The Tag Knows You

— A Text Prediction Model of Tags, Based on Reddit.com.

Produced by Bald Daddy

QIN Jiali, YE Qianying, GAO Ge, ZHAO Feier

All Resources Based on

01 Background

02 Project Objective

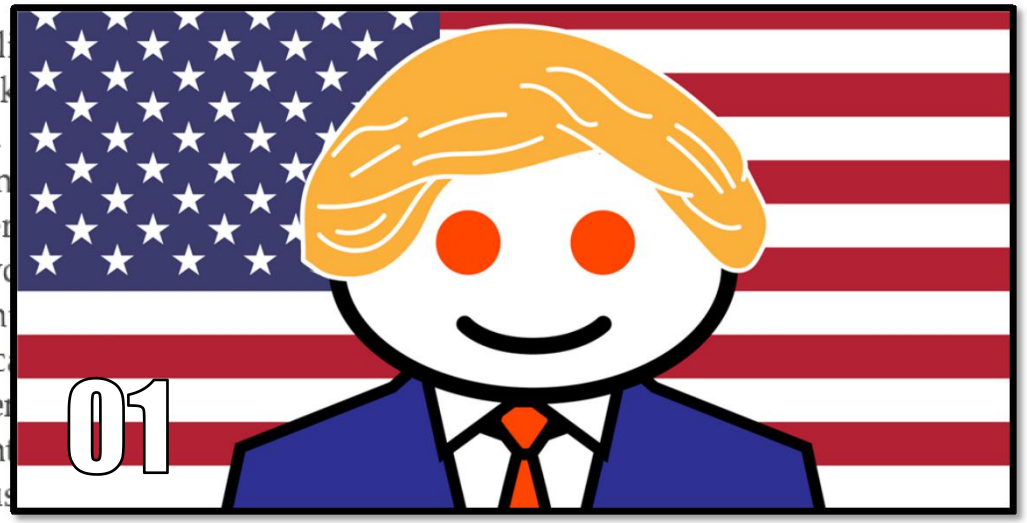
03 Data Preparation

04 Create A Model

05 Prediction

06 Conclusion

All Resources Based on



reddit reall
includ track
givin track
replacemen
usual exper
guy also w
drama righ
cleveland c
managemer
celtic bright
thunder rus
washington wizard wall best pg east 9 philadelphia 76er process work
10 portland trailblaz dame cj reach anoth level 11 toronto raptor recent
struggl concern 12 minnesota timberwolv spooki spook 13 memphi as
grizzli solid team long 14 indiana pacer pg13 15 la clipper father time a

Background

Project Objective

Data Preparation

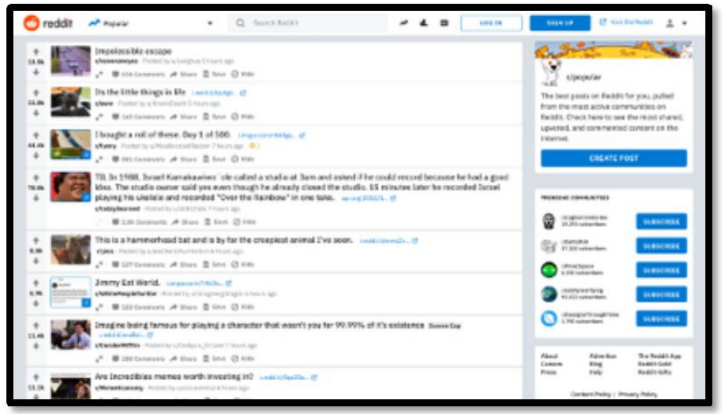
Create A Model

Prediction

Conclusion

THETAG KNOWS YOU

reddit



- An American social news aggregation, web content rating, and discussion website. Reddit had 542 million monthly visitors (234 million unique users), ranking as the **#6 most visited website** in U.S. and **#21** in the world
- Registered members submit content to the site such as **links, text posts, and images,** which are then **voted up or down** by other members.
- a variety of topics including **news, science, movies, video games, music, books, fitness, food, and image-sharing**.
- With the multireddits, users see top stories from a collection of **subreddits**.
- Reddit released its **"spoiler tags"** feature in January 2017. The feature warns users of potential spoilers in posts and pixelates preview images.

reddit realli surpr
includ track info l
givin track info se
replacement. think
usual experi artisa
guy also would m
drama right none.
cleveland cavali le
management. 5 ut
celtic bright futur
thunder russ/adar
washington wizar
10 portland trailbl
struggl concern 12
grizzli solid team

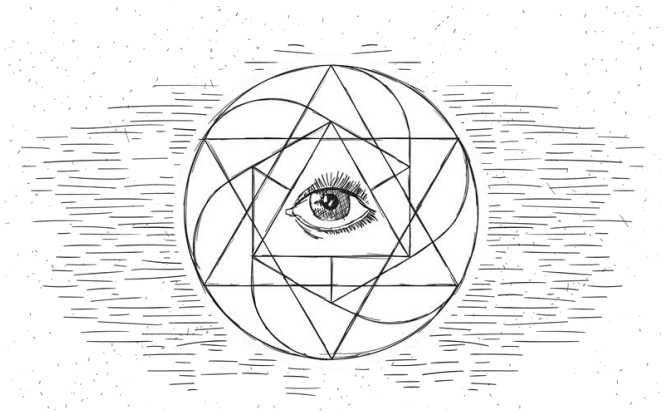
reddit

reddit realli surpr
includ track info li
givin track info se
replacement. think
usual experi artisa
guy also would m
drama right none.
cleveland cavali le
management. 5 ut
celtic bright futur
thunder russ/adar
washington wizar
10 portland trailbl
struggl concern 12
grizzli solid team

All Resources Based on

reddit reall
includ track
givin track
replacemen
usual exper
guy also wo
drama righ
cleveland c
managemer
celtic bright
thunder rus
washington wizard wall best pg east 9 philadelphia 76er process work 10 portland trailblaz dame cj reach anoth level 11 toronto raptor recent struggl concern 12 minnesota timberwolv spooki spook 13 memphi as grizzli solid team long 14 indiana pacer pg13 15 la clipper father time a

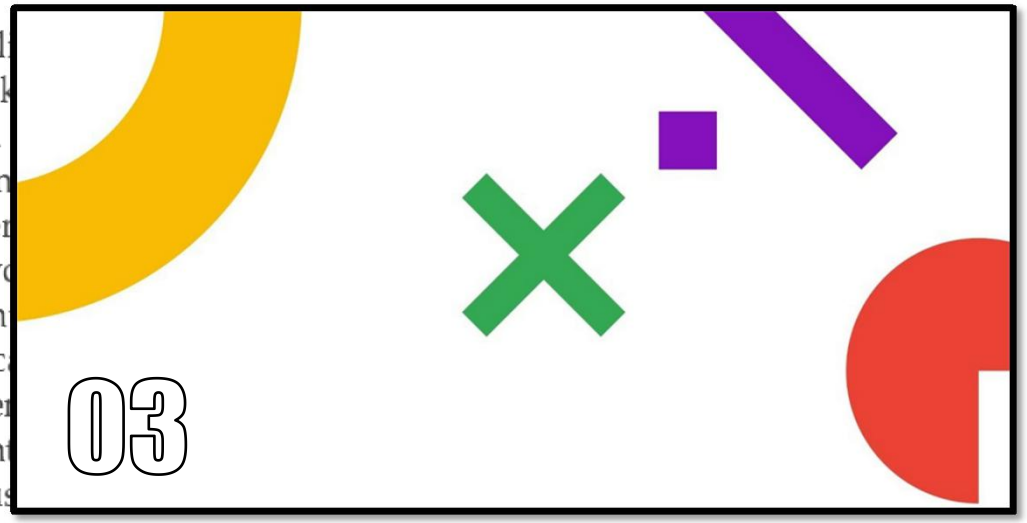
Background
Project Objective
Data Preparation
Create A Model
Prediction
Conclusion



To build a system to predict the tags on the social platform according to the text users have edited.

ycl check tube clear ovul n't latest appoint got doctor basic hand prescript provera also clomid explain take n't order blood test see ovul though suppos n't 'll period. n't
xactli 'll ovul 3 year sinc 've cycl honestli ca n't rememb cycl normal like n't know clomid make ovul around d14 like normal n't want stress get ovul test strip check sex
12 d20 like cover ovul timefram read everi day otherwis sperm immatur abl reach egg. 'm pretti piss appoint tbh differ doctor last time seem want room soon possibl nu
ppoint tell afterward expect bloat extrem pm happen took advic know tri worri import 're move week 'll probabl stress worri girlfriend parent gave set bar glass 30 year
ke vintag style narrow thick bottom thought cool gift christma parti last night unfortun idiot friend friend alway guy tri forc larg old fashion style ice cube one glass n't
mash ice cube glass broke set one glass anyon know kind glass want tri find replac sinc girlfriend realli like glass end world sinc parent 30 year got 3 month ago suck on
roken first time use highbal glass find one like http //m.imgur.com/8bsw6lv http //m.imgur.com/kaenanl first purchas cue mcdermott lucki brand cue play cue cue ' pr
ertain like feel wood wood joint cue also quit light 18oz mayb even 17oz ' also look jump break cue mcdermott call stinger ' comfort form power want get break ' us
ove it. question guy think switch ld shaft/cu ' post state ' still beginn stage think would best time switch ever right believ ' go stick medium tip fine let learn use eng

All Resources Based on



reddit reall
includ track
givin track
replacemen
usual exper
guy also w
drama righ
cleveland c
managemer
celtic bright
thunder rus
washington wizard wall best pg east 9 philadelphia 76er process work
10 portland trailblaz dame cj reach anoth level 11 toronto raptor recent
struggl concern 12 minnesota timberwolv spooki spook 13 memphi as
grizzli solid team long 14 indiana pacer pg13 15 la clipper father time a

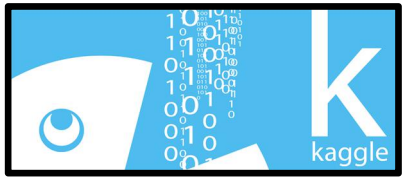
Background
Project Objective
Data Preparation
Create A Model
Prediction
Conclusion

THETAG KNOWS YOU

line. right left forward face wit
inexplic ignor pass news. rival
nti like banner realli might one
pictur contest taker'm process
mp far heater wood black sand
hwat shrimp guess 'm look tip
ani call robin care soon launch
realli red robin bird less dress
robin vector color e63946 must
d need smile one way another.
nt recent start watch watch last
ep symbiot warm know sysmt
n't get host 'm sure explain one
an much like sunburnt skin n't
g restor retain 24/7 7 month hi
point 2010 farm 2 wfe 1 search
ver upgrad psconfig command
erfectly. use differ tool roiscan
c instal working. oputils.vb tri

Original Data

Dataset from

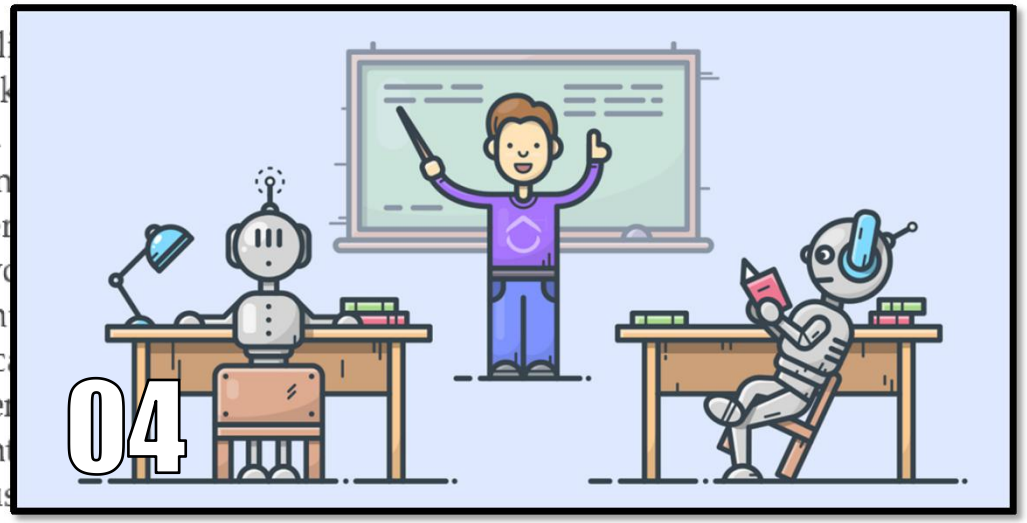


line. right left forward face wit
inexplic ignor pass news. rival
nti like banner realli might one
pictur contest taker'm process
mp far heater wood black sand
shwat shrimp guess 'm look tip
ani call robin care soon launch
realli red robin bird less dress
robin vector color e63946 must
d need smile one way another.
nt recent start watch watch last
ep symbiot warm know sysmt
n't get host 'm sure explain one
an much like sunburnt skin n't
g restor retain 24/7 7 month hi
point 2010 farm 2 wfe 1 search
ver upgrad psconfig command
perfectly. use differ tool roiscan
c instal working. oputils.vb tri

A id	A subreddit	A title	A selftext
post id	the subreddit the post was posted to	title of post (newline characters and tabs are converted to < lb \> and < tab \> , without spaces respectively)	text content of post (newline characters and tabs are converted to < lb \> and < tab \> , without spaces respectively)
1013000 unique values	1013 unique values	1002124 unique values	1013000 unique values

The dataset consists of **1.013M self-posts**, posted from **1013 subreddits** (1000 examples per class). For each post the subreddit, the title and content of the self-post are given.

All Resources Based on



reddit reall
includ track
givin track
replacemen
usual exper
guy also w
drama righ
cleveland c
managemer
celtic bright
thunder rus
washington wizard wall best pg east 9 philadelphia 76er process work 10 portland trailblaz dame cj reach anoth level 11 toronto raptor recent struggl concern 12 minnesota timberwolv spooki spook 13 memphi as grizzli solid team long 14 indiana pacer pg13 15 la clipper father time a

All Resources Based on

```
import nltk
from nltk.tokenize import word_tokenize
from nltk.stem.porter import *
from nltk.corpus import stopwords

from sklearn.naive_bayes import MultinomialNB
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.feature_extraction.text import CountVectorizer

nltk.download('punkt')
nltk.download('stopwords')
```

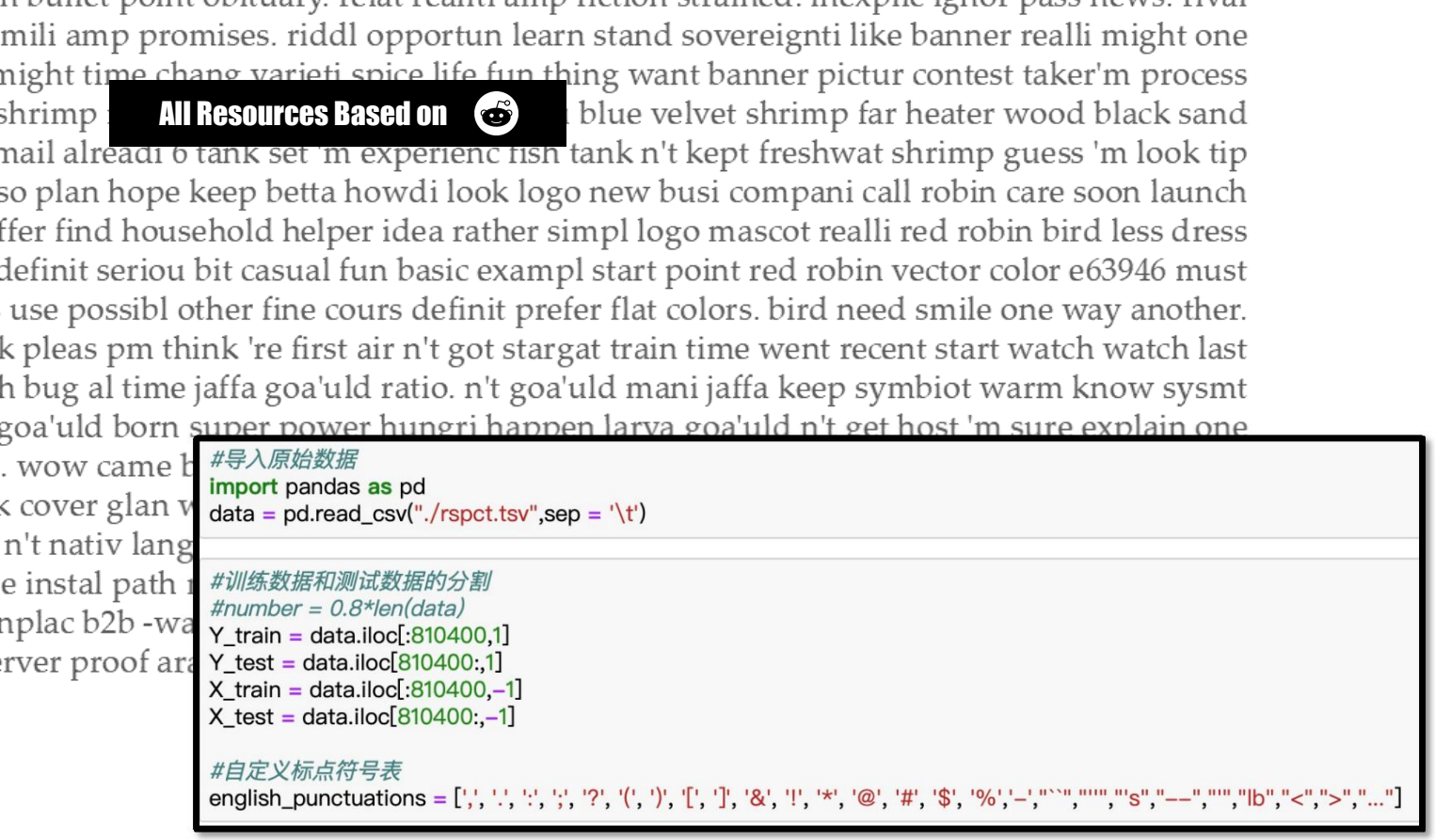
```
[nltk_data] Downloading package punkt to /Users/evondeng/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]     /Users/evondeng/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

import nltk *# Natural Language Toolkit*

Tokenize	**分词**
Stem.porter	**提取词干**
Stopwords	**停用词** *# Stopwords: I, you...*

model of naive bayes

txt feature extraction



import nltk *# Natural Language Toolkit*

Tokenize	**分词**
Stem.porter	**提取词干**
Stopwords	**停用词** *# Stopwords: I, you...*

model of naive bayes

txt feature extraction

All Resources Based on

```
#训练 文本数据的预处理和存储
for sentence in X_train:
    token=word_tokenize(sentence)
    tokens = []
    for t in token:
        t = t.lower()
        if t not in stopwords.words('english'):
            if t not in english_punctuations:
                stemmer = PorterStemmer()
                tokens.append(stemmer.stem(t))
    tokens = " ".join(tokens)
    csvFile = open("new_sentence_train.csv", "a", encoding="UTF-8")
    writer = csv.writer(csvFile)
    data = [
        tokens
    ]
    writer.writerow(data)
    csvFile.close()
```

```
# 测试 文本数据的预处理和存储
for sentence in X_test:
    token=word_tokenize(sentence)
    tokens = []
    for t in token:
        t = t.lower()
        if t not in stopwords.words('english'):
            if t not in english_punctuations:
                stemmer = PorterStemmer()
                tokens.append(stemmer.stem(t))
    tokens = " ".join(tokens)
    csvFile = open("new_sentence_test.csv", "a", encoding="UTF-8")
    writer = csv.writer(csvFile)
    data = [
        tokens
    ]
    writer.writerow(data)
    csvFile.close()
```

a. **Divide the data set by 8 : 2 - Train & Test**
b. **Filtered punctuation**
c. **Depunctuate, extract stem**
4. **Save as csv.**

All Resources Based on

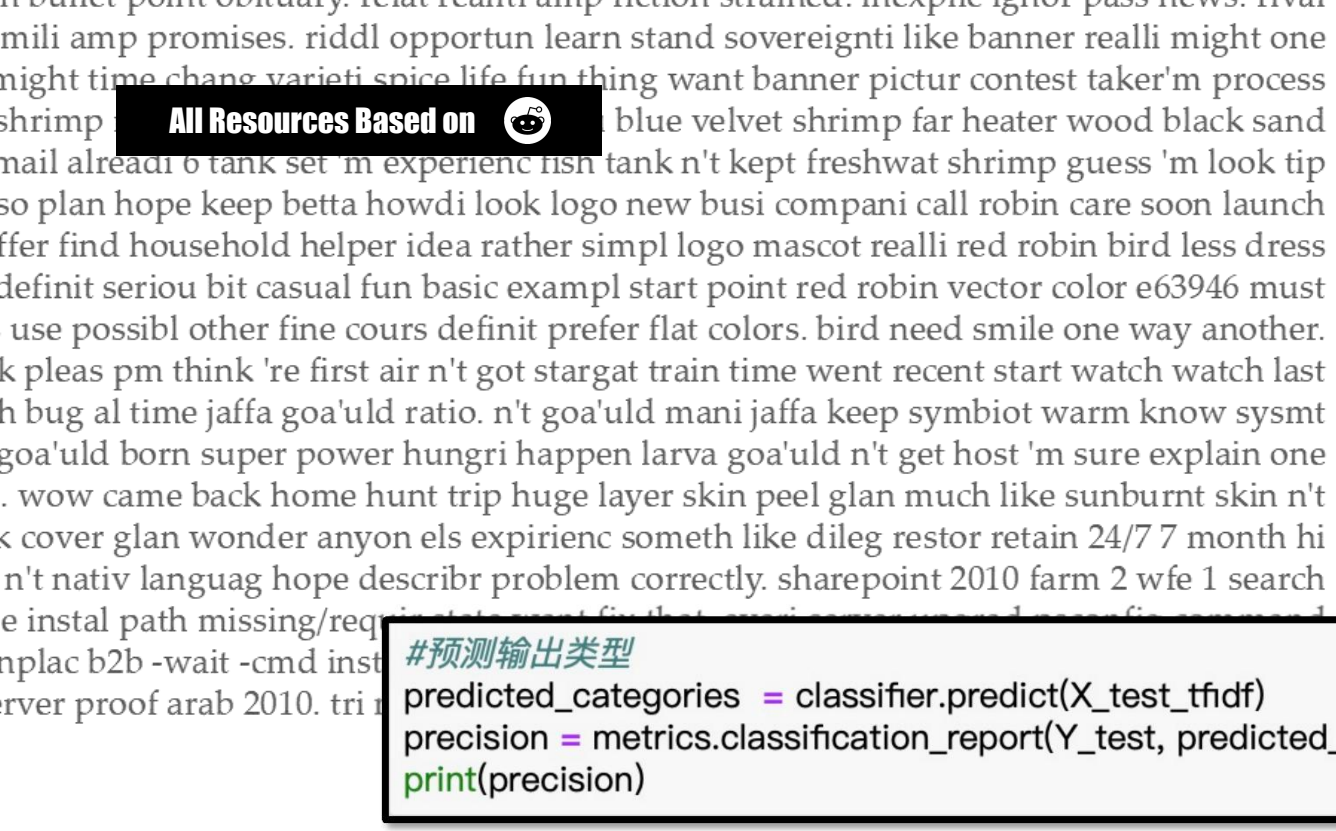
```
#从以保存的csv文件中取出数据
data1 = pd.read_csv("./new_sentence_train.csv",sep = ',')
data2 = pd.read_csv("./new_sentence_test.csv",sep = ',')
filtered_sentence_train = data1.iloc[:,0]
filtered_sentence_test = data2.iloc[:,0]
```

```
#定义文本特征提取器和tf-idf变换器
vectorizer=CountVectorizer()
transformer = TfidfTransformer()

# 用tf-idf变换器变换训练数据
X_train_tfidf = transformer.fit_transform(vectorizer.fit_transform(filtered_sentence_train))

# 多项式朴素贝叶斯分类器
classifier = MultinomialNB().fit(X_train_tfidf, Y_train)

#用tf-idf变换器变换测试数据
X_input_termcounts = vectorizer.transform(filtered_sentence_test)
X_test_tfidf = transformer.transform(X_input_termcounts)
```

tf-idf - weighting

a. Positive correlation with the frequency of occurrence in a document.

b. Negative correlation with the frequency of occurrence in the whole corpus.



to predict the output

All Resources Based on

Project Objective

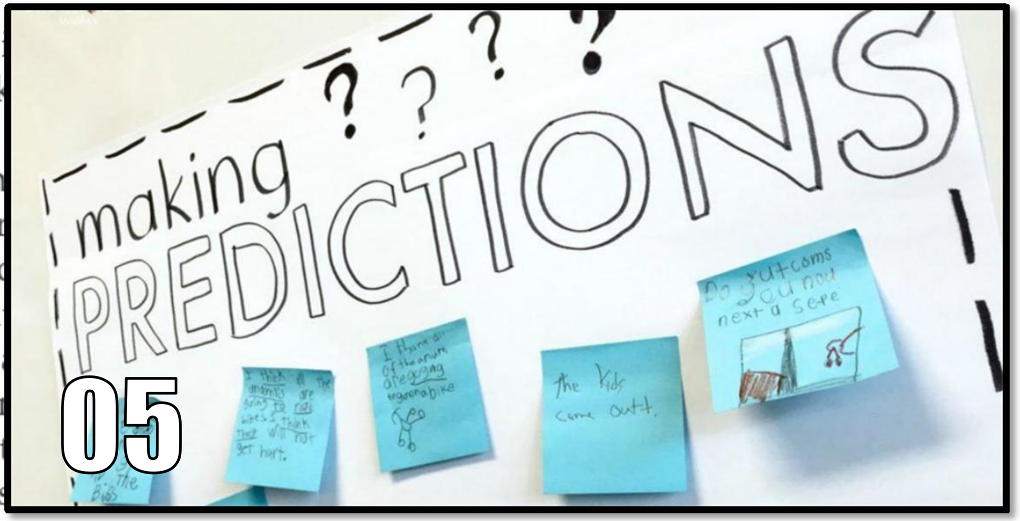
Data Preparation

Create/Test A Model

Prediction

Conclusion

reddit reall
includ track
givin track
replacemen
usual exper
guy also wo
drama righ
cleveland c
manageme
celtic brigh
thunder ru



washington wizard wall best pg east 9 philadelphia 76er process work 10 portland trailblaz dame cj reach anoth level 11 toronto raptor recent struggl concern 12 minnesota timberwolv spooki spook 13 memphi as grizzli solid team long 14 indiana pacer pg13 15 la clipper father time a

THETAG KNOWS YOU

keep improv big guy. 6 boston celt
ard wall best pg east 9 philadelphia
pook 13 memphi grizzli solid team
rade korver plan begins. 18 milwau
oung get bump potential. 22 detroi
n helm 27 chicago bull go plan 28 s

	precision	recall	f1-score	support
13ReasonsWhy	0.73	0.80	0.76	200
3Dprinting	0.72	0.60	0.66	200
3d6	0.44	0.94	0.60	200
4Runner	0.77	0.67	0.72	200
7daystodie	0.74	0.58	0.65	200
90DayFiance	0.95	0.44	0.60	200
ABDL	0.87	0.55	0.67	200
ABraThatFits	0.70	0.95	0.81	200
ACL	0.62	0.95	0.75	200
ACT	0.77	0.70	0.74	200
ADHD	0.38	0.50	0.43	200
APStudents	0.82	0.79	0.80	200
ASUS	0.72	0.46	0.56	200
AcademicPsychology	0.49	0.54	0.51	200
Accounting	0.84	0.38	0.52	200
Adobelllustrator	0.57	0.68	0.62	200
Adoption	0.57	0.84	0.68	200
workflow	0.87	0.80	0.83	200
wownoob	0.77	0.84	0.80	200
writing	0.41	0.48	0.44	200
xxfitness	0.37	0.61	0.46	200
xxketo	0.69	0.69	0.69	200
yandere_simulator	0.98	0.81	0.89	200
ynab	0.85	0.91	0.88	200
yoga	0.84	0.71	0.77	200
yorku	0.97	0.52	0.67	200
zootopia	0.91	0.68	0.78	200
micro avg	0.71	0.71	0.71	202600
macro avg	0.77	0.71	0.72	202600
weighted avg	0.77	0.71	0.72	202600

r.com/a/l4evp chase gg 55 enchantr
sold gitd harley 10 stingray 5 purp
en goblin 22 gitd black suit spiderm
st coupl day today fell complet http
nyon know need repot whatev els
user least account pretti posit guy l

All Resources Based on

Data Preparation
Create/Test A Model
Prediction
Conclusion

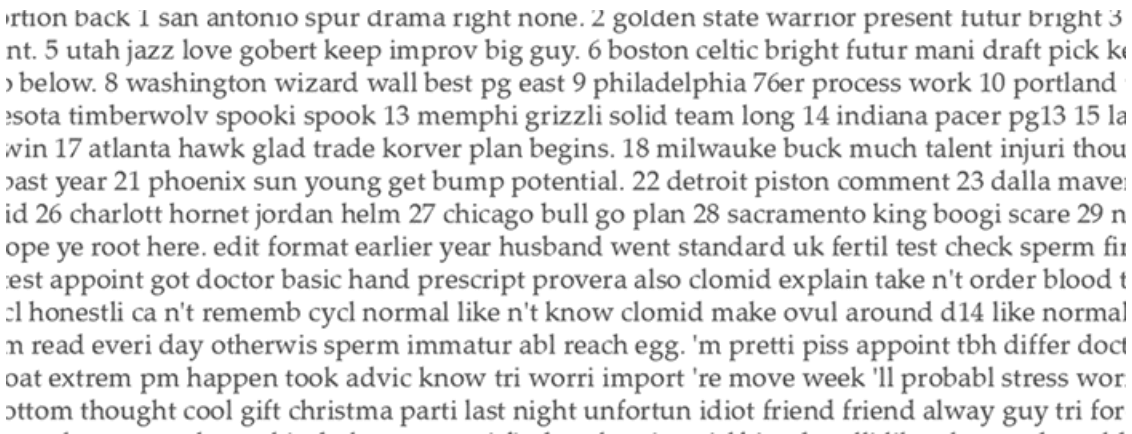
reddit reall
includ track
givin track
replacemen
usual exper
guy also w
drama righ
cleveland c
managemer
celtic bright
thunder rus
washington wizard wall best pg east 9 philadelphia 76er process work 10 portland trailblaz dame cj reach anoth level 11 toronto raptor recent struggl concern 12 minnesota timberwolv spooki spook 13 memphi as grizzli solid team long 14 indiana pacer pg13 15 la clipper father time a

Overall, both precision and recall value is very high, so we can say that the model is somehow good.

limitation

We propose an assumption that we can construct a library in which some keywords can be added to a certain area;

For example we can define “lipstick” or “fragrance” to cosmetic area. That may improve the precision of the predict system of social platform tag.




That's all Folks!