Introduction to Natural Language Processing: Takeaways

Syntax

- Splitting an array into random train and test subsets:

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(counts, submissions["upvotes"],
test_size=0.2, random_state=1)
```

- Initializing the LinearRegression class:

```
from sklearn.linear_model import LinearRegression
clf = LinearRegression()
```

- Predicting using a LinearRegression model:

```
predictions = clf.predict(X_test)
```

Concepts

- Natural language processing is the study of enabling computers to understand human languages. Natural language processing including applications such as scoring essays, inferring grammatical rules, and determine emotions associated with text.
- A bag of words model represents each piece as a numerical vector.
- Tokenization is the process of breaking up pieces of text into individual words.
- Stop words don't tell anything about the document content and don't add anything relevant. Examples of stop words are 'the', 'an', 'and', 'a', and there are many others.
- To calculate prediction error, we can use the mean squared error. The mean square error penalized errors further away because the errors are squared. We often use the MSE because we'd like all our predictions to be relatively close to the actual values.

Resources

- Natural Language Processing
- Bag-of-words model