Applied

Text Classification

Guest lecture by Saul Johnson - Exercises

Getting Warmed Up

1. Head to this page (you'll need to create a Kaggle account if you haven't already) and download the SMS spam/ham dataset as a CSV: https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset
2. Split the CSV down into separate text files organised by subfolder. Use Python, like we did in class.
3. Now, run some text classification models over this data. Try training Naive Bayes, SVM and Random Forest each at least 5 times. Which gives you the highest accuracy on average?

From Unigrams to N-Grams

1. Right now, we're training on unigrams (single words). What is the main disadvantage of this approach? Discuss, then we'll do so as a class.
2. How can we adapt our pipeline to use bigrams/trigrams in addition to unigrams? *Hint: check out the ngram_range parameter of CountVectorizer.*

A Much Larger Dataset

1. Now, download the much larger IMDB positive/negative movie review dataset from this page: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews
2. Prepare it just as before by breaking down the large CSV file into individual text files.
3. Try training a classifier as before? Is it too slow? Write a Python script to pick only 500 files from each directory to use for training instead of the full 25,000.

Enhancing Tiny Datasets

1. Download the tiny dataset from here and try training different models with it a few times each. Notice anything about the accuracy of those models? What do you think is causing this?
2. Research "data augmentation" and read this article on automatically paraphrasing text using Python.
3. Apply this technique to increase the size of the dataset, and try training your models again. Notice an improvement?

Note: You are encouraged to tackle pre-processing these datasets in Python (e.g. converting from CSV to separate text files). If you get stuck at any of the data processing steps, however, I've made the ready-to-use datasets available for download here:

- **SMS spam:** https://sauljohnson.com/downloads/datasets/sms_spam.zip
- **Movie reviews:** https://sauljohnson.com/downloads/datasets/movie_reviews.zip
- **Main courses and desserts:** https://sauljohnson.com/downloads/datasets/mains_and_desserts.zip

Delivered as part of a Guest Lecture at NHL Stenden Leeuwarden - 25/11/2022