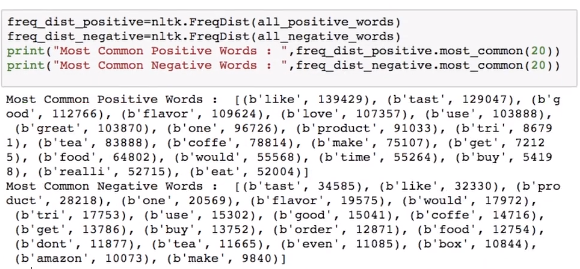
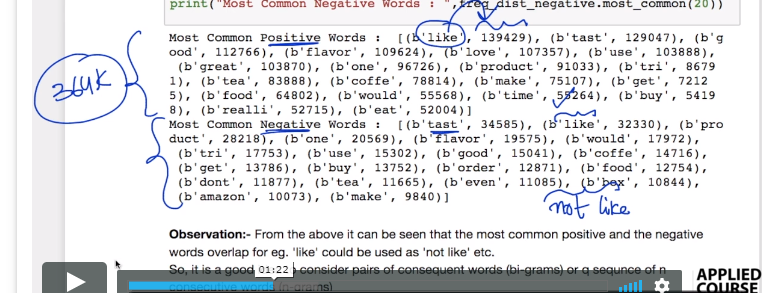
Till now what we have done is we have distinguished all the positive words and all the negative words.

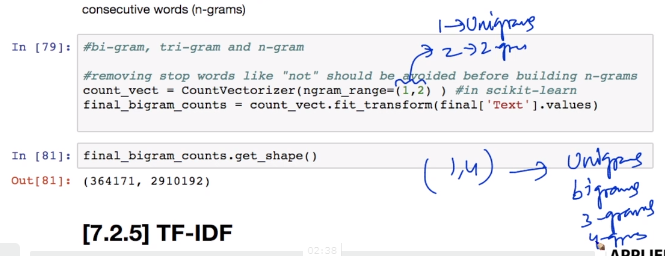
Now first thing we will do is find the most frequent positive word and most frequent negative word.



When we observe above output we see that “**like”** is occurring very frequently in both positive and negative which actually doesn’t make any sense. Because in negative reviews it must be somewhat “**Not like”** .



To resolve this issue we need **n-grams**

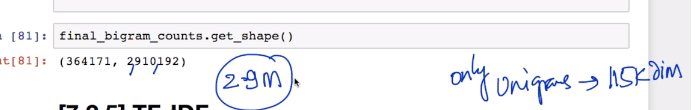


ngram\_range is basically a tuple with minimum and maximum value be from where and till where.

i.e., if it is (1,4) so it will consider all unigrams, bigrams, trigrams and 4-grams.

When we were only considering unigrams in bag of words we were getting 115281

But when we are **using bigrams we are 2.9 million dimensions**



**Comments:**

* "not" is not present in cleaned Text.how can n-grams identify the negative words?

It is because the word "not" and "very" are already present in the english stopwords of nltk library. As we are directly removing all te stopwords that are associated with nltk library for english language, we do not see the words "not" and "very" in the cleaned form of the text.

When we are building uni-grams, it is always better to remove the words "not" and "very" as they do not add much meaning. But when we are building n-grams(n>1) then it is recommended to keep the words "not" and "very" in the text and remove the remaining stopwords.

* max\_df is used for removing terms that appear **too frequently**, also known as "corpus-specific stop words". For example:

max\_df = 0.50 means "ignore terms that appear in **more than 50% of the documents**".

max\_df = 25 means "ignore terms that appear in **more than 25 documents**".

The default max\_df is 1.0, which means "ignore terms that appear in **more than 100% of the documents**". Thus, the default setting does not ignore any terms.

* min\_df is used for removing terms that appear **too infrequently**. For example:

min\_df = 0.01 means "ignore terms that appear in **less than 1% of the documents**".

min\_df = 5 means "ignore terms that appear in **less than 5 documents**".

The default min\_df is 1, which means "ignore terms that appear in **less than 1 document**". Thus, the default setting does not ignore any terms.