Lab 3

DATA 441/641

Applied Natural Language Processing

Binary classification of COVID-19 tweets using tf-idf vectorization & PCA

1. Import modules and data (2 points)

Import the content from the data file that is provided to you on Canvas.

2. Clean and normalize text (5 points)

Remove the pound sign from hash tags, drop words with fewer than 2 characters, and drop any punctuation words. Perform lemmatization.

3. Instantiate vectorizers and topic models. (3 points)

For this part use PCA to reduce your feature dimensions. Remember to create a sparse-to-dense transformer.

4. Set up a cross validation scheme, optimize the hyperparameters of SVM, and print out the accuracy, precision, recall, and f1 score. (10 points)

What are the parameters that provide the best classification performance?