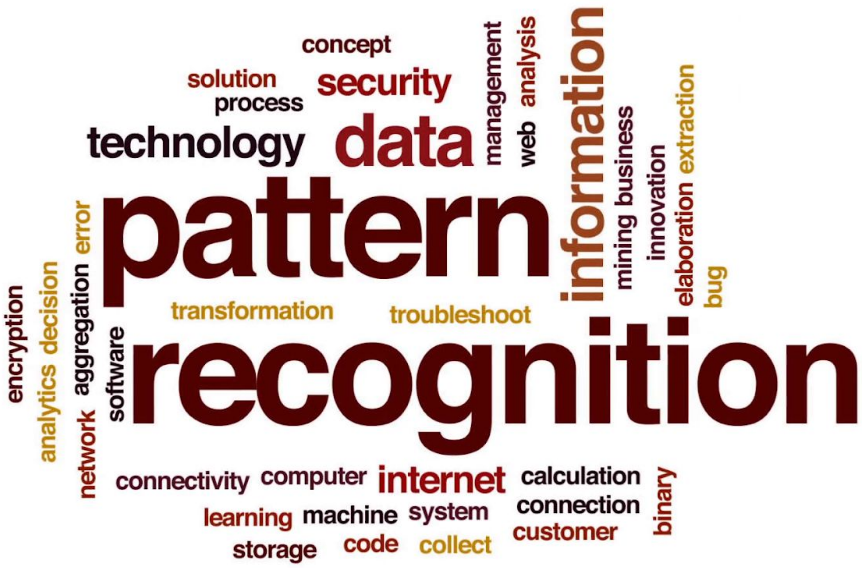

Text Classification

Take seven different samples of Gutenberg digital books, that are of seven different genres and authors, that are semantically different. Separate and set aside unbiased random partitions for training and test.

The overall objective is to produce classification predictions and compare them; analyze pros and cons of algorithms and generate and communicate the insights.

Prepare the data: create random samples of 200 documents of each book, representative of the source input.
Preprocess the data; prepare the records of 150 words records for each document, **label** them as a, b and c etc. as per the book they belong to.
Transform to BOW, and TF-IDF, n-gram, etc.
Train a machine that can tell which author (or genre), when asked!
Evaluation: Do ten-fold cross-validation.
Perform **Error-Analysis**: Identify what were the characteristics of the instance records that threw the machine off.
Document your steps, explain the results effectively, using graphs.
Verify and validate your programs; Make sure your programs run without syntax or logical errors.

Rubric: (accounts for 20% of the final grade.)

Choose data of your choice, (labeled data) 1%
Preprocessing and Data Cleansing 1%
Feature Engineering 2%
Use SVM, Decision Tree, k-Nearest Neighbor 2%
Perform Evaluations, 2%
Compare and decide which algorithm is performing as the champion model 0.5%
Perform Error Analysis, 2%
Perform Visualizations, Graph the results 2%
One or Two pages of presentation 0.5%
Report, detail explanations 2%