NLP Questions

1. You just trained a model to classify soft drinks according to their flavor-profile. The results are below. What has the model learned about the correspondence between the item name and the classification? Can you find any problems with what the model has learned? How would you address this problem?

Ans- **The model has learned that the flavor profile of a soft drink can be predicted based on the item name. Specifically, the model has learned that drinks with certain brand names are associated with specific flavors, such as Pepsi and Coke being associated with cola flavor and 7up and Sprite being associated with lemon-lime flavor.**

**One problem with the model's learning is that it may not be accurate for all soft drinks. For example, there may be drinks that are not accurately classified by the model, such as drinks with unique flavors or new flavors that the model has not encountered before.**

**To address this problem, the model may need to be trained on a larger dataset with a wider variety of soft drinks and flavors, or more features (such as ingredient lists) may need to be included to improve accuracy. Additionally, the model may need to be regularly updated and retrained to adapt to changes in the soft drink market and new products**.

|  |  |
| --- | --- |
| **item\_name** | **flavor\_profile** |
| 7up can 355ml | lemon-line |
| pepsi bottle 14oz | cola |
| 7up bottle 500ml | lemon-line |
| 7up can 500ml | lemon-line |
| pepsi bottle 20oz | cola |
| coke bottle 14oz | cola |
| pepsi can 20oz | cola |
| coke bottle 20oz | cola |
| sprite bottle 500ml | lemon-line |
| sprite can 500ml | lemon-line |
| sprite can 355ml | lemon-line |
| coke can 20oz | cola |
| sprite bottle 355ml | lemon-line |
| 7up bottle 355ml | lemon-line |
| pepsi can 14oz | cola |
| coke can 14oz | cola |

1. You’ve been given a product list of ~16,000 carbonated soft drinks (see carbonated\_soft\_drinks.csv) and you’re asked to build a model that can extract the product’s brand (e.g., Coke, Pepsi Sunkist), when available. What type of model would you use? If you choose a supervised-learning model, where will you get your training data?

**I tried to preprocess data using NLP as text data can be preprocessed by removing stop words, stemming or lemmatizing words, and converting text to numerical data using techniques like bag-of-words or word embeddings but while training on models I was unable to calculate accuracy for a supervised learning model but later on, I used function definition to extract brand names with the help of nested loop and match=re.search function from the import re module and then I used the supervised learning model which provided me with 99.68% accuracy using Naive bayes algorithm. Also, I manually imported brands = ["Coca-Cola", "Pepsi", "Mountain Dew", "Dr. Pepper", "Sprite", "Fanta"] in list format as training data.**