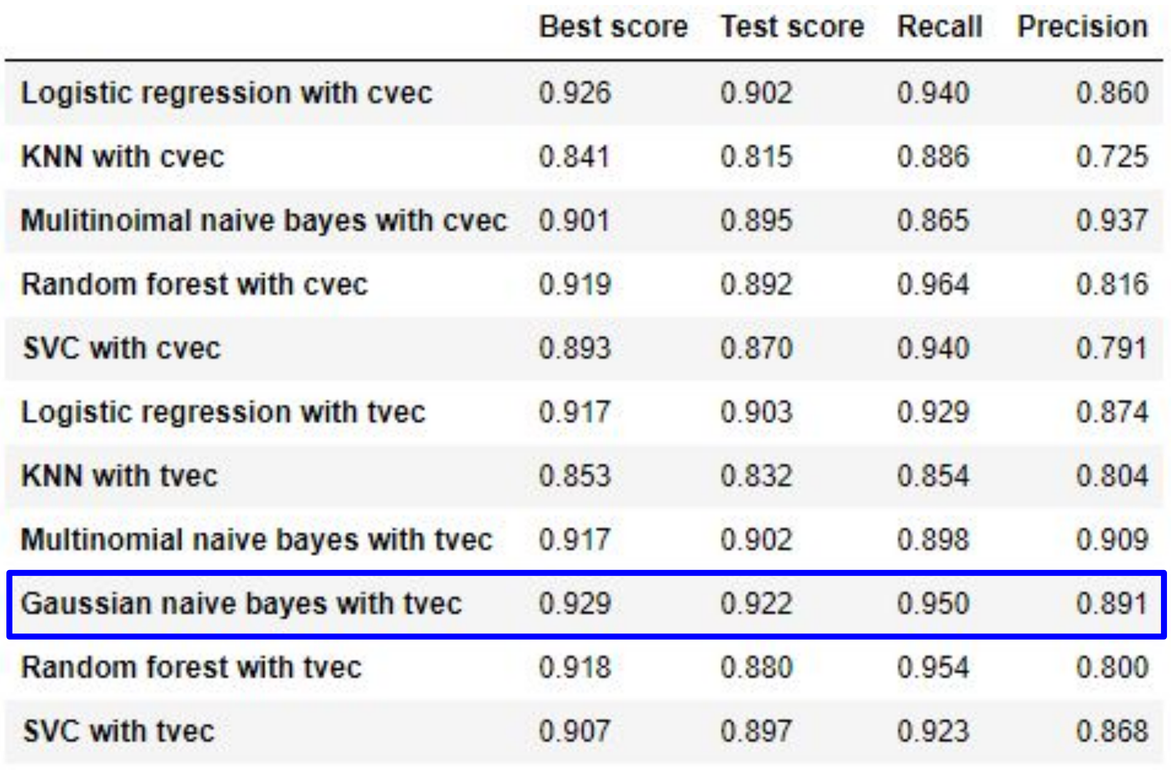
- Modeling using CountVectorizer

  ⇒ Logistic regression, knn, multinomial naive bayes, random forest

- Modeling using TfidfVectorizer

  ⇒ Logistic regression, knn, multinomial/Gaussian naive bayes, random forest

Model Evaluation



	Best score	Test score	Recall	Precision
Logistic regression with cvec	0.926	0.902	0.940	0.860
KNN with cvec	0.841	0.815	0.886	0.725
Mulitinoimal naive bayes with cvec	0.901	0.895	0.865	0.937
Random forest with cvec	0.919	0.892	0.964	0.816
SVC with cvec	0.893	0.870	0.940	0.791
Logistic regression with tvec	0.917	0.903	0.929	0.874
KNN with tvec	0.853	0.832	0.854	0.804
Multinomial naive bayes with tvec	0.917	0.902	0.898	0.909
Gaussian naive bayes with tvec	0.929	0.922	0.950	0.891
Random forest with tvec	0.918	0.880	0.954	0.800
SVC with tvec	0.907	0.897	0.923	0.868

Best model

Web Application

Web application to judge which subreddit the text is from

- The framework is Flask
- Using the best model (Gaussian naive bayes) to judge

Summary and future perspective

- Collected data from the sewing and the 3Dprinting subreddits
- The unique frequent words for the sewing are ‘sew’, ‘fabric’, etc.
- The unique frequent words for the 3Dprinting are ‘print’, ‘filament’, etc.
- The best model was Gaussian naive bayes with tvec (score 0.92)
- If the app would be for business and used for long, the model should be trained continuously(*1).
- BERT(Bidirectional Encoder Representations from Transformers) is a way of NLP, which understands a context, would improve the prediction(*2).

Reference

1*. https://towardsdatascience.com/why-machine-learning-models-degrade-in-production-d0f2108e9214

2*. https://en.wikipedia.org/wiki/BERT_(language_model)