IST736 Text Mining

HW6

CLARIFIED in Blue

Benoulli and Multinomial Naïve Bayes in Sci-kit Learn

1. Use the required format for this assignment
2. For this assignment you will use the Benoulli (so here – be sure to use the binary=True in the CountVectorizer). So, you will need to create two dataframes – one for Bernoulli and one for the normal frequency counts. For those who want to do more, you can also use the TfidfVectorizer and create a third dataframe that is normalized. This is fast to do and will create an interesting opportunity to compare, binary, count, and tf-idf normalized. However, it is optional.
3. Next, on the frequency data frame (and on the tfidf if you create one) use the Multinomial Naïve Bayes algorithms
4. **Datasets:**
   1. First – create your own small, easy, predictable dataset (as either csv or corpus – your choice). Code the above to assure that you code works. You may choose to share this step in the Analysis section under models and methods.
   2. Next, use any labeled dataset (50 rows or more so you get interesting results). You can use csv or corpus. You can use the lie data or the sentiment data from class (and can repurpose any intros that you wrote). You can also find or use and API to get data. There are many levels here and it depends on how advanced you want to go. While it is not required, it would be great for you to try to use an API, clean up the data, and then use the methods on it.

Remember that the Benoulli model takes Boolean vectors as input,

NOTE: A Boolean Vector is the same as \*binary\* data. This means that instead of each word (in each column) being counted up for each document, rather, there is a 0 if the word is NOT in the doc and 1 if it is.

and the Multinomial model takes frequency vectors as input.

NOTE: This means that it uses the normal counts (and also can work on tf-idf or other normalized counts).

NOTE: You will need to create a Training Set and a Testing Set AFTER you clean up and fully prep your data. Remember that Naïve Bayes (as well as Bernoulli) needs to be trained first and then tested. In both cases REMOVE AND KEEP the labels ☺