Natural Language Processing

Natural Language Processing Intuition

Why does the Bag of Words model replace all the capital letters by the lower cases?

Many computer languages, in particular Python and R, treat capitals and lower case letters as completely different symbols. So if we won’t convert everything to lower cases then we are to take into account that some words may contain capitalized letters and include additional code for them. Much more simpler to replace all capitals by lower cases.

What is sparsity?

sparsity occurs when you have lots and lots of zeros in your matrix (therefore called a sparse matrix). So when we reduce sparsity, that means we reduce the proportion of zeros in the matrix.

Natural Language Processing in Python

Could you please explain what ’quoting = 3’ means?

quoting = 3 will ignore the double quotes in the texts. The number of observations in the dataset didn’t match the number of texts which was due to a splitting anomaly with the double quotes. The potential problem is that if you have double quotes in a review, this review can accidentally be separated and considered as another review with no outcome. So to make sure we avoid this kind of situation we ignore the double quotes by using quoting = 3.

How can we download all the NLTK material in one shot?

To download all the material in NLTK in one shot you need to open a terminal and run:

pip install -U nltk

Could you please explain what is PorterStemmer()?

PorterStemmer() applies stemming to the words in your texts. So for example, "loved" will become "love" after the stemming.

How to do NLP in French or other languages?

Check if there are some NLP classes specific to your own language, as it is the case in French for example:

from nltk.stem.snowball import FrenchStemmer stemmer = FrenchStemmer()

and for the stopwords, include this in your code:

set(stopwords.words(’french’))

Why did we change the reviews from lists of words back to strings?

It is done because the fit\_transform() method from the CountVectorizer() class requires strings to work.

What is the difference between "TF-IDF" and "CountVectorizer"?

The difference is that TF-IDF applies normalization, unlike CountVectorizer. Otherwise it is the same, if you input the parameter ’normalize = None’, then it’s very similar.

To use TF-IDF, it’s the same as for the other NLP techniques, you just import the class you want (say TfidfVectorizer from ’sklearn.feature\_extraction.text’), then you create an object of this class while inputing the parameters you want to preprocess your texts.

Why do we delete ’NOT’, given that "Crust is good" is not the same as "Crust is not good"? It also works the other way: "Crust is bad" is not the same as "Crust is not bad". Hence, this is not usually taken into account. Hence we believe/hope that on average, the number of misconceptions caused by this would be less in total.

After we are done with cleaning and getting our bag of words model, why are we doing classification?

We are doing classification on the reviews to predict the outcome of new ones, exactly like sentiment analysis. Only to make these predictions in the best conditions, we need to apply the bag of words model first. That will prepare the data in a correct format for it to be fitted to the classifier.

What model seems to be the best solution to the Homework Challenge?

Definitely Naive Bayes, which by the way is usually recommended first when doing Natural Language Processing.