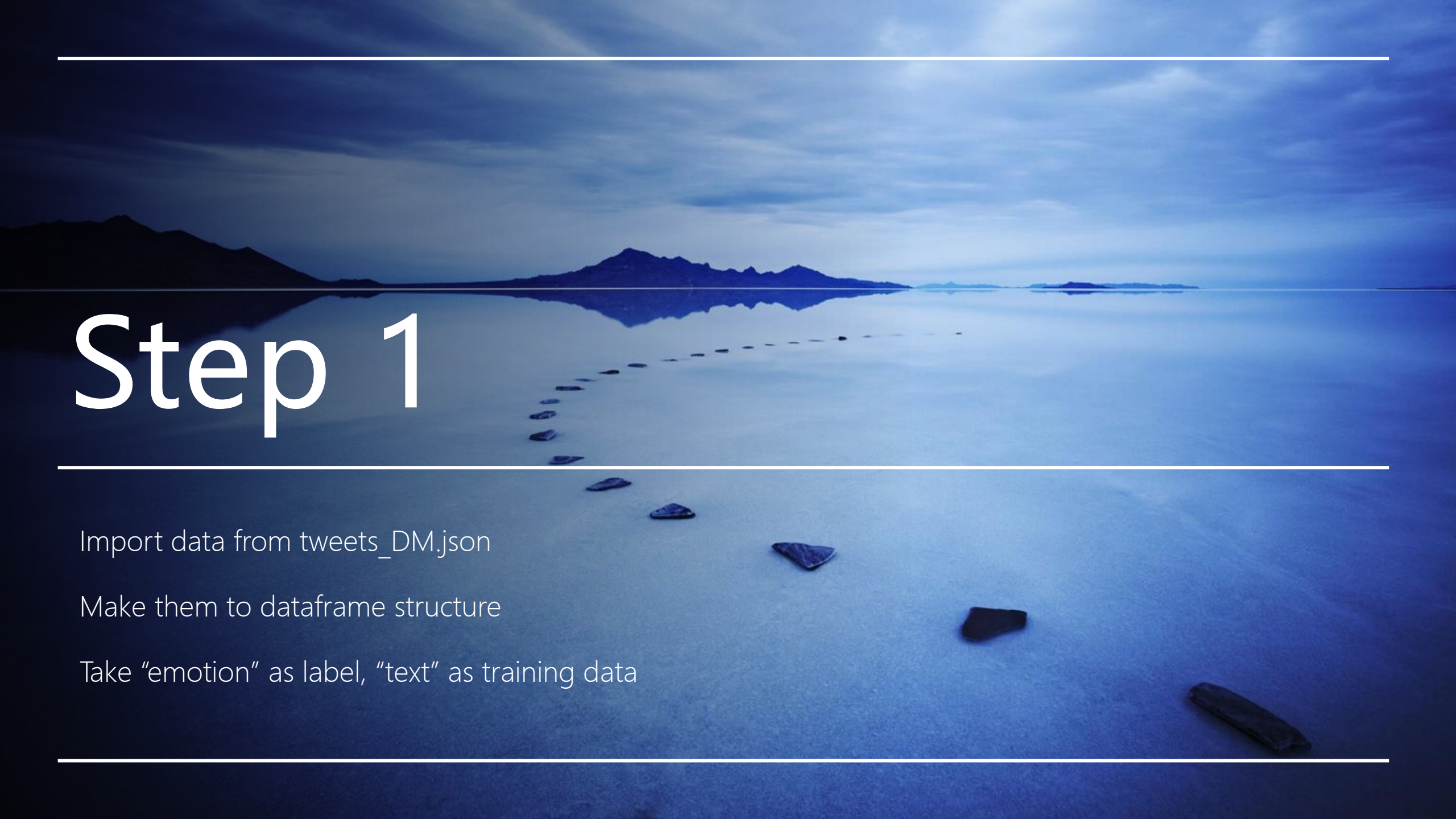
Kaggle report

110062466 孫槐駿



Step 1

Import data from tweets_DM.json

Make them to dataframe structure

Take “emotion” as label, “text” as training data

Then

I was wondering which way should I do to preprocess the training data?

First

Using bag of words

Using BOW_500 to represent text

RESULT

1.

0.29558	0.29427

Second

I want to move on to get better grades

But, HOW?

How about decision TREE?

Then

I used DecisionTreeClassifier to fit train and label

After that, I predict the test data

And upload to kaggle

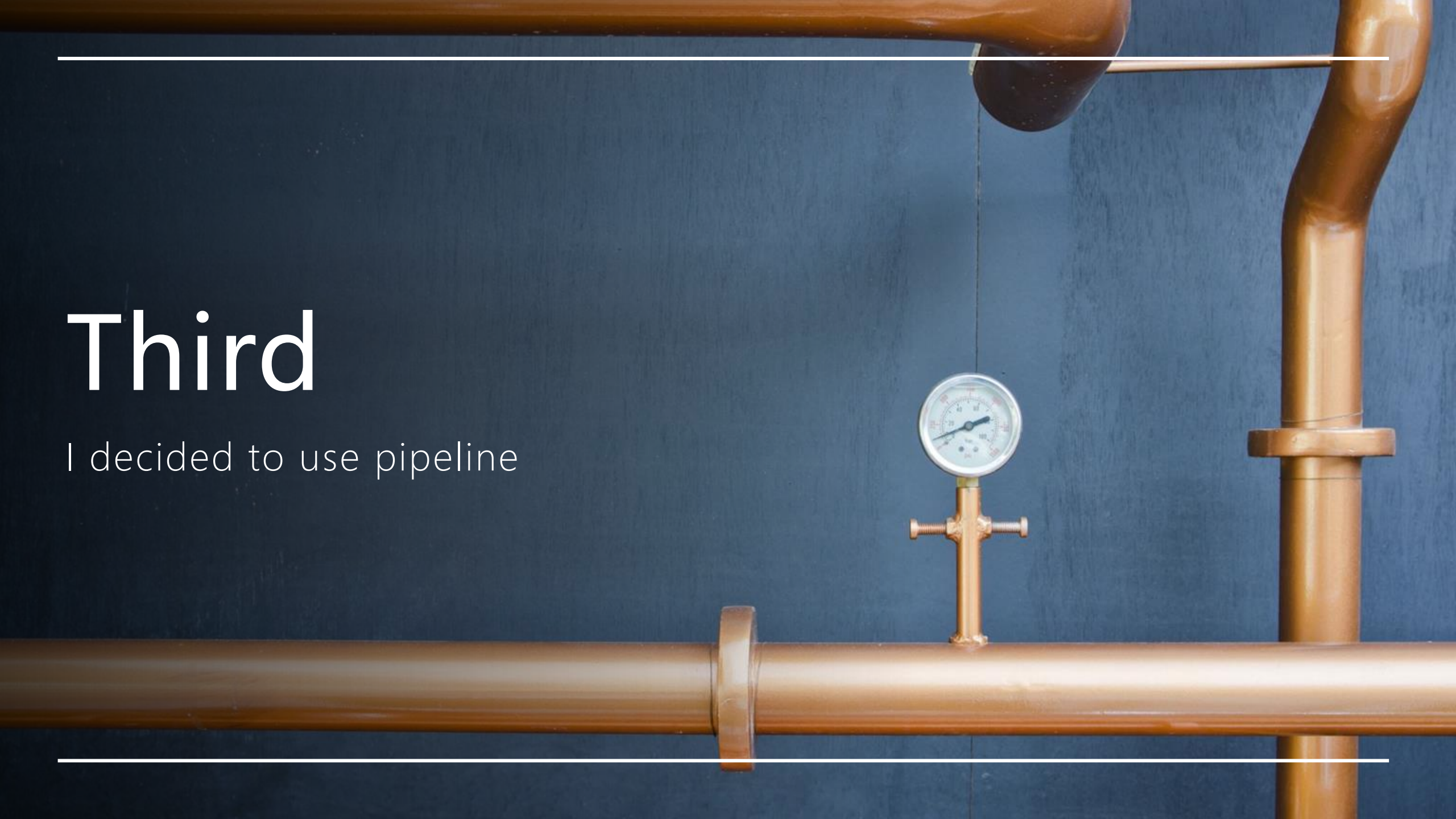
So

Better, But

Not That Better

0.38305

0.38374



Third

I decided to use pipeline

Step 2

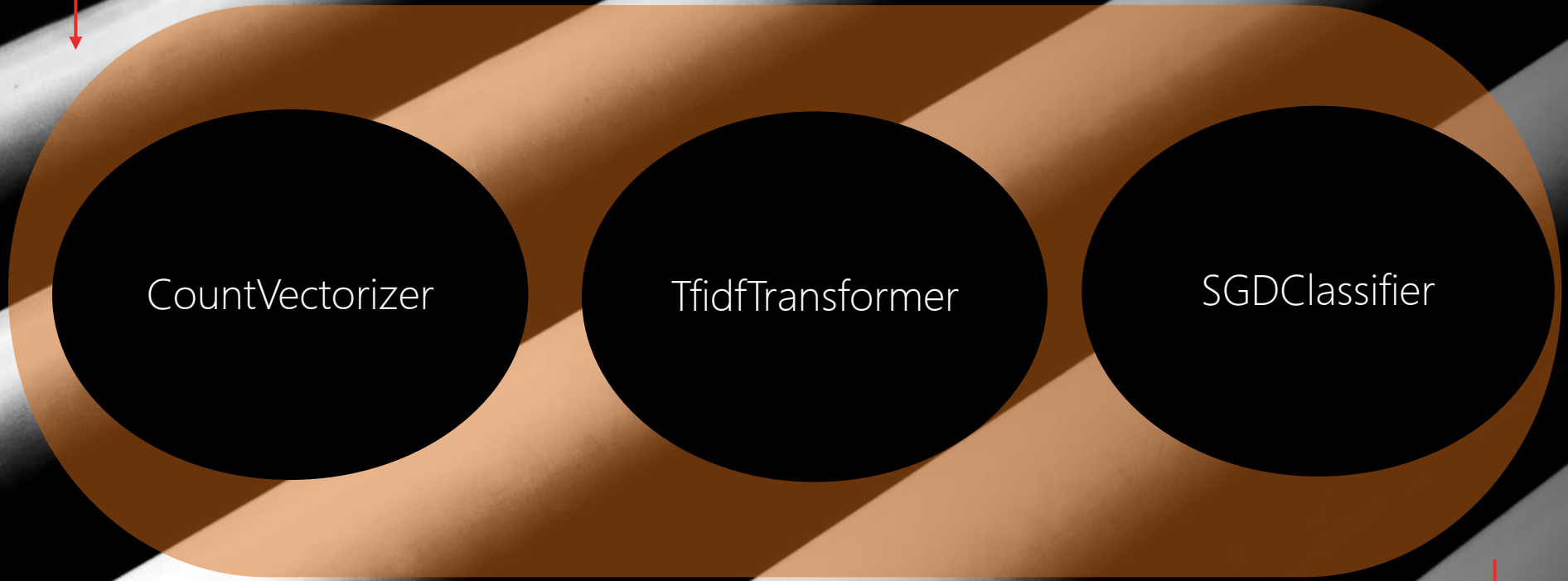
I installed nfx

Which is a good way to eliminate bad words, like user handles, stop words, and punctuations

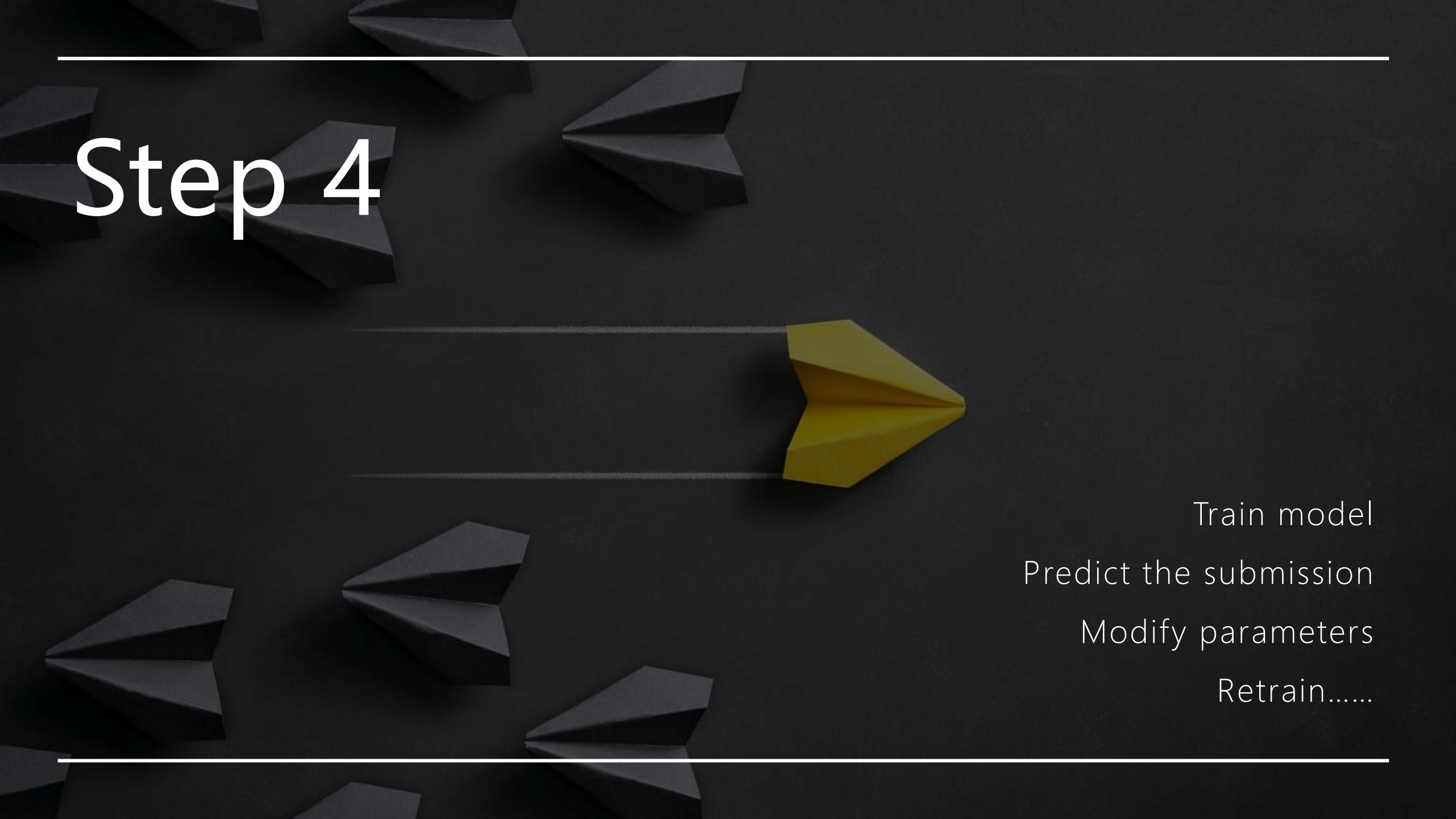
Step3

Installed Pipeline, TfidfTransformer, and SGDClassifier

PIPELINE



Step 4



Train model

Predict the submission

Modify parameters

Retrain……

Finally

Conclusion

In this competition

I've learned how to build Bag of words, decision tree, and pipeline model

Before that, I was wondering why I could train model in sklearn hundreds of and thousands of times?

But in this competition, I found out it was a good way to modify the parameters with the model in pipeline to increase the accuracy