Sentiment Analysis with Logistic Regression

Supervised ML and Sentiment Analysis

Diagram

Description automatically generated

Sparse representation (Feature extraction based off a vocabulary)

Sparse representation first builds a vocabulary based off a training set and then extracts features based off the sparse representation.

Graphical user interface, application

Description automatically generated with medium confidence

However the feature vector would equal the size of your entire vocabulary:  
Graphical user interface, text

Description automatically generated

This would result in large training and prediction times.

Positive and negative counts

One way of sentiment analysis is positive and negative count as shown below. This is done by calculating the word frequency in classes.

Graphical user interface, text, table

Description automatically generated

`Table

Description automatically generated

Above is a frequency dictionary, which maps a word and the class to the number of times that word showed up in a class.

Feature Extraction with Frequencies

A picture containing text

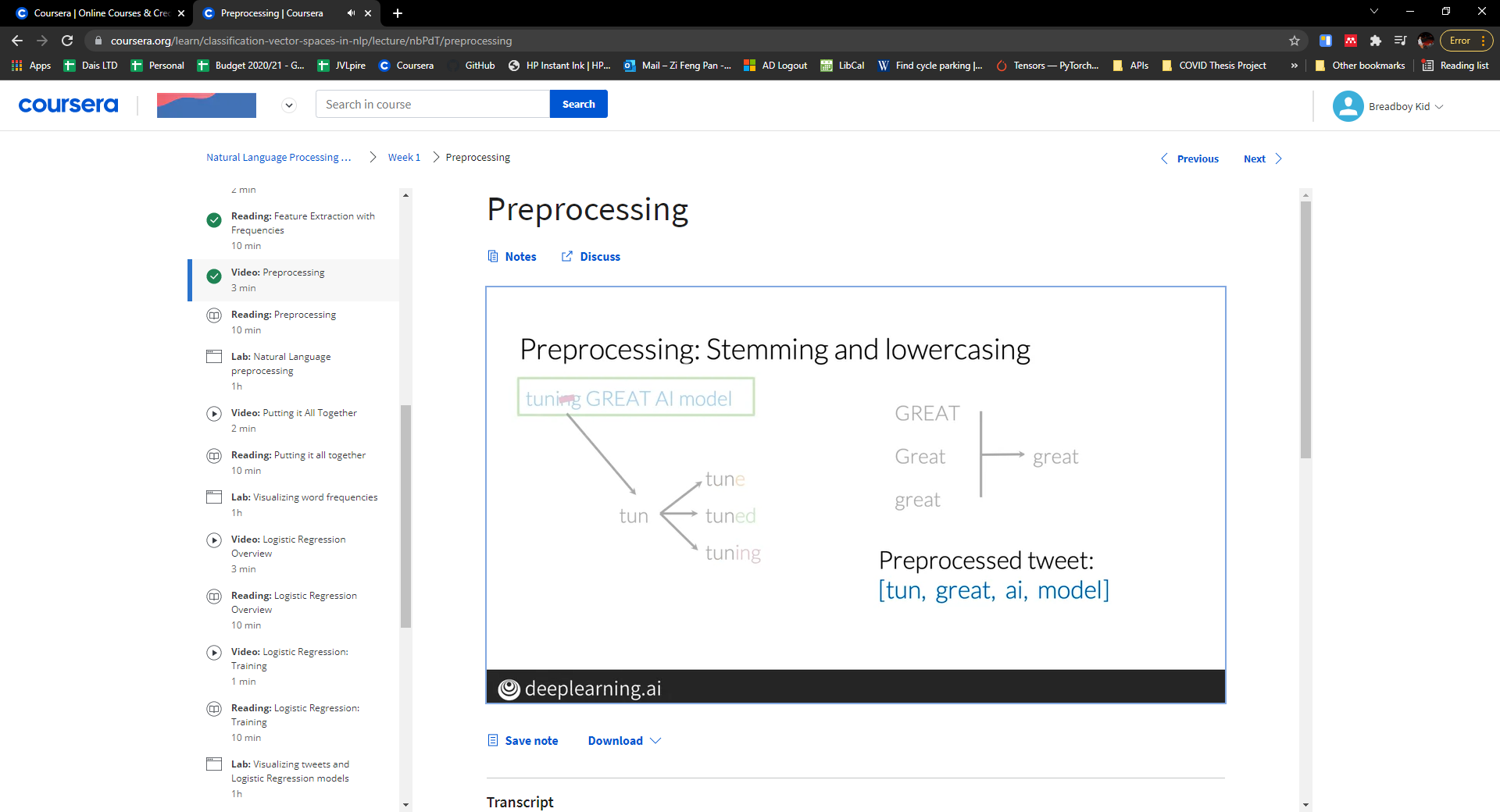
Description automatically generated

Here instead of having a vector of dimension V (your vocab), you only need a vector of dimension 3.

Pre-processing

Stemming

Stemming and lowercasing to reduce and match common words together.



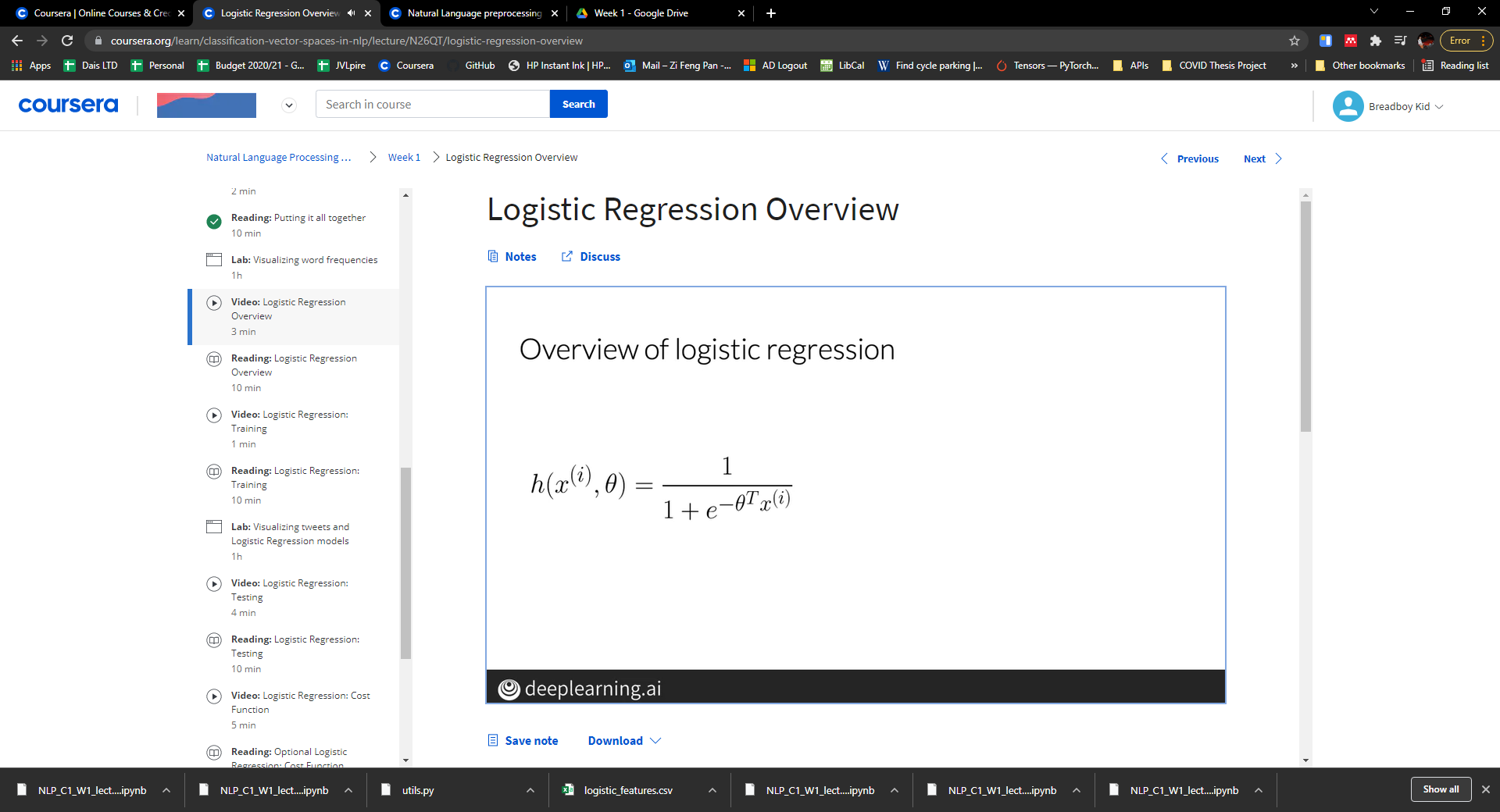
Stop words

To pre-process a text you need to remove stop words. This can also be done with punctuations, but only if these are deemed unimportant for the contexts.

A picture containing diagram

Description automatically generated

Logistic Regression Overview

#