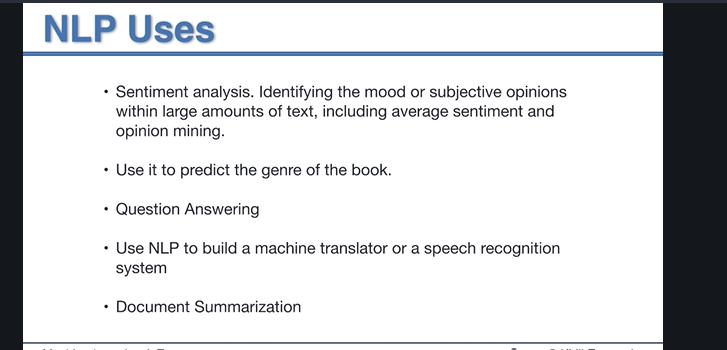
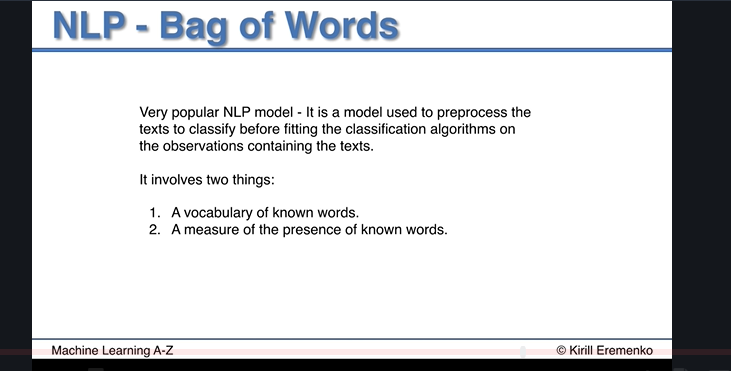
Speaking of classification algorithms, most of NLP algorithms are classification models, and they include Logistic Regression, Naive Bayes, CART which is a model based on decision trees, Maximum Entropy again related to Decision Trees, Hidden Markov Models which are models based on Markov processes.

A very well-known model in NLP is the Bag of Words model. It is a model used to preprocess the texts to classify before fitting the classification algorithms on the observations containing the texts

Uses of NLP (Nature Language processing):

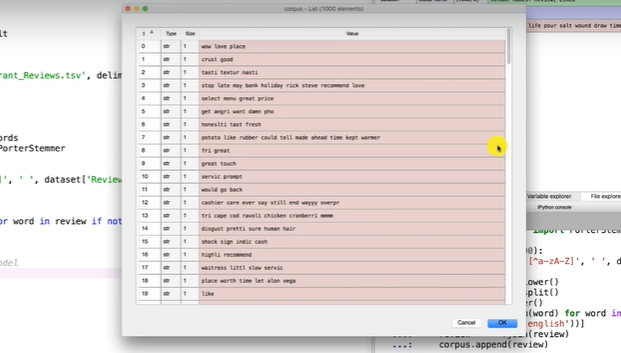




We used tsv file here for because csv contains “,” which means something.

Before processing the data, we need to clean the text, eg “on”, “the”, “and” etc, since they are not relevant word that can help to predict whether the review is positive under this context. Also, make sure every character is lowercased etc.

We need to filter unwanted text with stopwords from NLTK. Then we need to do “stemming” to store the root of the word, eg “loved”, “loving” to “love”



\*\*Note: In python, the compiler loops through “set” quicker than “list”, so turn the stopwords into a set in the loop so that it will run faster.

\*\*Note: Corpus is a collection of text

Bag of words is about creating one column for each word. It is a sparse matrix, and sparsity means a lot of 0 in the table, which is something we would like to reduce.

In this scenario, kernel SVM works the best.

With 0.25 test data:

Accuracy: 0.736

Precision: 0.734

Recall: 0.789

With 0.20 test data (this is a more secure result based on low precision):

Accuracy: 0.735

Precision: 0.715

Recall: 0.8058