## Solution description

Our solution is to create an analysis of bodies of text by classifying keywords in the text that signify the text belongs to or is describing a specific category or genre of writing. We will do this using the sklearn python library, namely the Pipeline module or the GridSearchCV module.

We will likely select a list of common stopwords such as ‘and’, ‘or’, ‘then’, ‘this’, ‘that’ in order to avoid confusing the model and decreasing the margin for error/confusion.

The data will require pre-processing as there are missing or incomplete values for many of the records, and some of the records have multiple genres, which will not only require us to split the genre string by commas into a list of strings, but design our model so that it can potentially receive multiple classifiers for one body of text; one solution might be to split one record with three categories into three separate records with one category each. It has also not been taken off the table that we could mix multiple datasets together in order to have a larger variation of writing styles and (hopefully) more accurate predictions.

We will be building a web API for our model, this will either be built using flask, or Django – both of which are python based web frameworks. Our GUI will either be built using either python’s ktinter GUI library or a web based GUI which could potentially be a lone html file.

## Data Description

Our data consists of bodies of text (in the case of our data, movie synopsis or a “plot”), and matching genres for the movies, movies which have a genre will be used as a training set. Some movies have missing categories, these records can be separated off and used as a practical test set as opposed to a prediction accuracy test set. The data includes these features:

* Release year
* Title
* Origin/ethnicity
* Director
* Cast
* Genre
* Wiki Page
* Plot

Some of the records do not include a plot, these records will be cleansed in the data pre-processing. There are 34894 records in total. 17% of the records belong to the drama genre, 17% have unknown genres, there are supposedly 2263 unique genres, however, this number counts two, comma separated genres as an entirely new genre. For example, if a record has “historical, thriller” as its genre, this will not count as thriller or historical, but an entirely new entity, the same rule applies for “short action/crime western” for example. All features apart from genre and plot will be dropped from the table.