In Class Activity

Text classification with Naïve Bayesian Model –Activity 5B

So, at the end of last class, what I really heard everybody saying was, we want a more complicated example. We can always make things more complicated!

What about "nonsense" words? What happens when there is no matching word found… that is… the user provided a nonsense word such as "Sharknado" or no word at all?

This is what the big kids call a "zero vector" when talking about vectorized text.

Given a zero vector (or a very sparse vector if the nonsense word is accompanied by other words not in the vocabulary), the Naive Bayes classifier will still make a prediction.

However, this prediction will be based solely on the prior probabilities of each class (i.e., the overall distribution of classes in the training data) because the input does not provide any additional information to sway the prediction towards one class or another.

Essentially, the model will default to its "best guess" based on the most common class or a distribution it learned during training. This is similar to what we saw in the k-NN model when K was set to be too large.

When the model fails to give an intelligent answer, the default response will be driven by the dominant (most popular, most numerous) traits of the data set.

So, if most of the training documents went to "Customer Support" then the text will be assigned to customer support. (Which, if you think about it, is not unreasonable!)

The practical outcome is that the prediction for a nonsense word input will not be meaningful or based on any real understanding of the input text. It highlights a limitation of such models when encountering out-of-vocabulary (OOV) words or phrases.

To improve handling of OOV words or enhance the model's ability to deal with unexpected inputs, you might consider tokenizing the input techniques like word embeddings (which can capture semantic similarities between words, including those not seen during training) or incorporating a wider variety of preprocessing steps to better manage rare or unknown words.

Instructions

1. Students should retrieve the activity5B.zip file and unzip the **activity5B** folder into your mis3536workspace. (There are three files, the Jupyter notebook, a csv file, and an excel file.)
2. Run the new and improved script. Your instructor will highlight its new and improved features. Take note of how we can observe the accuracy of the model.
3. Break into groups.

4. We are going to use our Bayesian classification script on a new data set. But first we will need to clean it.

   Our new data set is for the BigBiz IT Services Company.

   This company provides a wide range of tech support and customer service for businesses and consumers alike. They specialize in troubleshooting IT issues, offering helpdesk support for both hardware and software problems. The company is also involved in e-commerce, supporting online stores with technical services and customer assistance. They handle service requests ranging from login issues, system malfunctions, and hardware setup, to more specific problems related to online store management, including payment issues, account access problems, and product-related inquiries. BigBiz IT Services clientele includes individuals, small businesses, and larger companies looking for reliable, round-the-clock customer service and technical solutions.

   Our goal is to automatically assign incoming problems to the correct support queue.

5. You will need to clean that spreadsheet a bit, change its data format to csv and identify which columns should be used for the "document" and the "classification". Recall that all the training data is "the corpus".

   DISCUSS: What other data cleaning needs to take place?

6. Keep working in your group. Clean up the data in Excel. Edit the Python script. Run the script (one cell at a time) and experiment with the resulting model by running / rerunning the third cell.

7. Examine the reports generated by the script. Use those to add new and improved custom "Stop Words"

   DISCUSS: What was your strategy for choosing stop words? Share what your custom "stop words" were and if you were able to improve the accuracy of the model.

8. Rerun your analysis and test it. Is your classification model better, or worse?

9. When you are done, every student should **upload their notebook, cleaned spreadsheet, and pdfs.**