Calofir Ionut

Homework 2

For this homework we had two main tasks:

- Classification
- Clustering

1. Dataset

The dataset was composed of samples of text from 20 different authors. There were 20 samples for each author.

Steps for processing the dataset:

- I **tokenized** the texts
- For every text I applied **nltk POS Tagger**
- I only kept **nouns** and **verbs**
- I applied **lemmatization** for every word
- I removed all the **stopwords** and the words smaller than 2 chars
- I split the dataset in 3 sets: **train**, **validation(15%)** and **test(15%)**

2. Classification

For this problem the task was to predict the author of a text. My approach was to try a **SVM** with different parameters.

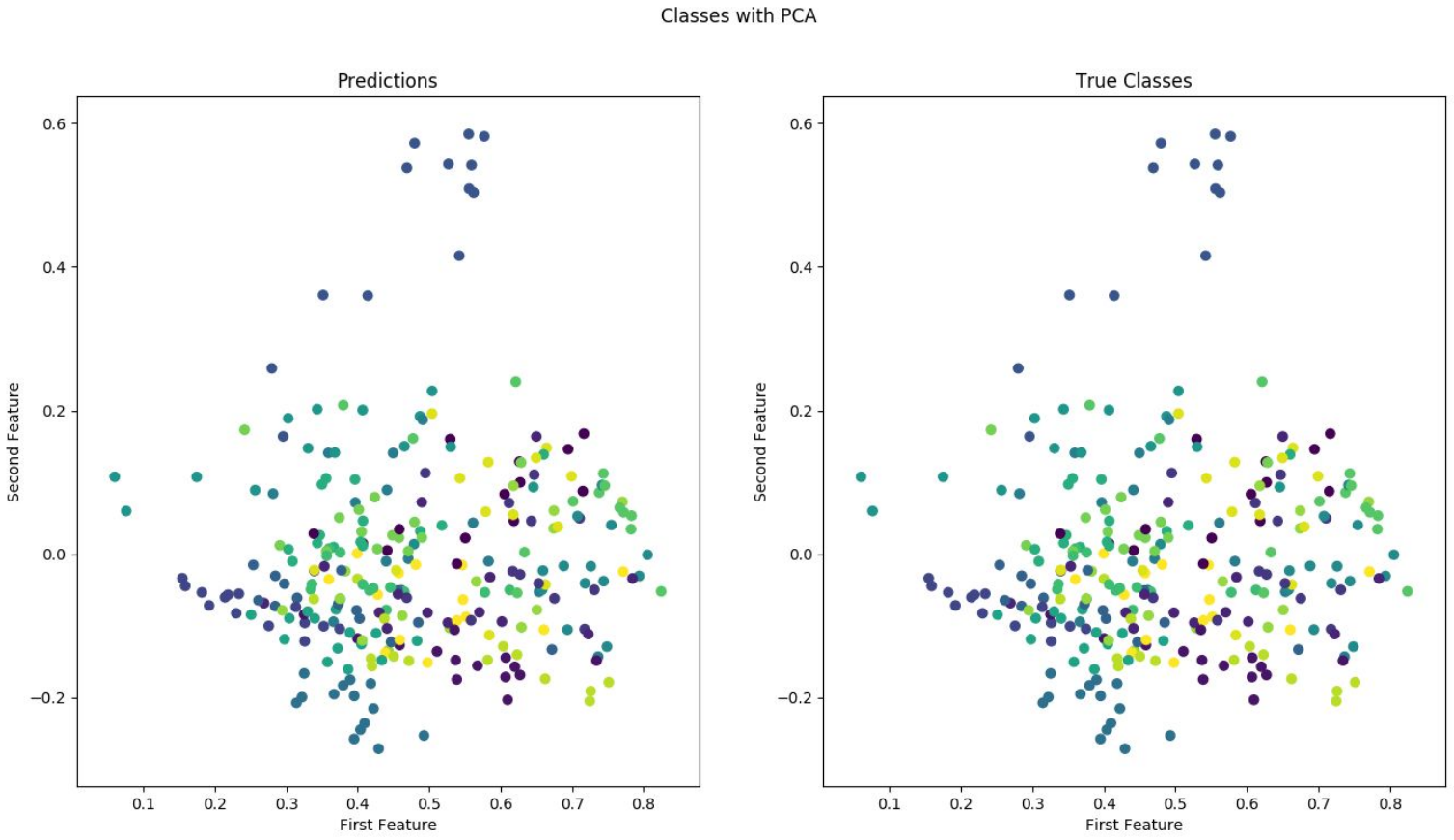
For text embedding I tried **TfIdf** and **CountVectorizer**. Another solution would have been **word2vec**, but unfortunately I didn't have enough RAM and I couldn't use **word2vec**.

Training

To decide which model is the best I implemented a **grid search** and I tested the classifier on the **validation set**. I set the random seed to be the same for every combination of hyperparameters so I can have the same validation test. I didn't use the grid search implemented in sklearn because of memory problem with my laptop.

The parameters were **TfIdf(True/False)**, **CountVectorizer(True/False)** (only one at a time), **PCA(True/False)** (I applied PCA after TfIdf or CountVectorizer), **SVM__C**, **SVM__kernel**, **SVM__gamma**. And the metrics were: **accuracy**, **precision**, **recall**, **f1**.

All the results can be found in **results/clf_results.csv**.



PCA with 2 components applied on the train set

3. Clustering

For this problem the task was to cluster the texts with 3 different algorithms (**KMeans**, **DBSCAN**, **Hierarchical(Agglomerative)**) and compare them.

For text embedding I tried **TfIdf**.

Training

To decide how the models perform I implemented a **grid search** so I can test different combinations of hyperparameters.

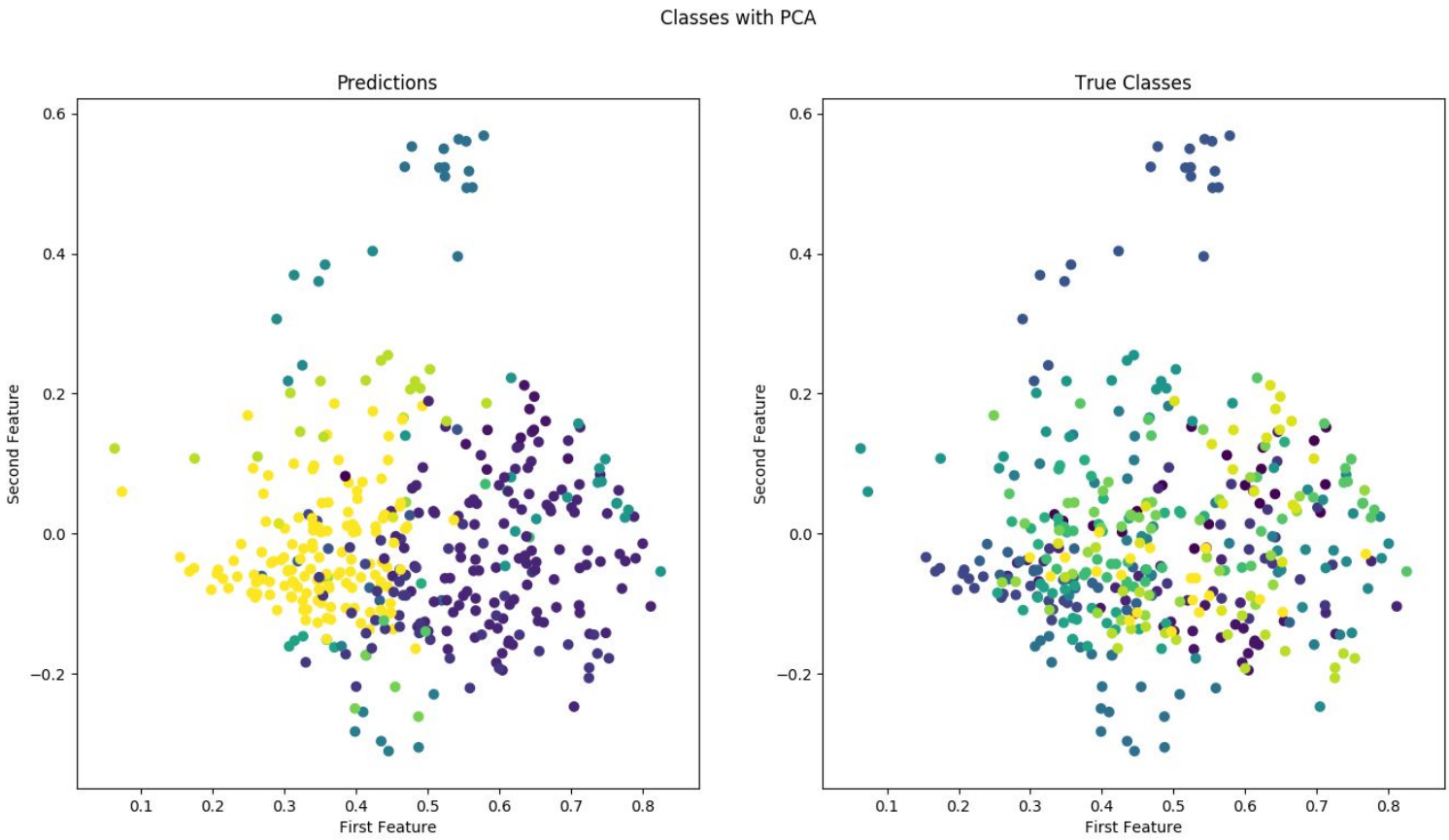
For **KMeans** the hyperparameters were **n_components** and **init**. And the metric was **silhouette**.

For **Hierarchical(Agglomerative)** the hyperparameter was **n_components**. And the metric was **silhouette**.

For **DBSCAN** the hyperparameter was **eps**. For different values I computed the **labels** so I can see how many clusters I have and how many noisy samples are present.

All the results can be found in **results/cluster_results.csv**.

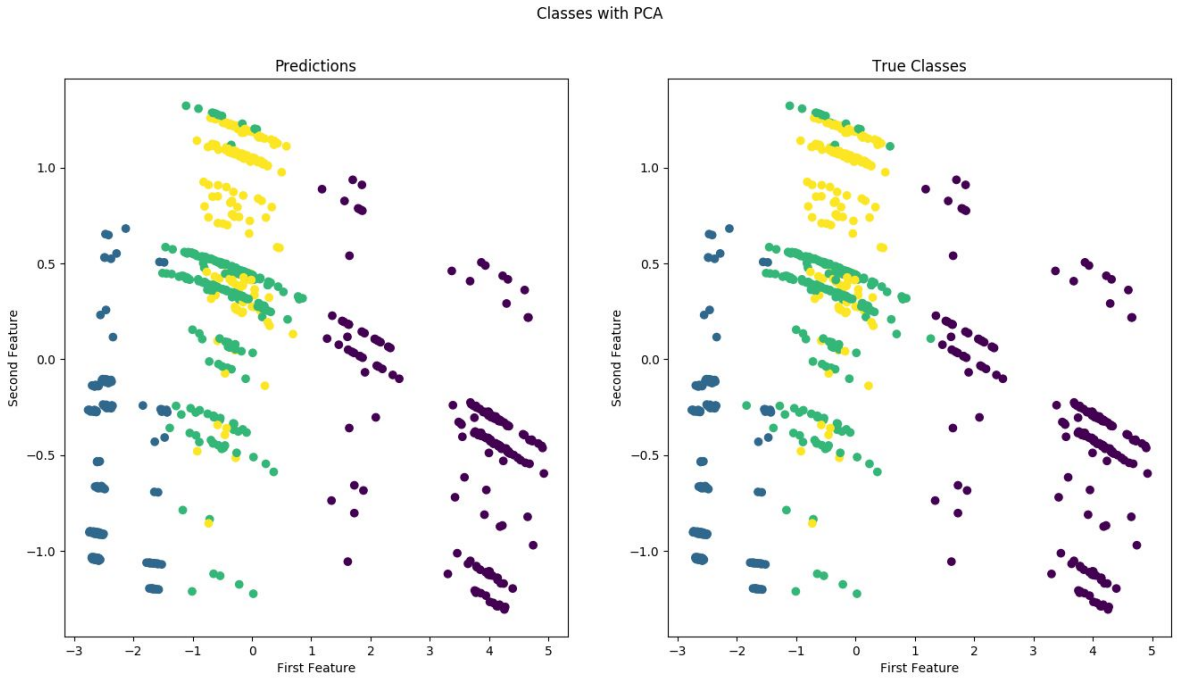
To better understand what the clusters represent I trained a **KMeans** with **20** clusters and **init=k-means++**. Then, for every cluster, I **counted** the number the authors that appear in that cluster. Also, for every cluster I computed the first **20** most frequent words. These results can be found in **results/cluster_data.csv**



PCA with 2 components applied on the whole set

4. Bonus

I also applied PCA on my first homework.



Classification (above) and Regression (below)