Comparison of r/WFH and r/digitalnomad using Natural Language Processing

Aridania Gerardo

Subreddit's Analyzed

r/WFH

- "Welcome to 'WFH - Working From Home,' the subreddit dedicated to those of us who work from home, be it for yourself or a company. Learn tips and tricks to make yourself more productive, avoid distractions and generally make your experience a more positive one."
- 23.6k Members
- Created Dec 8, 2010

r/digitalnomad

- "Digital Nomads are individuals that leverage technology in order to work remotely and live an independent and nomadic lifestyle."
- 1.4m Nomads
- Created Oct 15, 2009

Extract Data using Reddit's API through PRAW

```
# Instantiate Reddit using PRAW.
# API Pull Set-Up of Comments with Praw
# reddit = praw.Reddit(
#     client_id=" ",
#     client_secret=" ",
#     password=" ",
#     user_agent="Comment Extraction (by u/USERNAME)",
#     username=" ",
#                         )
```

Testing Models for the Best Classifier: Naive Bayes Classifier Model

Naive Bayes Classifier Model

```
mnb = MultinomialNB(alpha = .7)
mnb.fit(X_train_counts, y_train)

print("Train data CV score:", cross_val_score(mnb, X_train_counts, y_train, cv= 5))
print("Test data score:", mnb.score(X_test_counts, y_test))
```

```
Train data CV score: [0.82699478 0.81208054 0.8380597  0.81791045 0.82910448]
Test data score: 0.8324750075734626
```

Naive Bayes Classifier was the best model with a test data score of 0.832475 while the other three models (random forest, extra trees, and knn) were a runner up or lacked satisfaction.

Top 10 WFH Features

r/wfh
- Work
- Business
- Serious

wfh	-7.713109
office	-5.478953
job	-4.187229
home	-3.059532
work	-2.980772
desk	-2.842666
company	-2.389099
jobs	-2.283791
working home	-2.112716
day	-2.041842

Top 10 digitalnomads Features

r/digitalnomads
- ❏ Travel
- ❏ Adventures

country	3.573275
airbnb	3.425779
month	2.863306
nomad	2.721290
places	2.641610
place	2.640761
city	2.639234
visa	2.624555
dn	2.479216
countries	2.418015

Trained Logistic Regression Classifier

Test String "data science" to See which Subreddit it would be Categorized Under

Logistic Regression Classifier

```
test_post = ["data science"]
tes_counts = vectorizer.transform(test_post)
print(log_reg.predict(test_counts))
```

```
['WFH']
```

The trained Logistic Regression Classifier used the the test string “data science” and placed it as if it would originate under the subreddit r/WFH.

Summary

Using Natural Language Processing methods,I was able to analyze both subreddit's r/WFH and r/digitalnomads to train a Natural Language Processing model to identify what subreddit a test string is more likely to originate from (subreddit group). With further effort,I could observe more specific subreddits or users and look at sentiment analysis and add a view of how individual users change over time their subreddit digital fingerprint.

A good next step could be to use PRAW to gather posts from more groups such as r/WFH, r/workfromhome, and r/digitalnomands and analyze the intensity of the interaction connections between these three subreddits and visualize the promising interconnections as the connections extend to outer groups