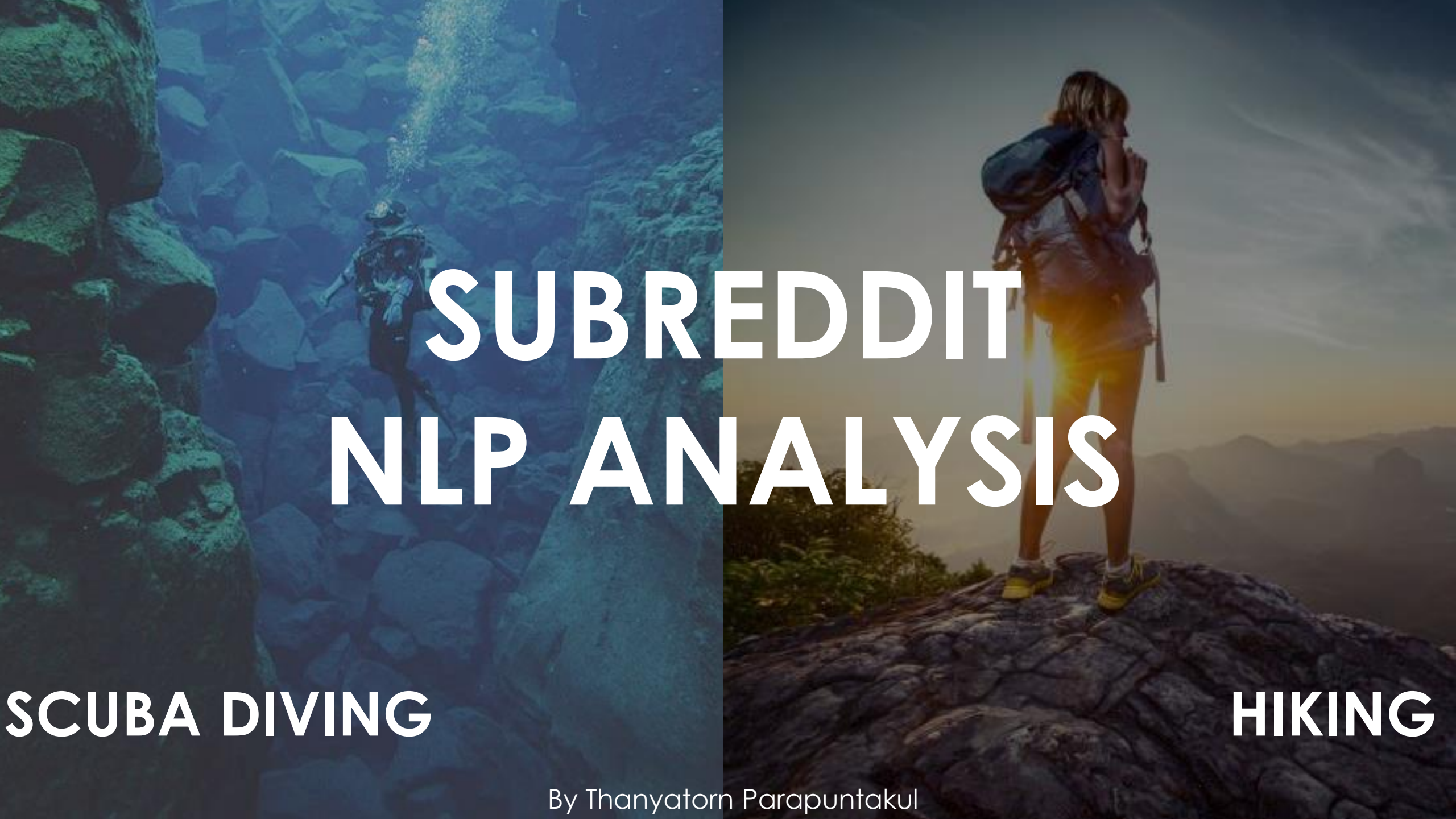


SUBREDDIT NLP ANALYSIS

SCUBA DIVING

HIKING

By Thanyatorn Parapuntakul

OVERVIEW

Reddit is a network of communities based on people's interests.

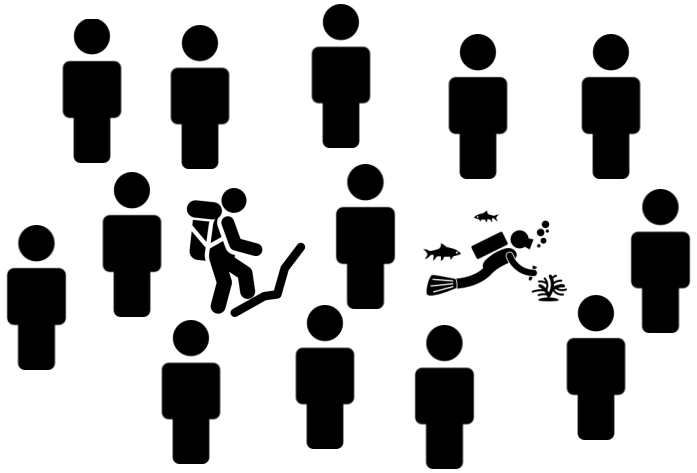
Collect 1,000 recent posts for **scuba diving** and **hiking subreddit**

The hiking's subreddit.
- 1.4 M members
- Since Oct 13, 2009

The scuba diving's subreddit.
- 95.6k members
- Since Mar 16, 2008

PROBLEM STATEMENT

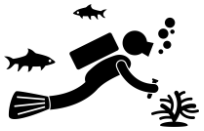
Scuba diving and **Hiking** are adventure sport that allow us to explore the world above and under water.
Therefore as a travel agency who needs to seek for adventure traveler among the people on the internet.
Therefore by knowing a dominant words that people use to discuss and search about scuba diving and hiking can helps the travel agencies spot their customers.



EDA

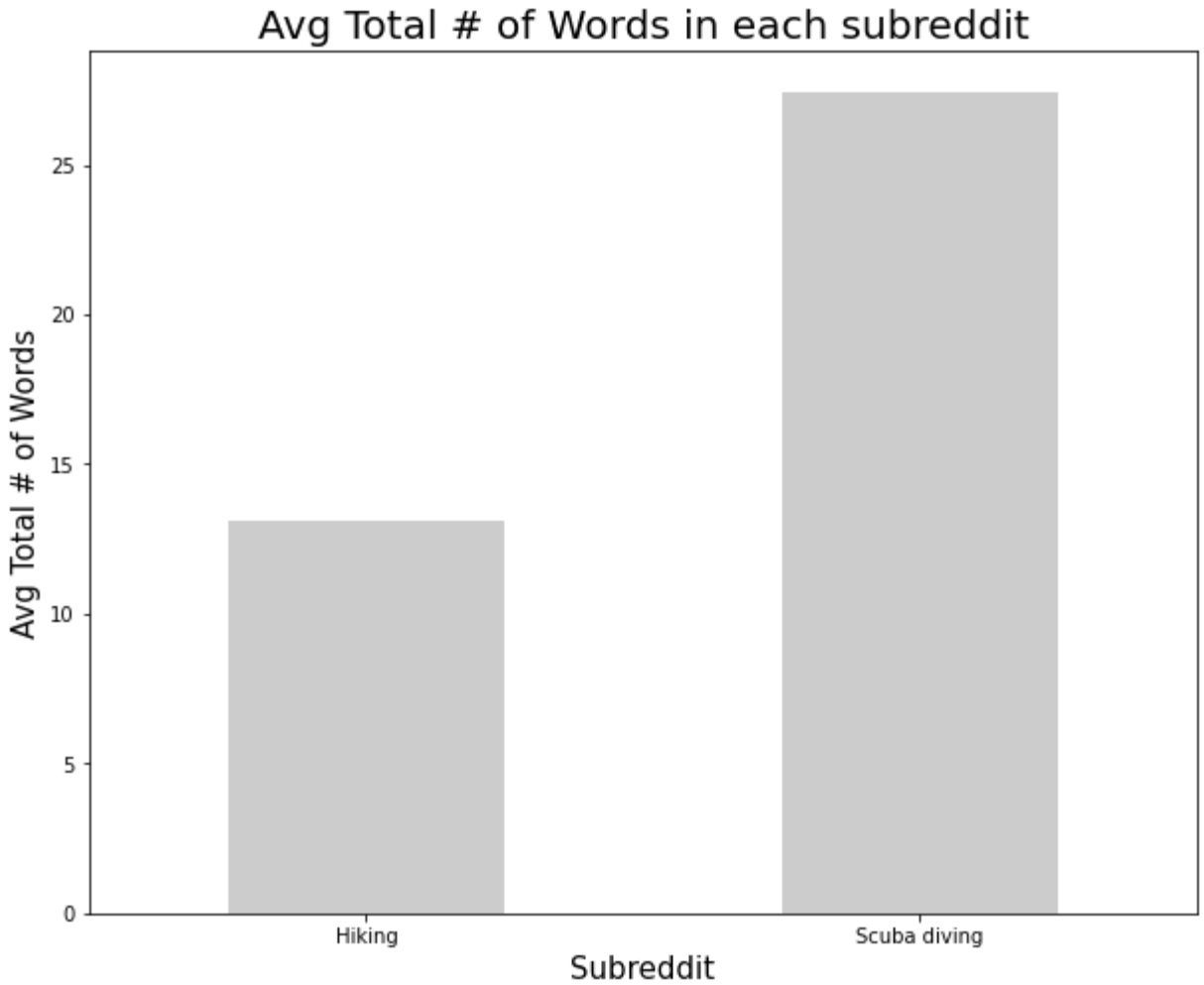
The hiking's subreddit.

- Avg.13 words per post

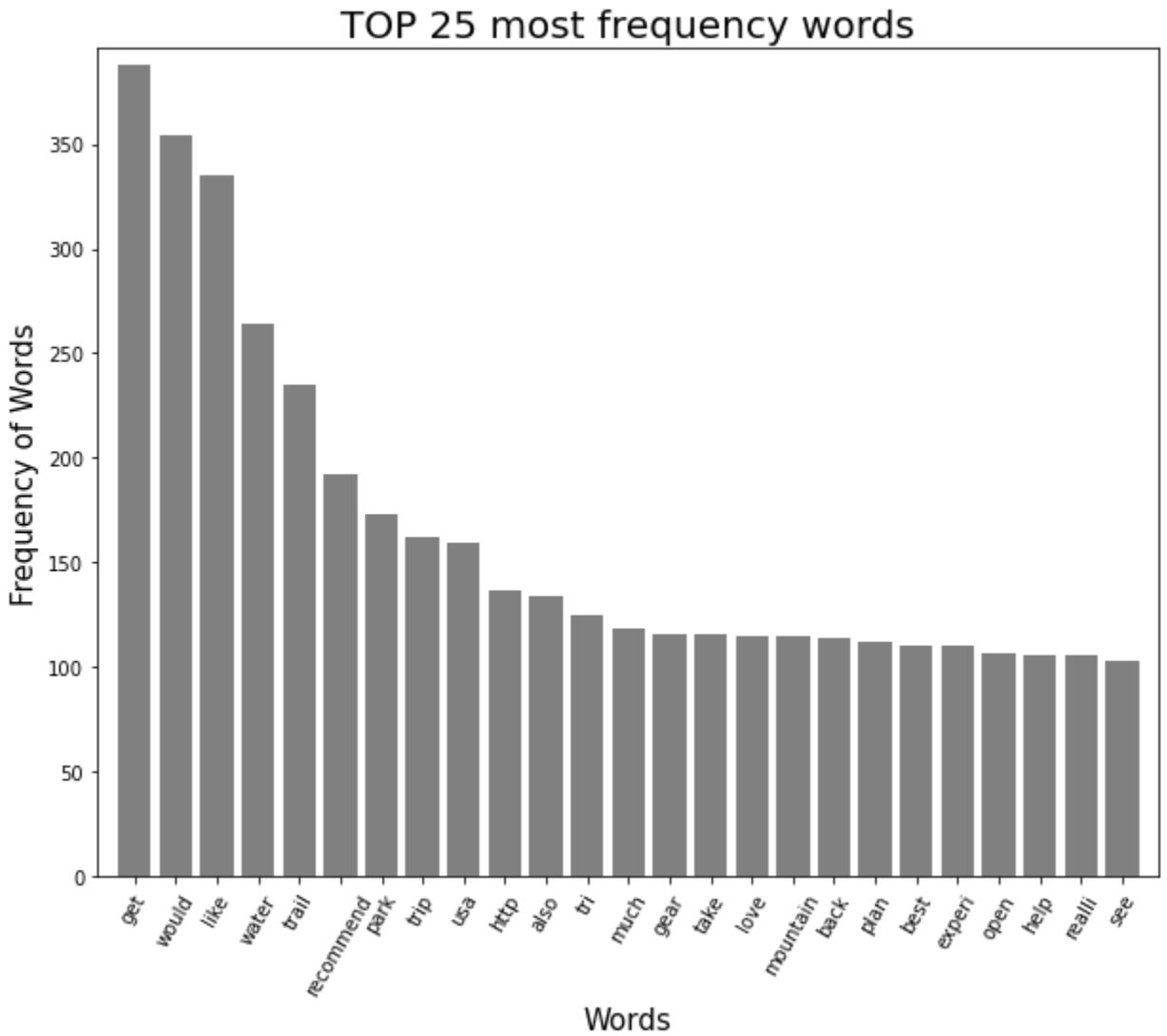


The scuba diving's subreddit.

- Avg.27 words per post



EDA

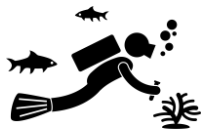


Top 3 most frequency is a general word

- Get
- Would
- Take

EDA

SCUBA DIVING



The scuba diving's subreddit.

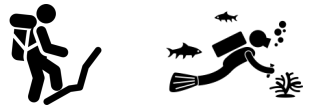
- Recommend
- Trip
- Help

HIKE

The hiking's subreddit.

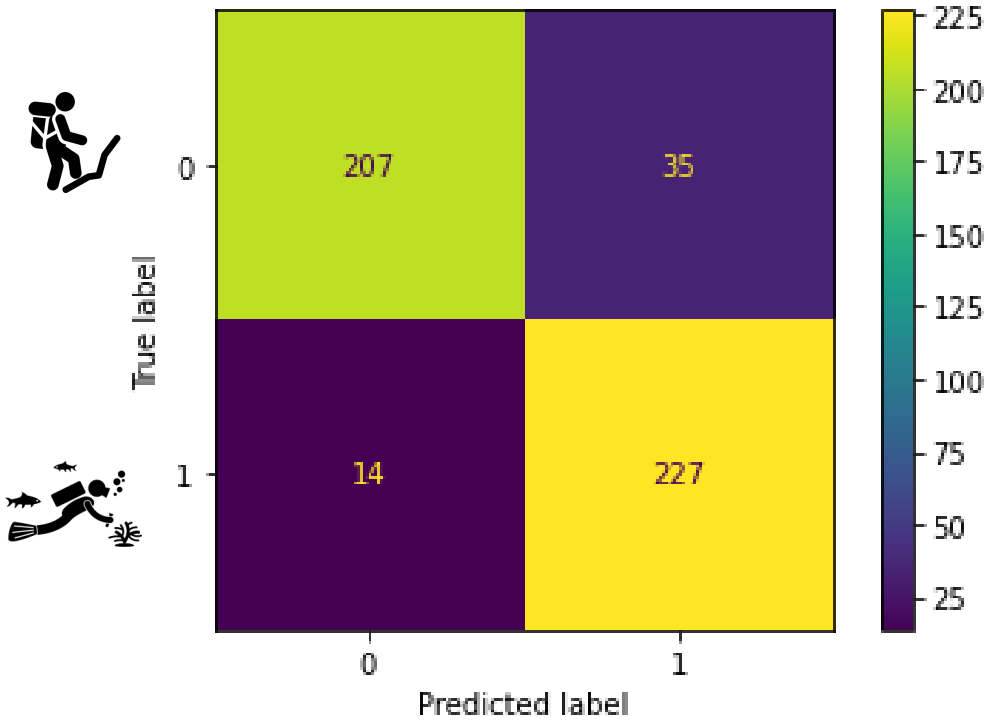
- Trail
- Mountain
- USA

MODELING



Model	Vectorizer	CV Score on Traing data set	Test Score
Logistic Regression	Count	89.35%	89.85%
Logistic Regression	Tfidf	89.62%	89.02%
Multinomial NB	Count	89.27%	89.02%
Multinomial NB	Tfidf	88.45%	88.61%
Binomial NB	Count	71.71%	71.42%
Binomial NB	Tfidf	71.71%	71.42%
KNeighbors NB	Count	72.40%	75.77%
KNeighbors NB	Tfidf	87.41%	89.23%

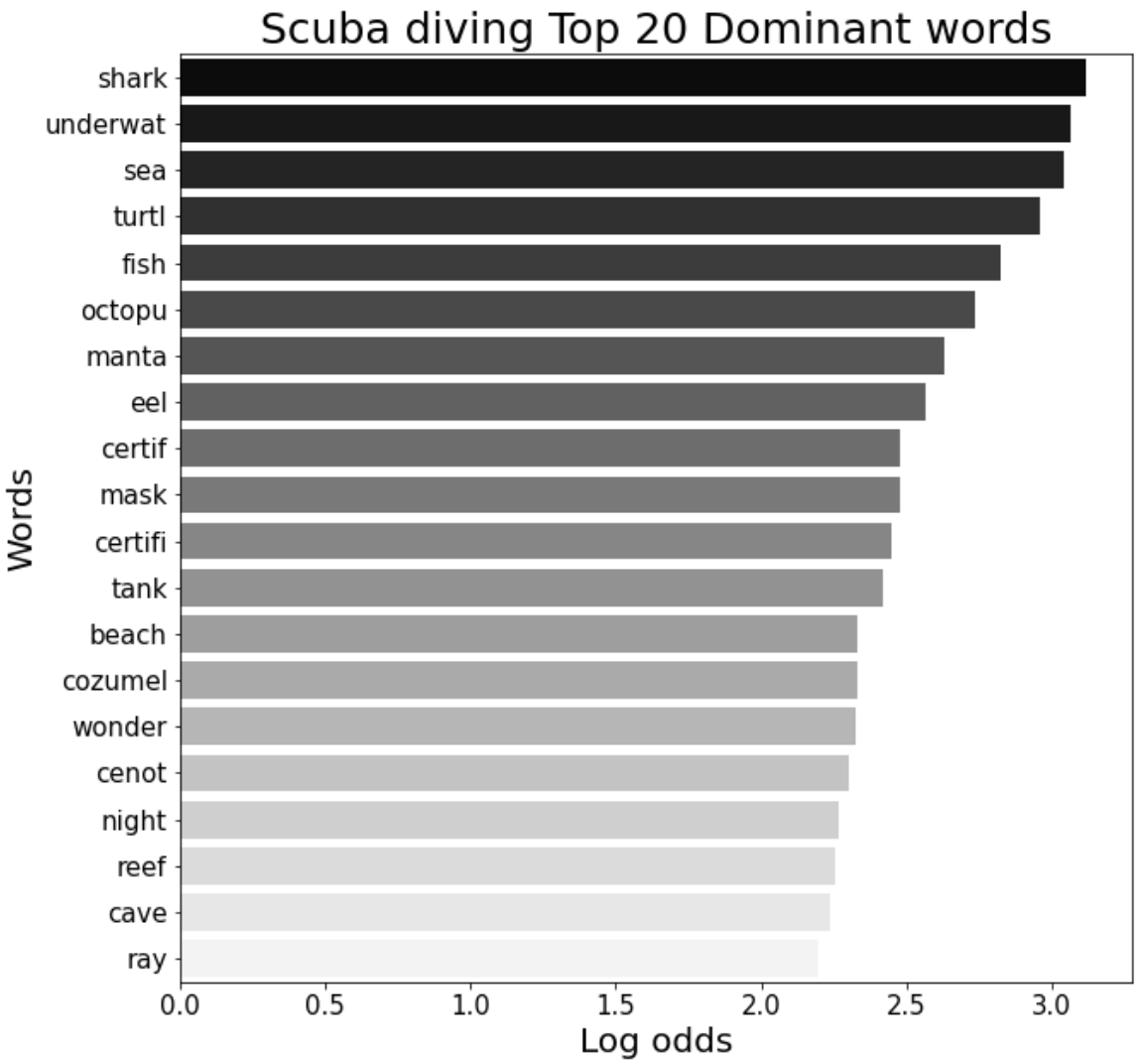
EVALUATION

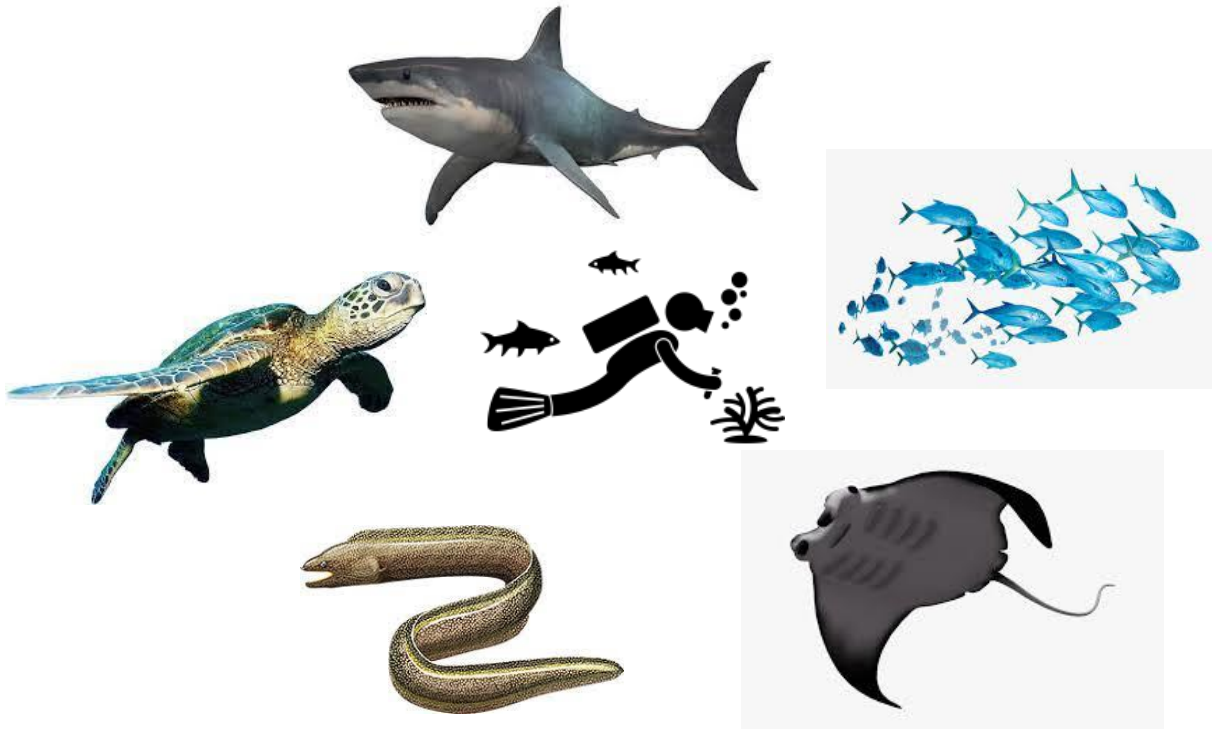


Accuracy: 0.8986
Sensitivity: 0.9419
Specificity: 0.8554
Precision: 0.8664

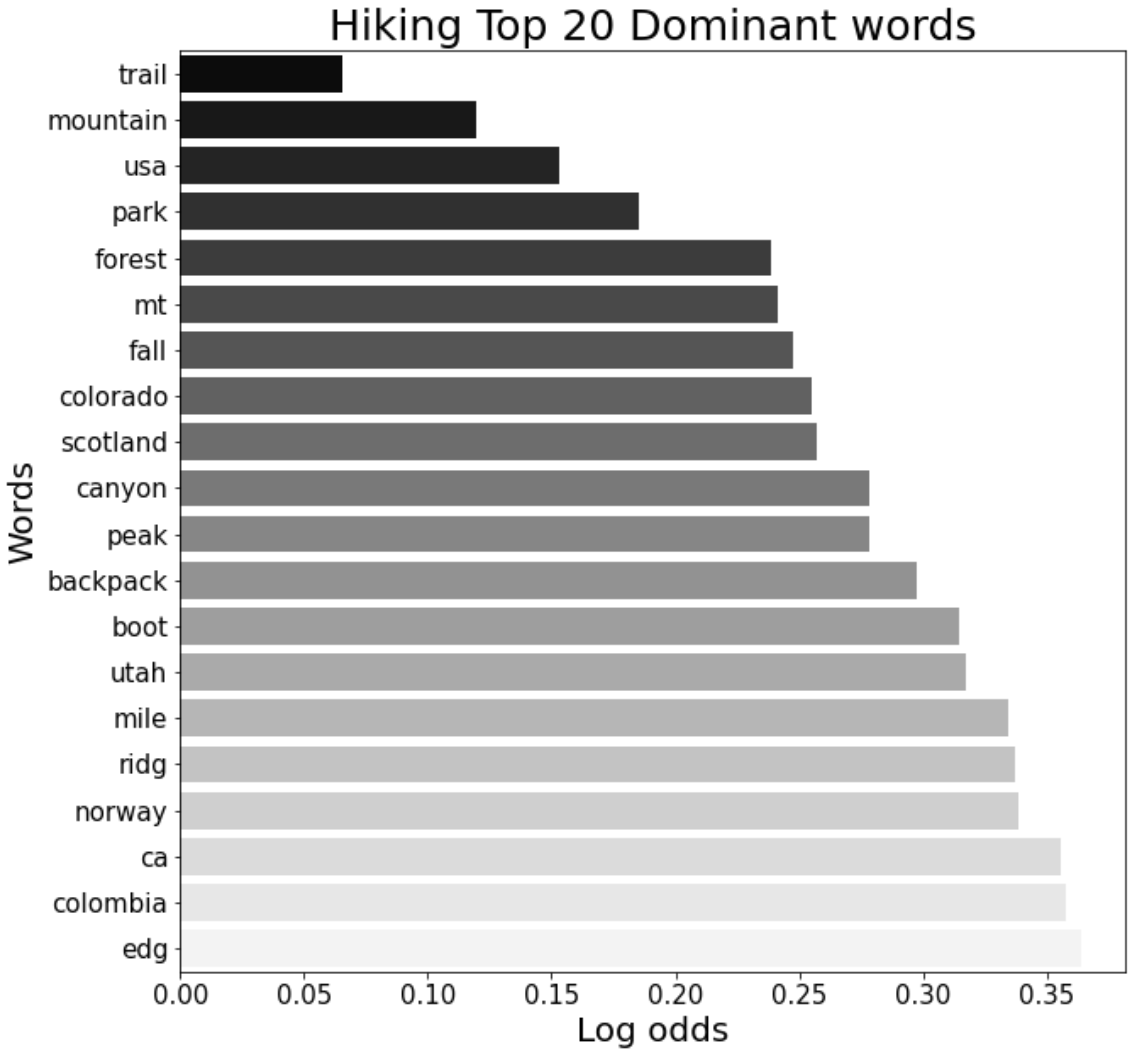
- Of all the posts, 89.86% are identify correctly.
- For Scuba diving posts 94.19% are identify correctly.
- For Hiking posts 85.54% are identify correctly.

RESULTS


Scuba diving Top 20 Dominant words
shark
underwat
sea
turtl
fish
octopu
manta
eel
certif
mask
certifi
tank
beach
cozumel
wonder
cenot
night
reef
cave
ray
Words
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Log odds



RESULTS


Hiking Top 20 Dominant words
trail
mountain
usa
park
forest
mt
fall
colorado
scotland
canyon
peak
backpack
boot
utah
mile
ridg
norway
ca
colombia
edg
Words
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Log odds

CONCLUSION

- Results from final model shows that posts in scuba diving and hiking subreddit are fairly different, and can classify with an accuracy of 89.86%

- The differences of the words are mainly due to creatures under the sea in scuba diving activity and how to describe location both activities

RECOMMENDATIONS

For further model prediction improvement

- Optimize stop words
- Increase number of posts for training data to have more words
- Include more text such as comments in each posts
- Use an image as a features


THANK YOU