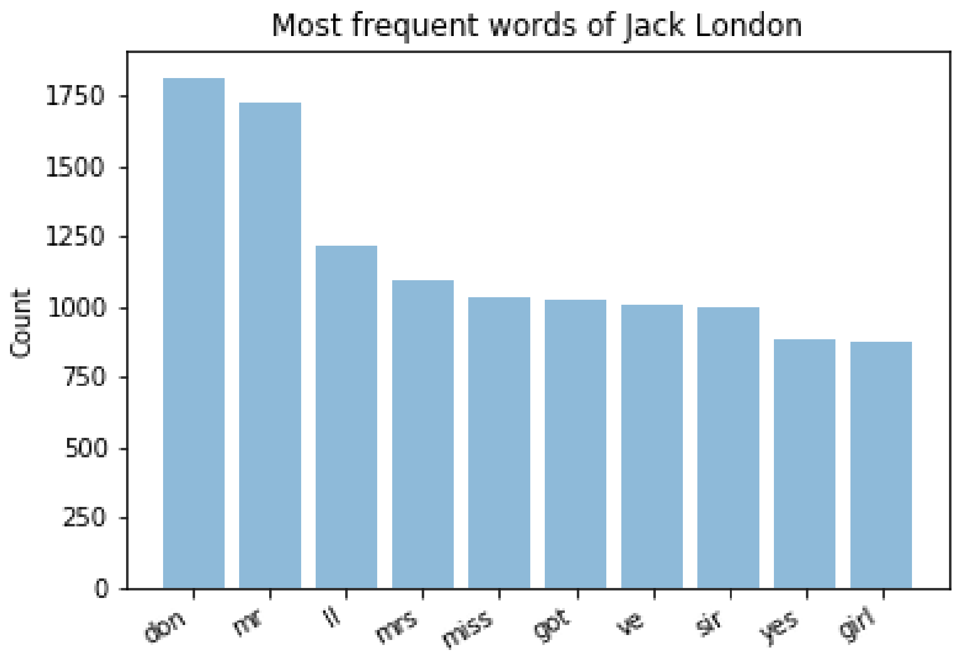
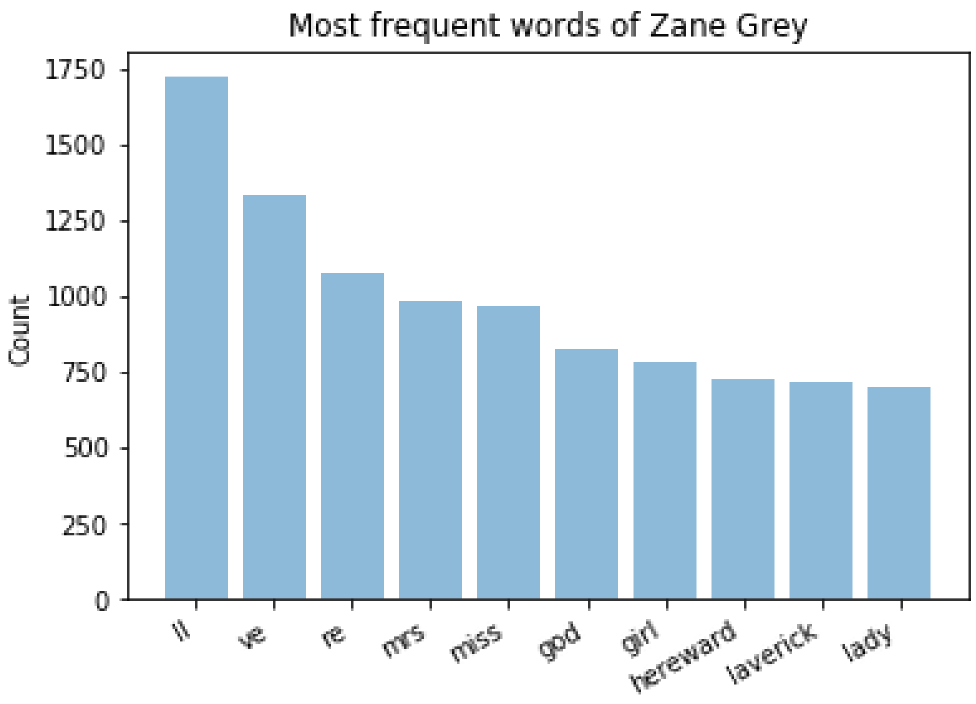
Homework 2: Support Vector Machines & Clustering

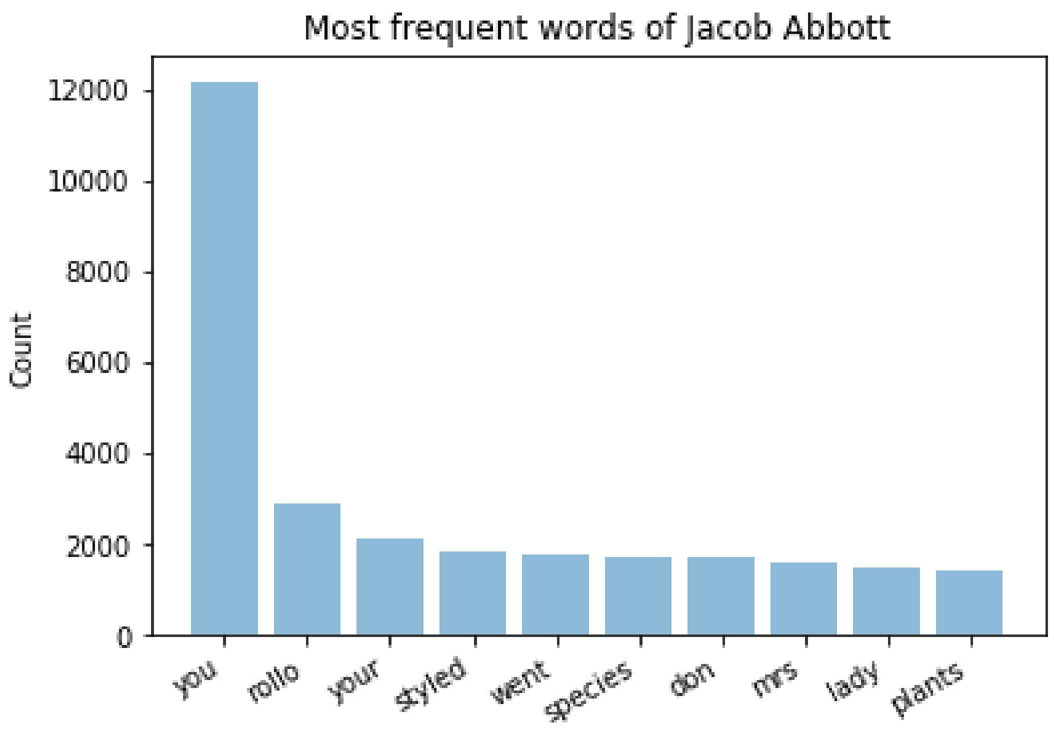
Dumitrescu Gabriel Horia

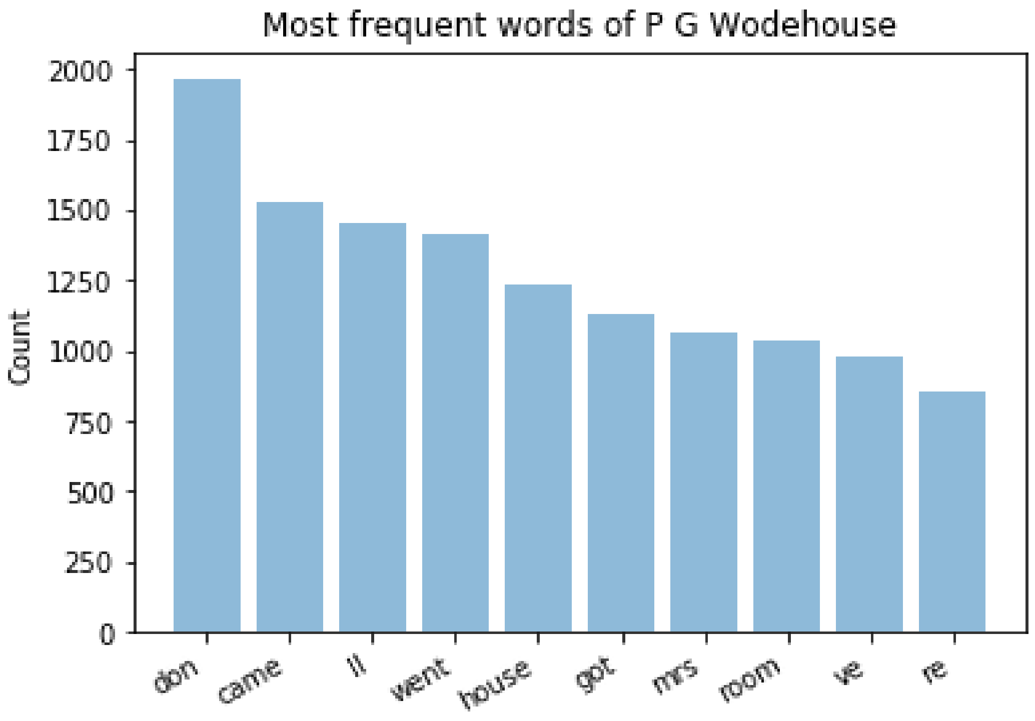
I)Data Visualization

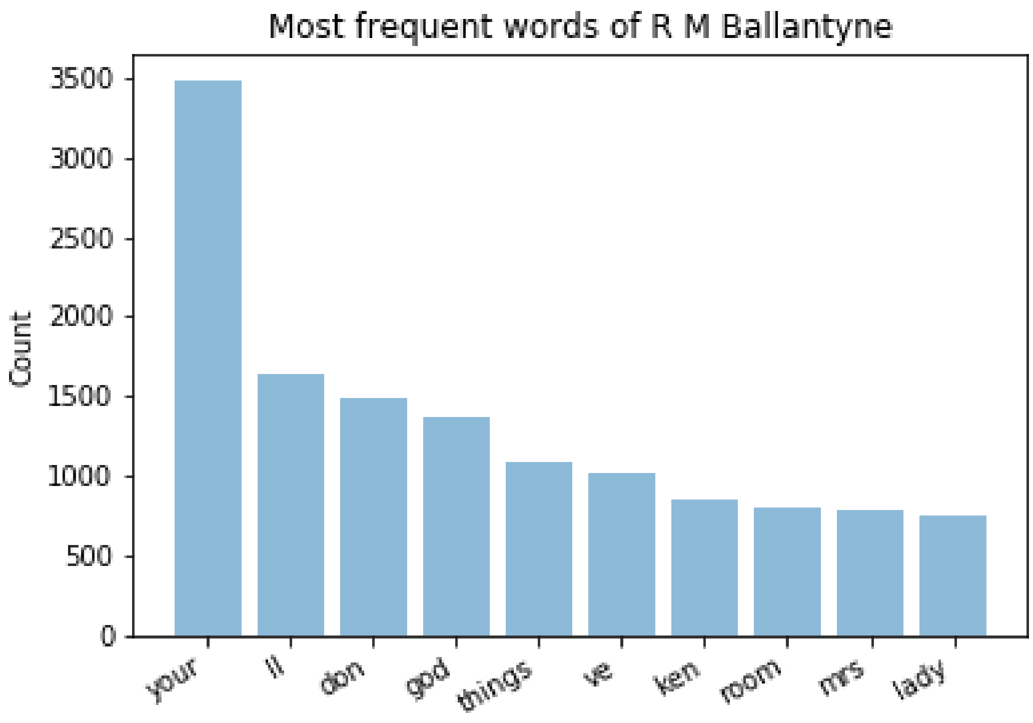
a) Common author words

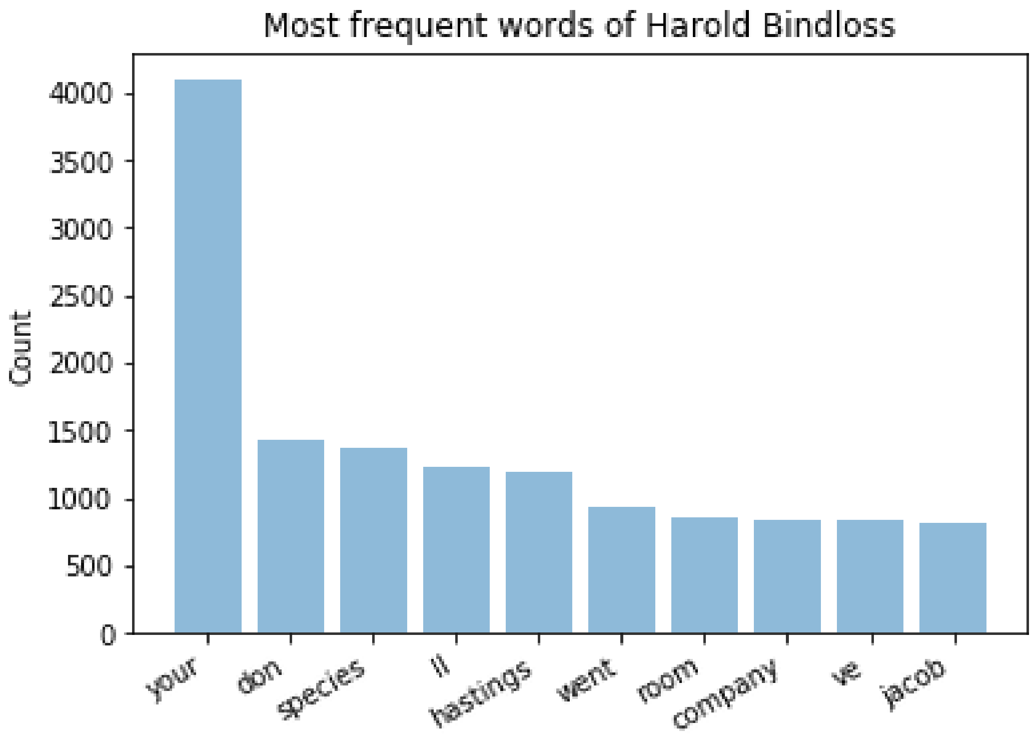


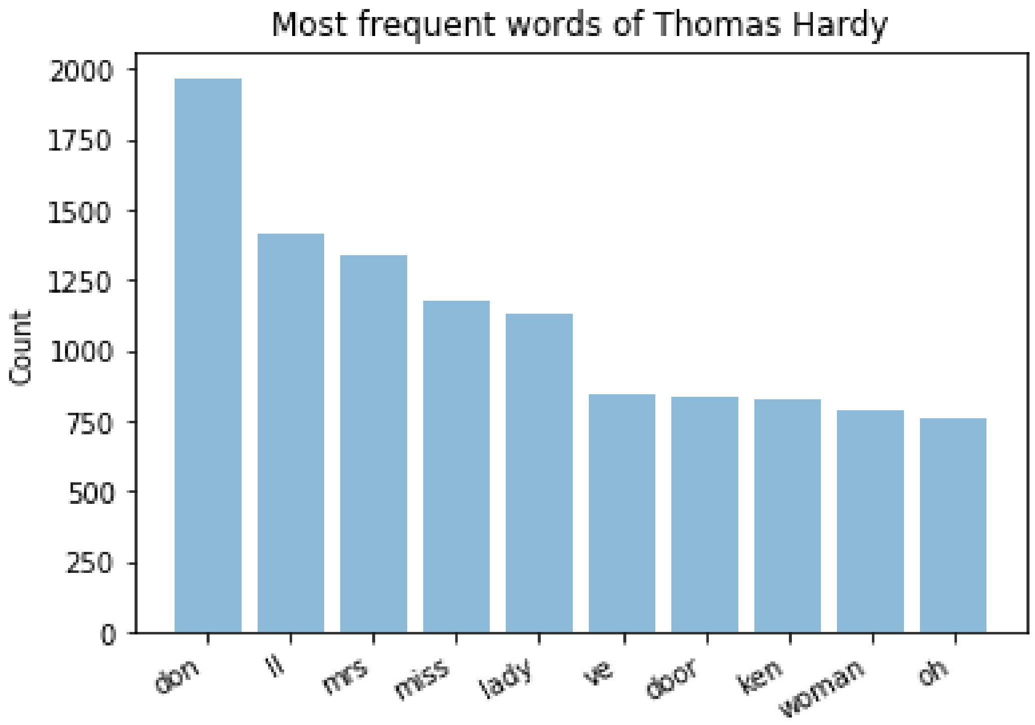


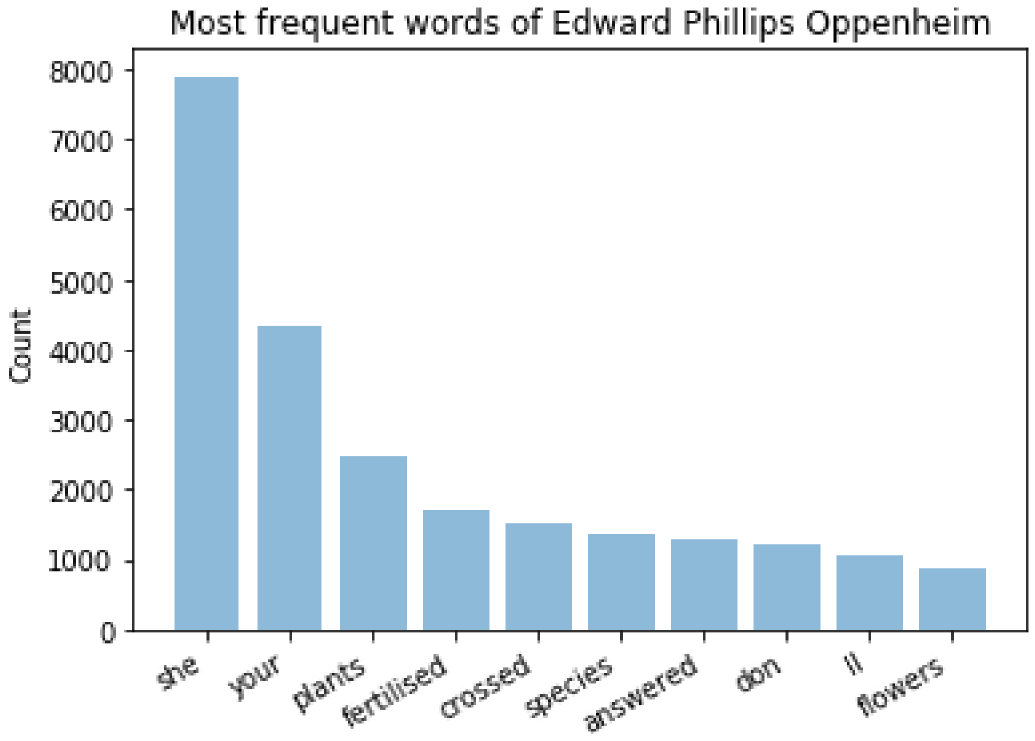

Most frequent words of Jacob Abbott
Count
0
2000
4000
6000
8000
10000
12000
you
rollo
your
styled
went
species
don
mrs
lady
plants

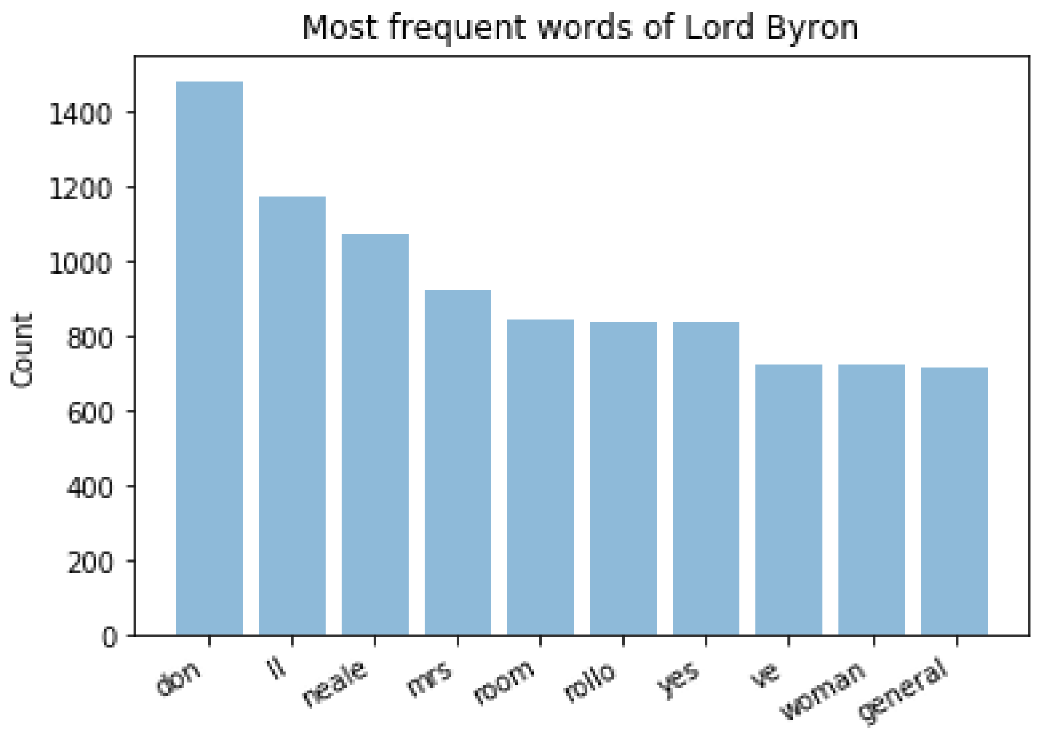

Most frequent words of P G Wodehouse
Count
0
250
500
750
1000
1250
1500
1750
2000
don
came
ll
went
house
got
mrs
room
ve
re

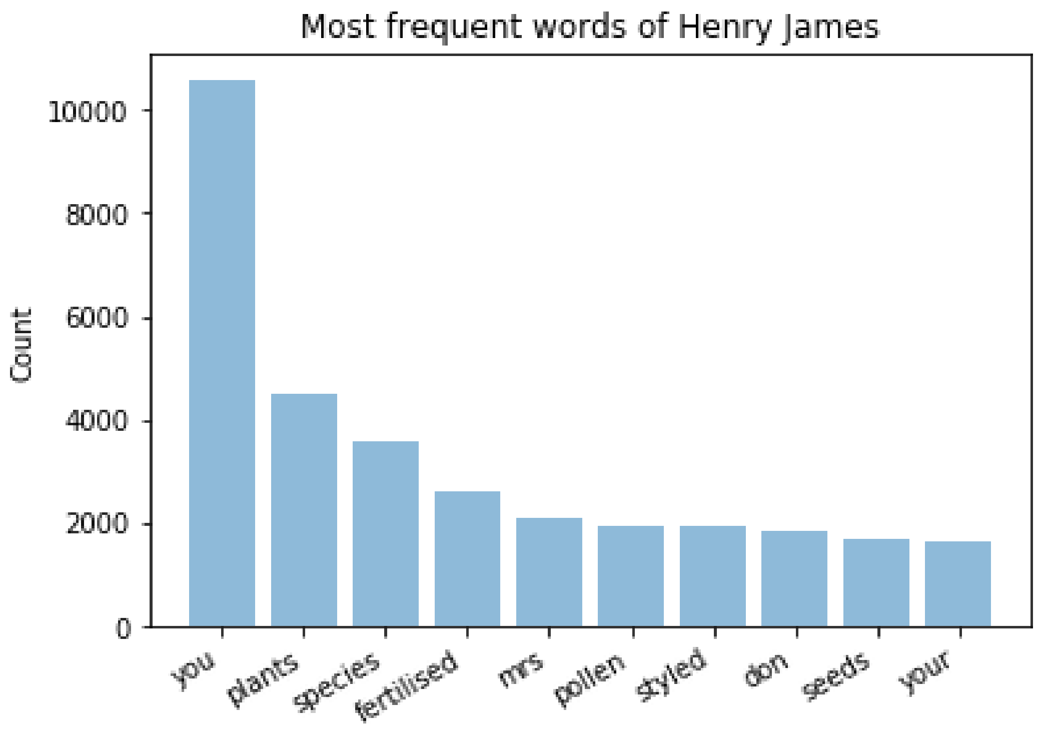

Most frequent words of R M Ballantyne
Count
3500
3000
2500
2000
1500
1000
500
0
your
ll
don
god
things
ve
ken
room
mrs
lady

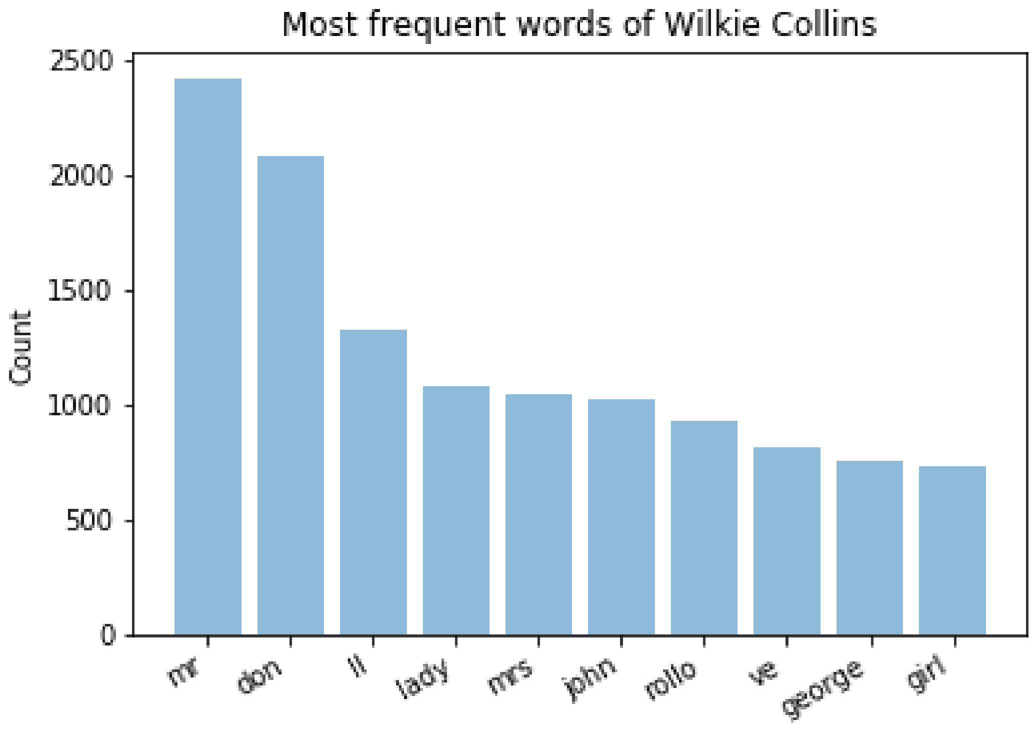

Most frequent words of Harold Bindloss
Count
4000
3500
3000
2500
2000
1500
1000
500
0
your
don
species
ll
hastings
went
room
company
ve
jacob

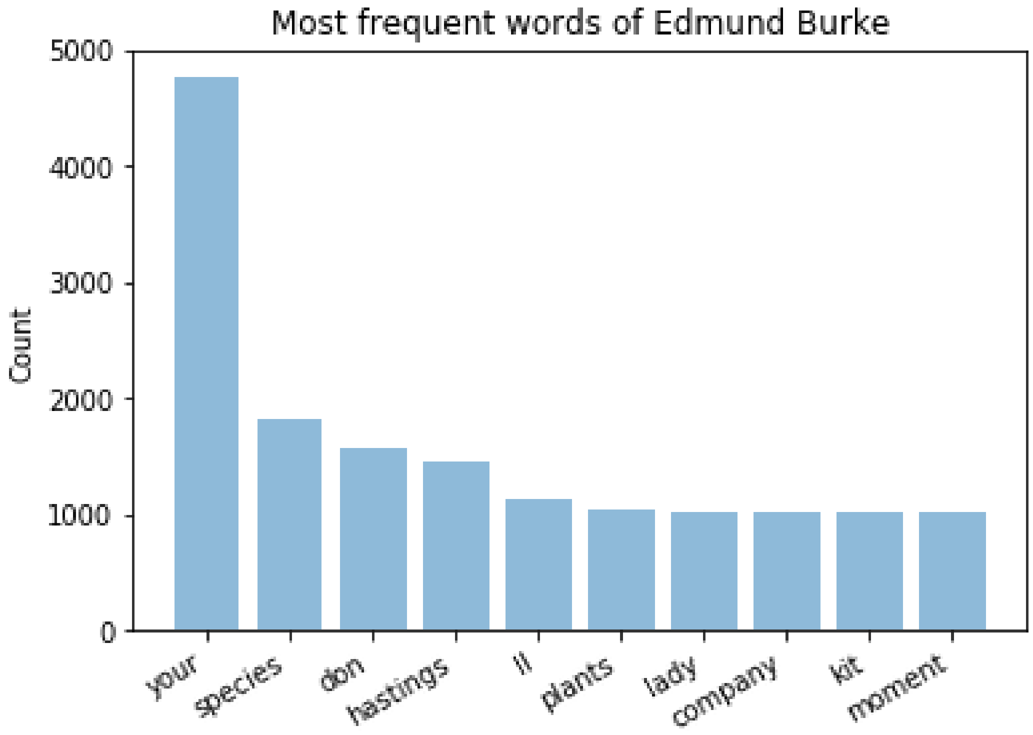

Most frequent words of Thomas Hardy
Count
0
250
500
750
1000
1250
1500
1750
2000
don
ll
mrs
miss
lady
ve
door
ken
woman
oh

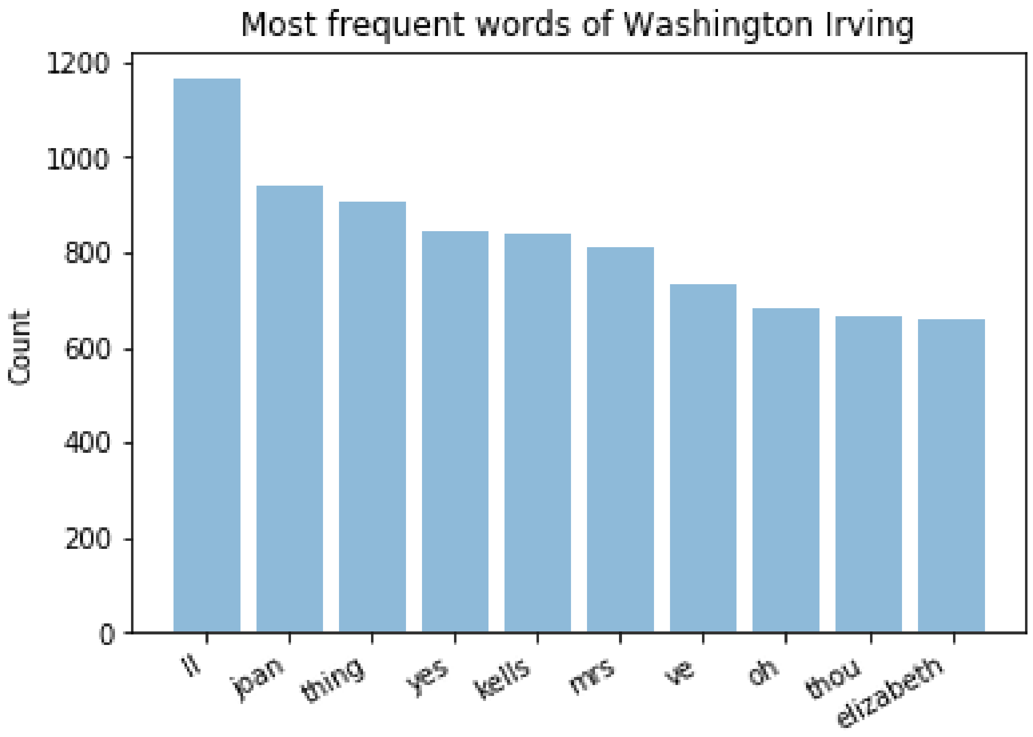

Most frequent words of Edward Phillips Oppenheim
Count
0
1000
2000
3000
4000
5000
6000
7000
8000
she
your
plants
fertilised
crossed
species
answered
don
ll
flowers

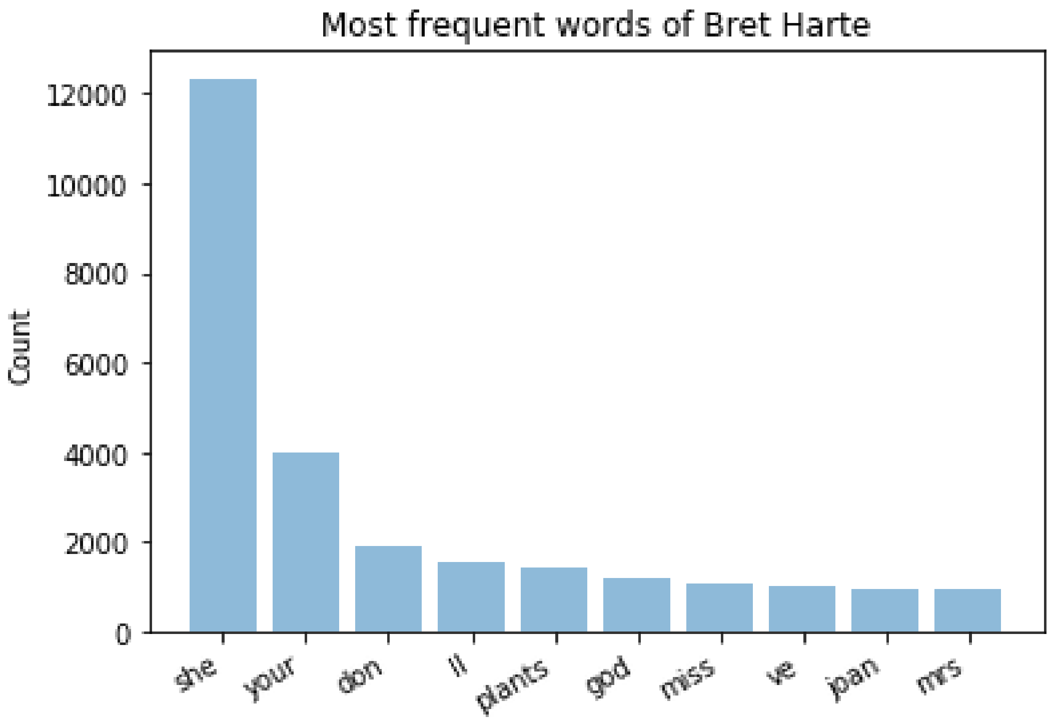

Most frequent words of Lord Byron
Count
0
200
400
600
800
1000
1200
1400
don
ll
neale
mrs
room
rollo
yes
ve
woman
general

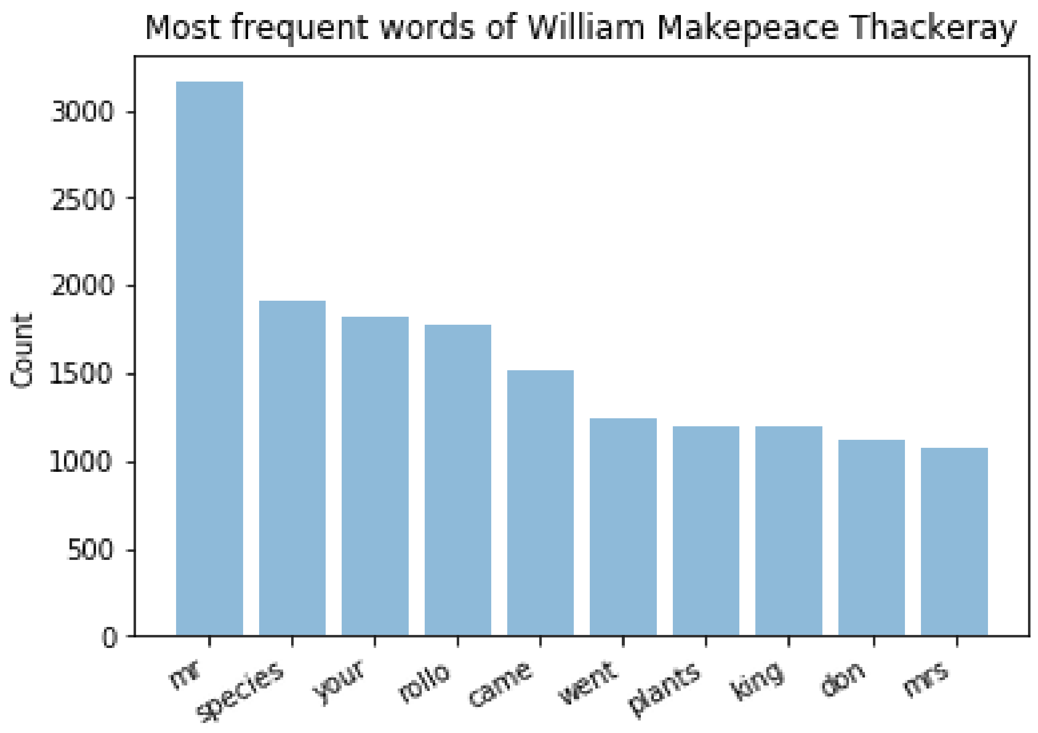

Most frequent words of Henry James
Count
0
2000
4000
6000
8000
10000
you
plants
species
fertilised
mrs
pollen
styled
don
seeds
your

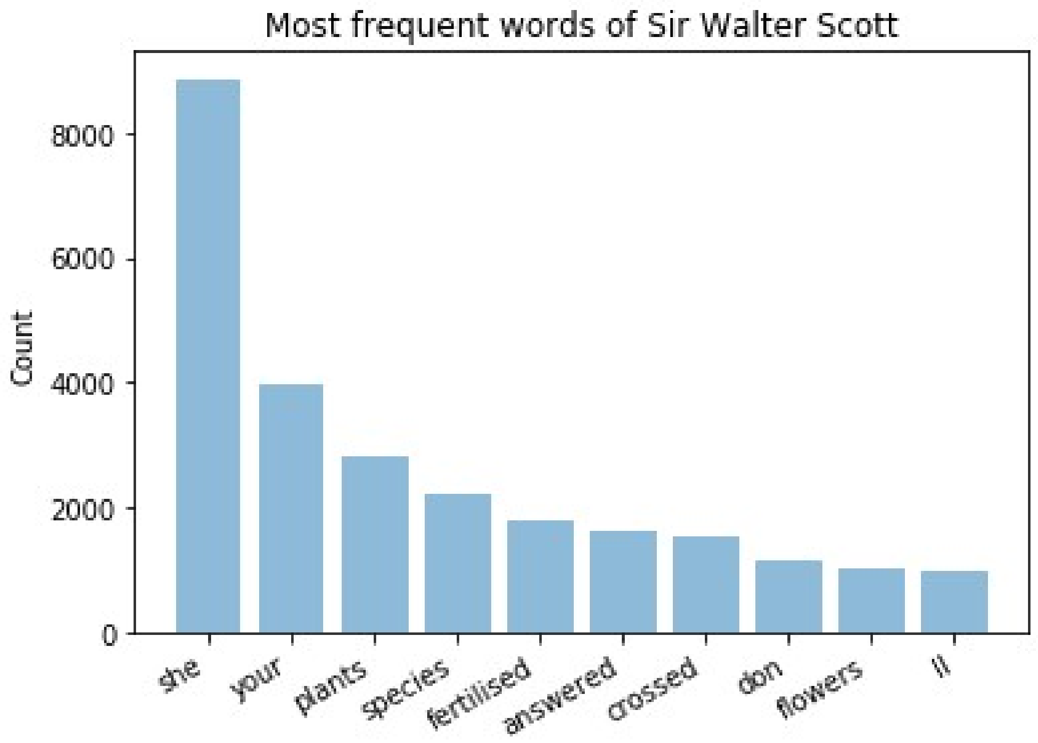

Most frequent words of Wilkie Collins
Count
2500
2000
1500
1000
500
0
mr
don
ll
lady
mrs
john
rollo
ve
george
girl

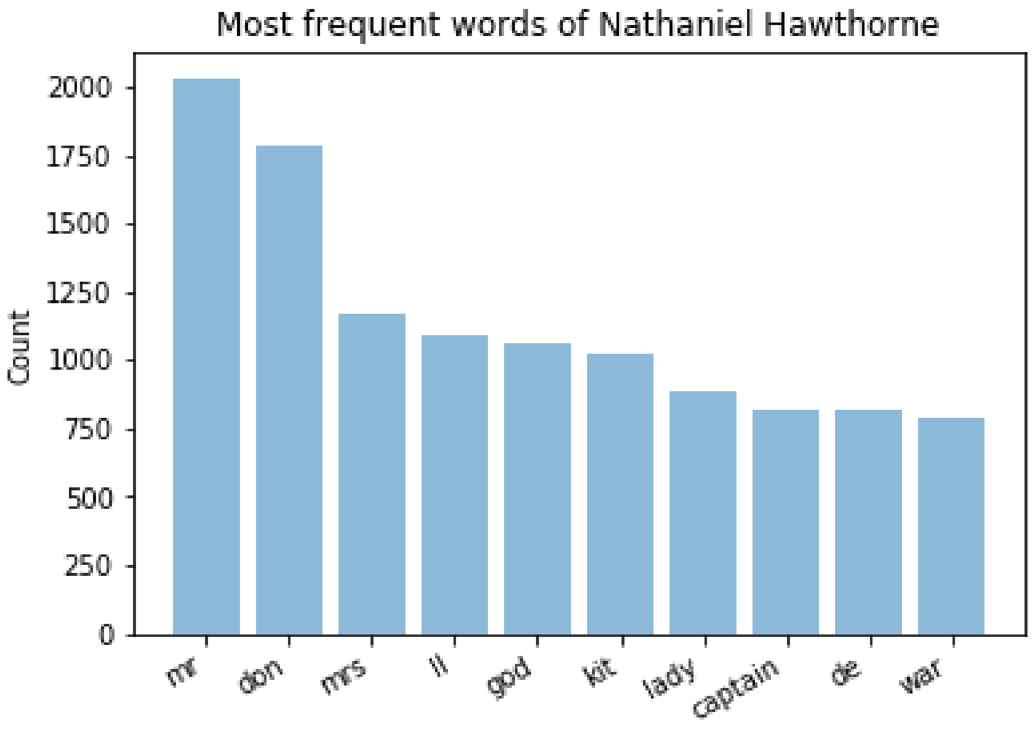

Most frequent words of Edmund Burke
Count
5000
4000
3000
2000
1000
0
your
species
don
hastings
ll
plants
lady
company
kit
moment


Most frequent words of Washington Irving
Count
0
200
400
600
800
1000
1200
ll
joan
thing
yes
kells
mrs
ve
oh
thou
elizabeth

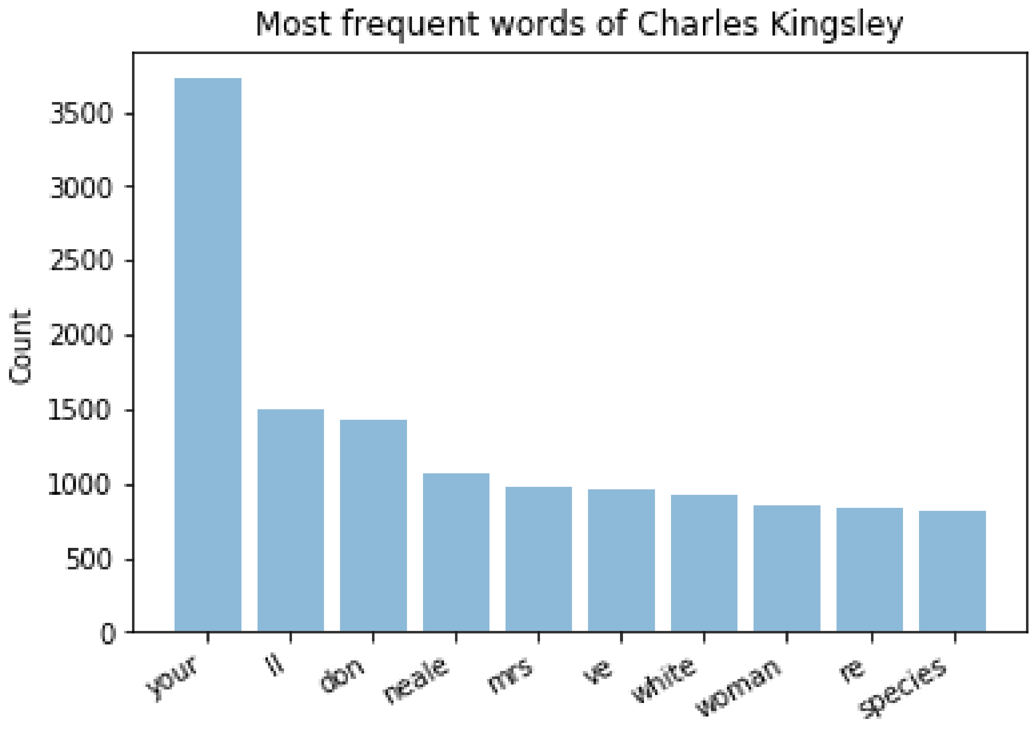

Most frequent words of Bret Harte
Count
0
2000
4000
6000
8000
10000
12000
she
your
don
ll
plants
god
miss
ve
joan
mrs


Most frequent words of William Makepeace Thackeray
Count
3000
2500
2000
1500
1000
500
0
mr
species
your
rollo
came
went
plants
king
don
mrs

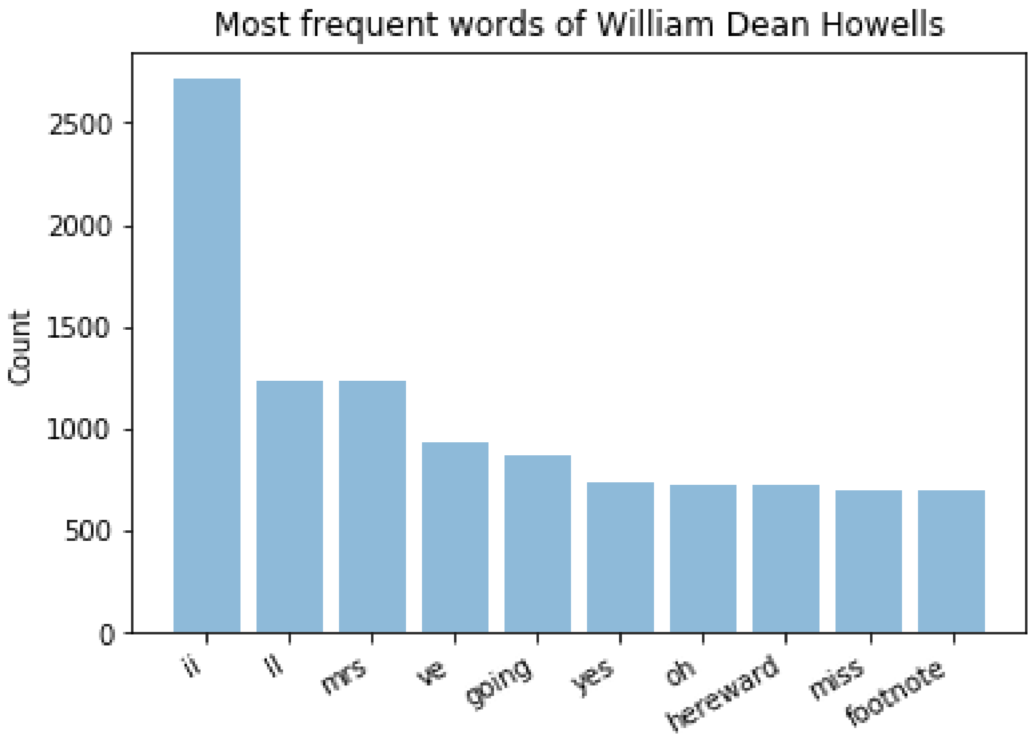

Most frequent words of Sir Walter Scott
Count
8000
6000
4000
2000
0
she
your
plants
species
fertilised
answered
crossed
don
flowers
ll



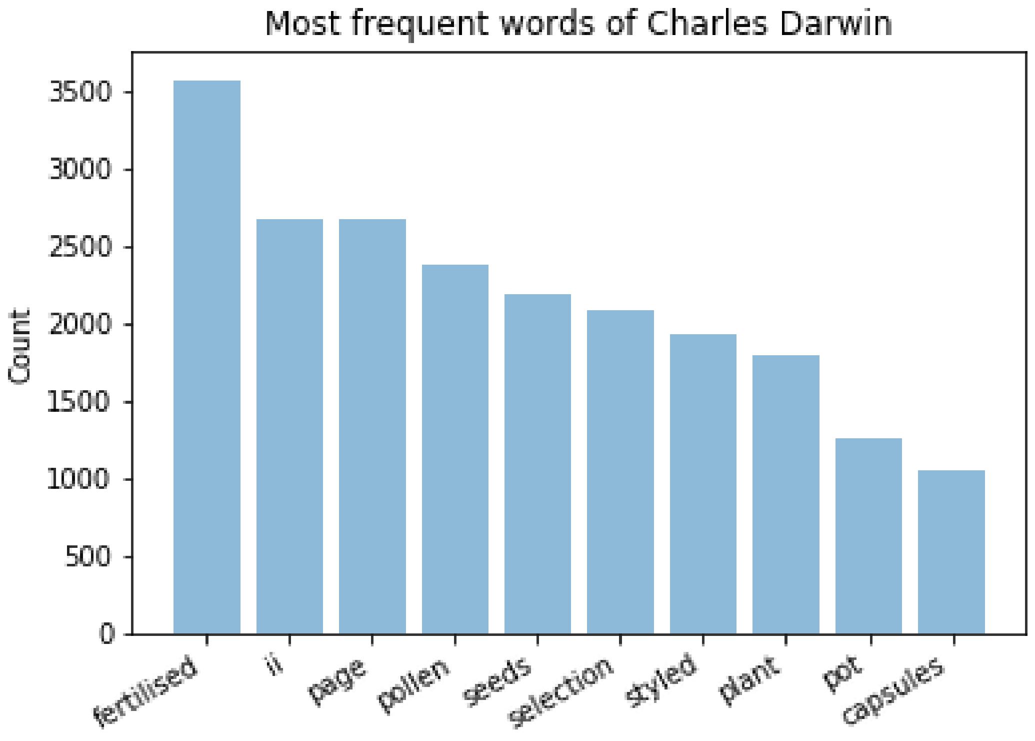
Most frequent words of Nathaniel Hawthorne
Count
0
250
500
750
1000
1250
1500
1750
2000
mr
don
mrs
ll
god
kit
lady
captain
de
war



Most frequent words of Charles Kingsley
Count
0
500
1000
1500
2000
2500
3000
3500
your
ll
don
neale
mrs
ve
white
woman
re
species

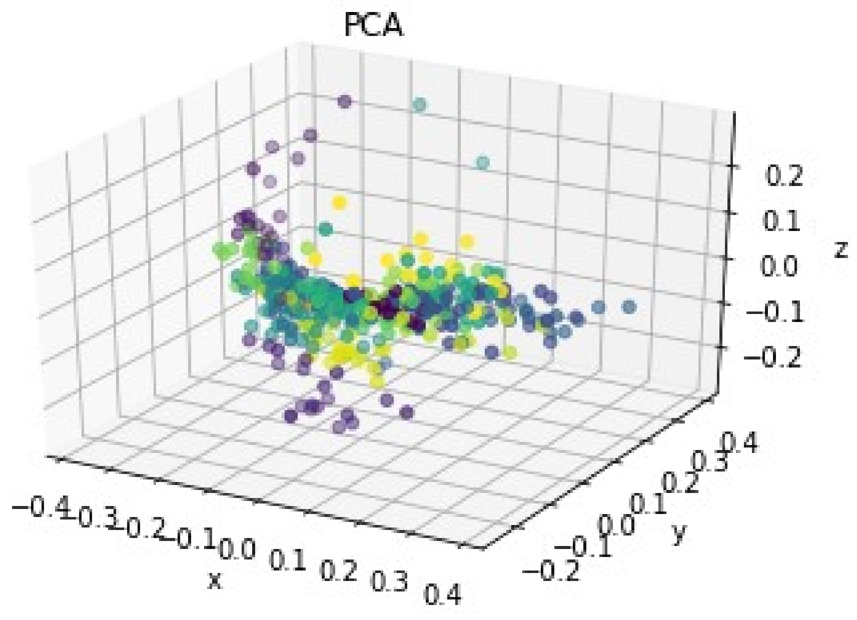


Most frequent words of William Dean Howells
Count
0
500
1000
1500
2000
2500
ii
ll
mrs
ve
going
yes
oh
hereward
miss
footnote

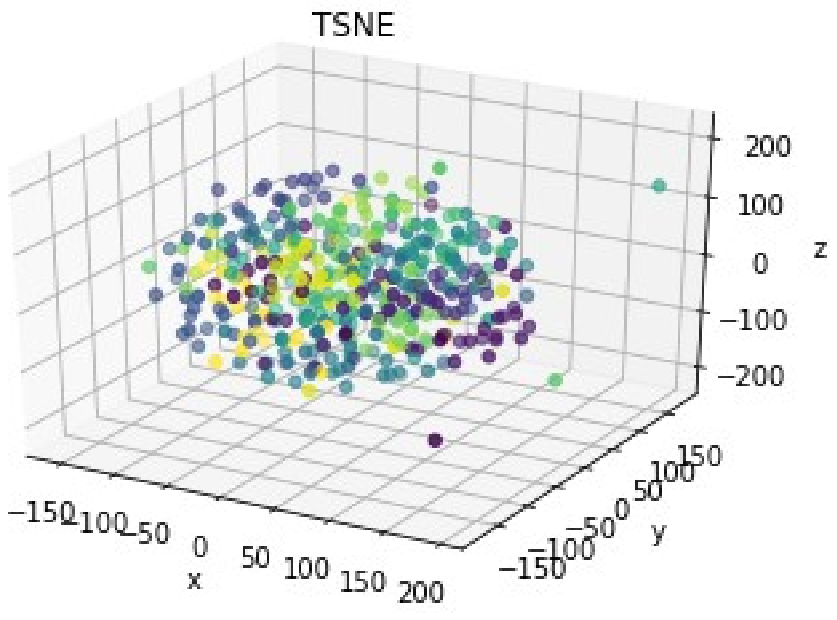


Most frequent words of Charles Darwin
Count
0
500
1000
1500
2000
2500
3000
3500
fertilised
ii
page
pollen
seeds
selection
styled
plant
pot
capsules

b)PCA Visualization(In jupyter notebook this can be rotated)



c)TSNE Visualization



II) Data preprocessing

a) I dropped the 'Unnamed: 0' column;

b) I created X, an array of 400 books and y, an array of their author names;

c) I encoded y with numeric values;

d) I represented the book texts in 3 ways:

-count vectorizer: it was useful for plotting the count of words ;

-tfid vectorizer: it was fast and yielded the best results for all of my models;

-word2vec: unfortunately this method was very expensive computationally. No matter how much I played with the parameters and tired to optimize the code, this process takes a lot of time and, most importantly, a lot of RAM. The only way I could get this to not receive a memory error was trimming it down so much, that it was too simple of a notation to represent the books (more on it’s performance later). In the future, I’ll use this for shorter documents, on computers with more ram and with stemming done beforehand. Another good solution is using sentence2vec or doc2vec instead;

e) I reduced the dimensionality of X to 400 (the number of samples), by using PCA. This increased performance, especially training time;

f) For classification, I split the data into : X_train, X_test, y_train, y_test with a ratio of 0.8 for training and 0.2 for testing. I used StratifiedShuffleSplit from sklearn to preserve the percentage of samples for each class;

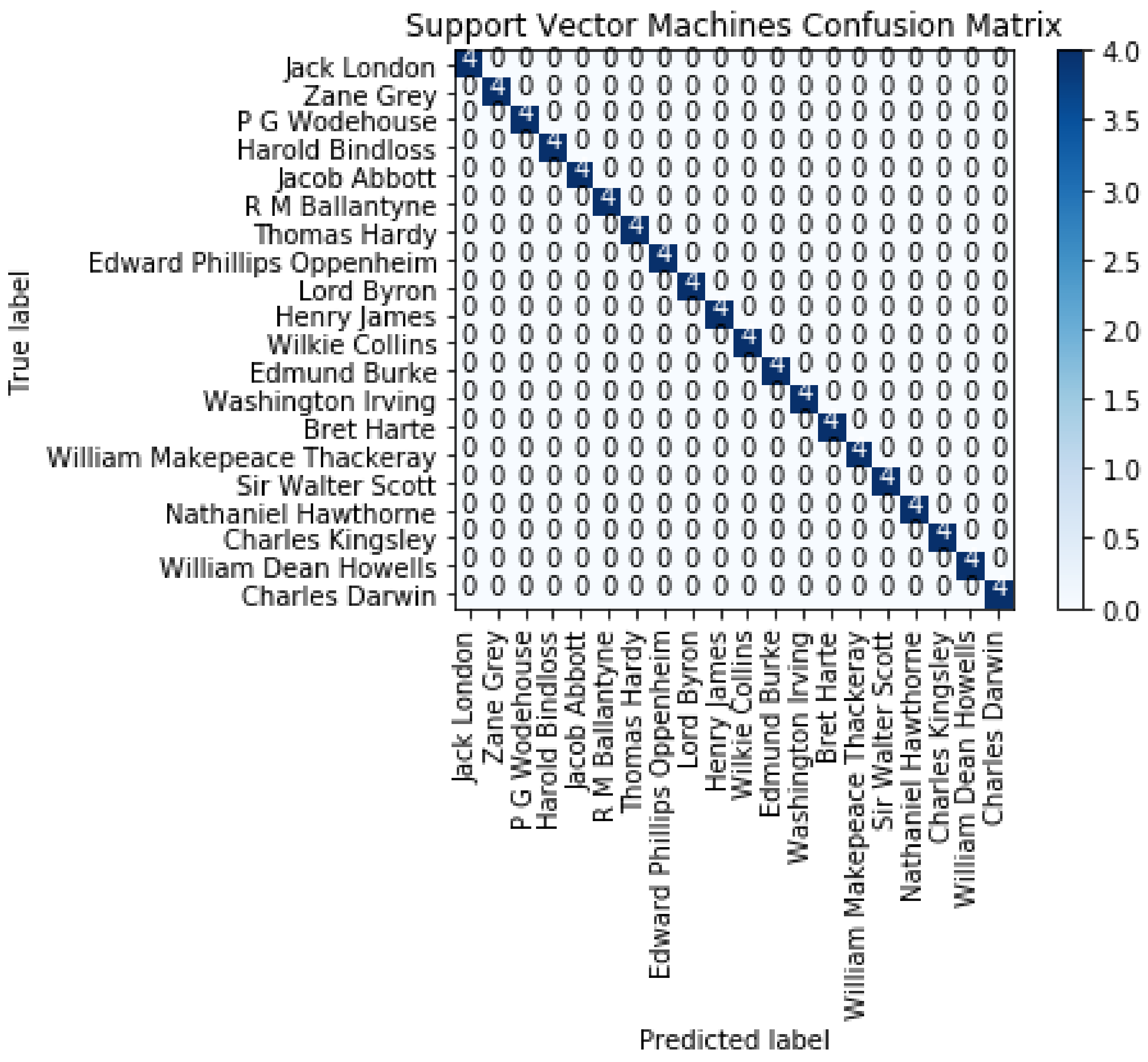
h) Tfid and count vectorizers removed accents and performed other character normalizations

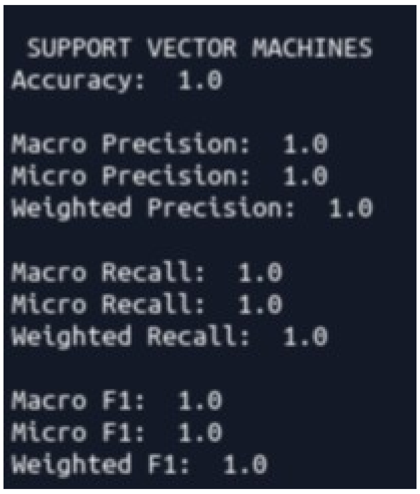
I) I stemmed all words. Warning: stemming takes a lot of time the first time.

III) Classification: Support Vector Machines

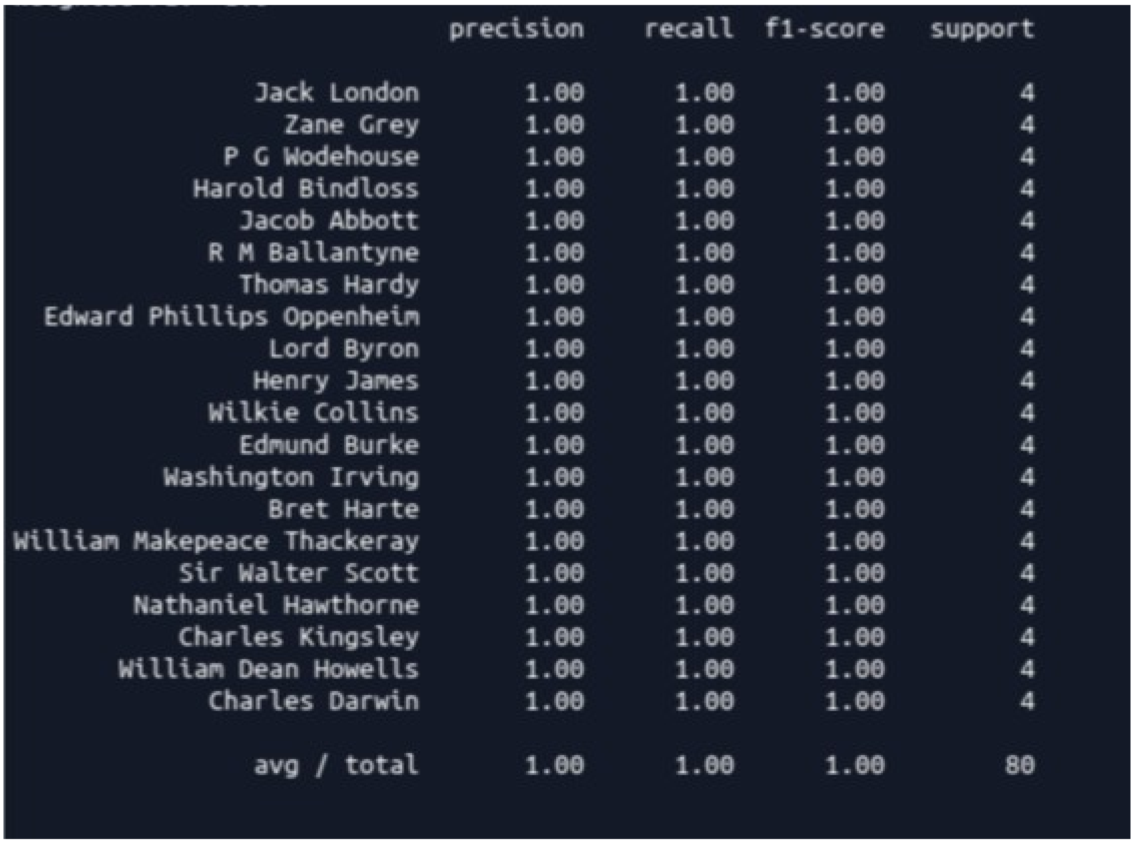
The default model had 85% accuracy. Extremely good for something so quick to write, especially compared to the models from the other homework.

Since the last time I used randomsearch to tune my parameters, this time I wanted to try gridsearch. After 2 failed nights, when my program didn’t fully compile, in my third attempt, it found, in 4 hours, parameters (penalty of 30 and ‘rbf’ kernel) with 100% accuracy. At first, I had to check my code because I didn’t believe it, I never witnessed this. I was ready to try bagging classifiers and other tricks, but, in the end, it would have been unnecessary. It’s not only perfect in accuracy, but in every other metric too.

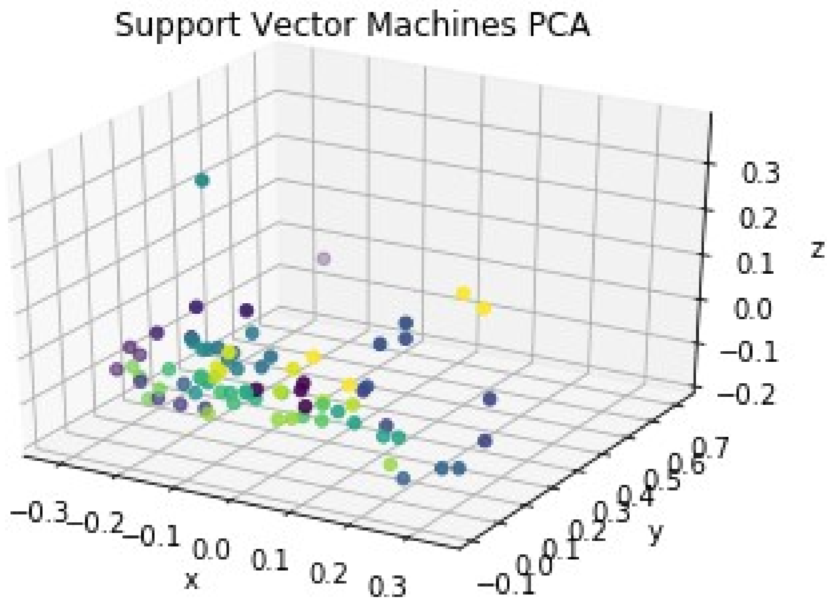


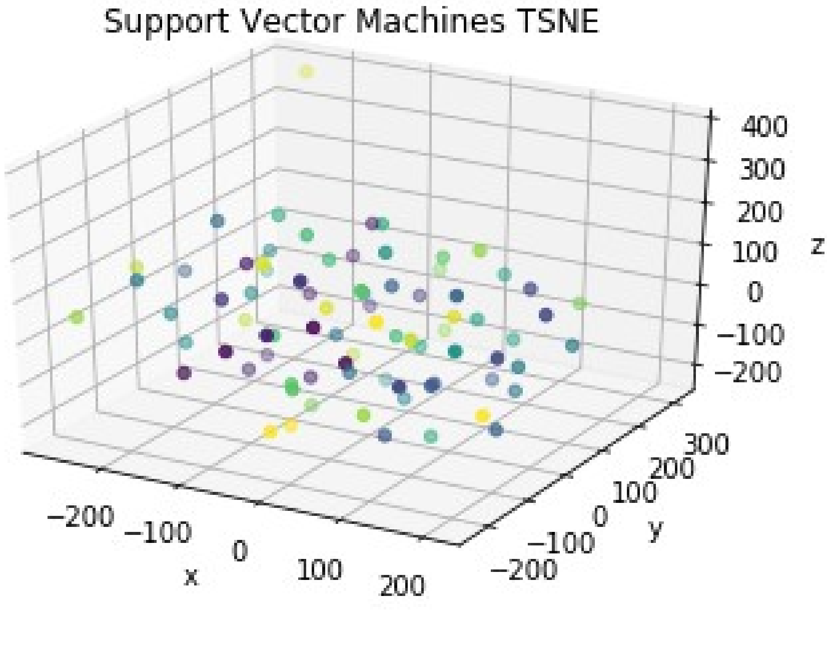


```
 SUPPORT VECTOR MACHINES
Accuracy:  1.0

Macro Precision:  1.0
Micro Precision:  1.0
Weighted Precision:  1.0

Macro Recall:  1.0
Micro Recall:  1.0
Weighted Recall:  1.0

Macro F1:  1.0
Micro F1:  1.0
Weighted F1:  1.0
```

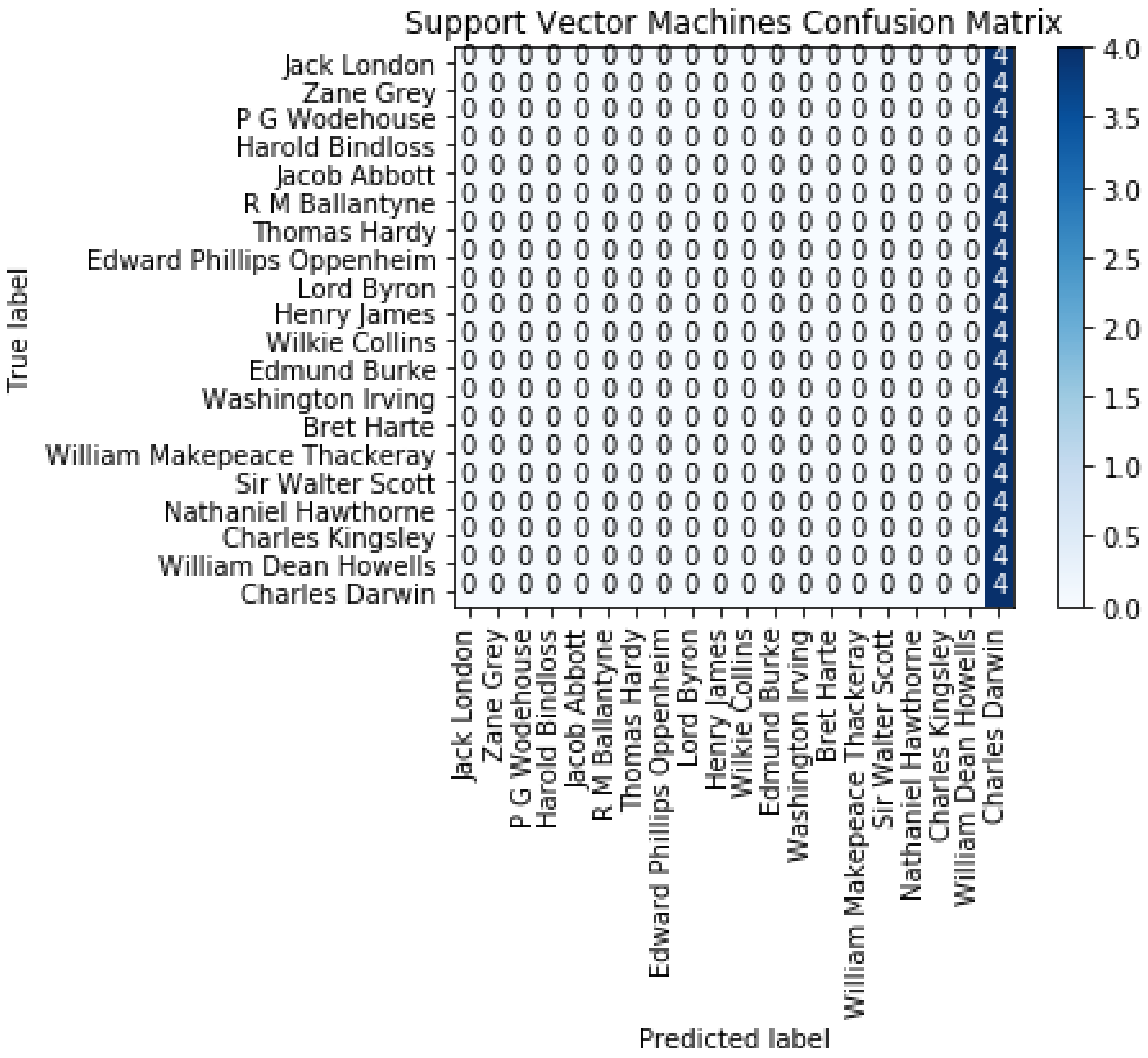


```
                             precision    recall  f1-score   support

                 Jack London       1.00      1.00      1.00         4
                   Zane Grey       1.00      1.00      1.00         4
               P G Wodehouse       1.00      1.00      1.00         4
             Harold Bindloss       1.00      1.00      1.00         4
                Jacob Abbott       1.00      1.00      1.00         4
              R M Ballantyne       1.00      1.00      1.00         4
                Thomas Hardy       1.00      1.00      1.00         4
    Edward Phillips Oppenheim       1.00      1.00      1.00         4
                  Lord Byron       1.00      1.00      1.00         4
                 Henry James       1.00      1.00      1.00         4
              Wilkie Collins       1.00      1.00      1.00         4
                Edmund Burke       1.00      1.00      1.00         4
           Washington Irving       1.00      1.00      1.00         4
                  Bret Harte       1.00      1.00      1.00         4
 William Makepeace Thackeray       1.00      1.00      1.00         4
            Sir Walter Scott       1.00      1.00      1.00         4
         Nathaniel Hawthorne       1.00      1.00      1.00         4
            Charles Kingsley       1.00      1.00      1.00         4
         William Dean Howells       1.00      1.00      1.00         4
              Charles Darwin       1.00      1.00      1.00         4

                 avg / total       1.00      1.00      1.00        80
```

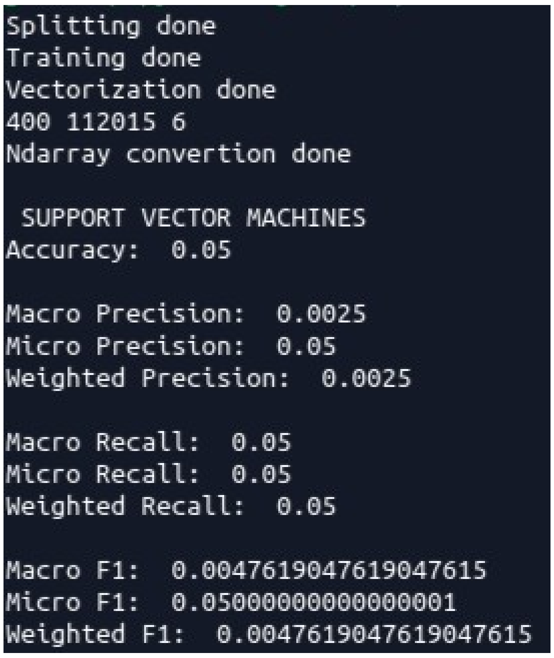

Support Vector Machines PCA
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
x
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
y
0.3
0.2
0.1
0.0
−0.1
−0.2
z


Support Vector Machines TSNE
−200
−100
0
100
200
x
−200
−100
0
100
200
300
y
400
300
200
100
0
−100
−200
z

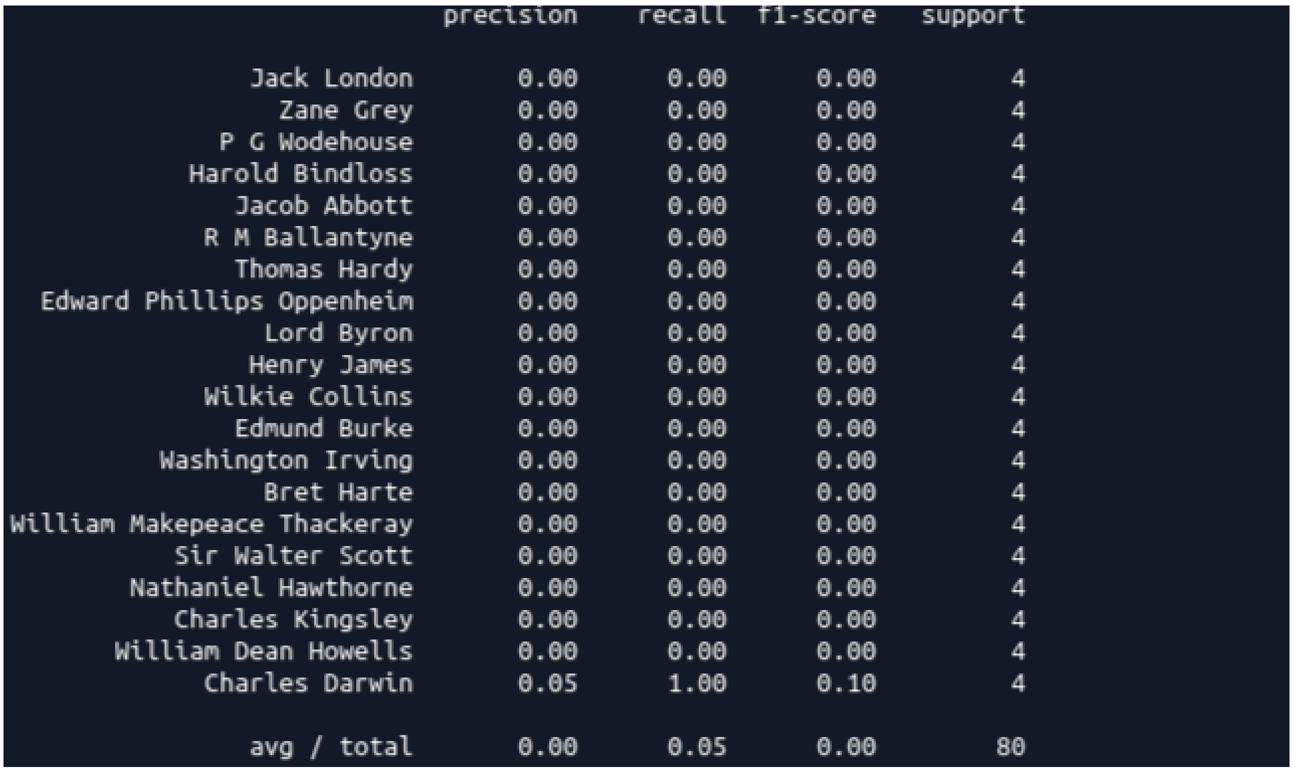
But, with the input from word2vec, even SVM had terrible performance. It only guessed Charles Darwin.



	Jack London	Zane Grey	P G Wodehouse	Harold Bindloss	Jacob Abbott	R M Ballantyne	Thomas Hardy	Edward Phillips Oppenheim	Lord Byron	Henry James	Wilkie Collins	Edmund Burke	Washington Irving	Bret Harte	William Makepeace Thackeray	Sir Walter Scott	Nathaniel Hawthorne	Charles Kingsley	William Dean Howells	Charles Darwin
Jack London	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
Zane Grey	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
P G Wodehouse	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
Harold Bindloss	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
Jacob Abbott	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
R M Ballantyne	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
Thomas Hardy	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
Edward Phillips Oppenheim	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
Lord Byron	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
Henry James	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
Wilkie Collins	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
Edmund Burke	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
Washington Irving	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
Bret Harte	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
William Makepeace Thackeray	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
Sir Walter Scott	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
Nathaniel Hawthorne	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
Charles Kingsley	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
William Dean Howells	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
Charles Darwin	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4



```
Splitting done
Training done
Vectorization done
400 112015 6
Ndarray convertion done

 SUPPORT VECTOR MACHINES
Accuracy:  0.05

Macro Precision:  0.0025
Micro Precision:  0.05
Weighted Precision:  0.0025

Macro Recall:  0.05
Micro Recall:  0.05
Weighted Recall:  0.05

Macro F1:  0.0047619047619047615
Micro F1:  0.05000000000000001
Weighted F1:  0.0047619047619047615
```

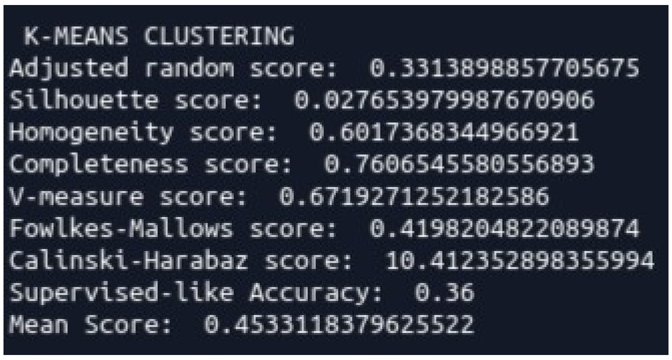


```
                           precision    recall  f1-score   support

              Jack London       0.00      0.00      0.00         4
                Zane Grey       0.00      0.00      0.00         4
            P G Wodehouse       0.00      0.00      0.00         4
          Harold Bindloss       0.00      0.00      0.00         4
             Jacob Abbott       0.00      0.00      0.00         4
           R M Ballantyne       0.00      0.00      0.00         4
             Thomas Hardy       0.00      0.00      0.00         4
  Edward Phillips Oppenheim       0.00      0.00      0.00         4
               Lord Byron       0.00      0.00      0.00         4
              Henry James       0.00      0.00      0.00         4
           Wilkie Collins       0.00      0.00      0.00         4
             Edmund Burke       0.00      0.00      0.00         4
        Washington Irving       0.00      0.00      0.00         4
               Bret Harte       0.00      0.00      0.00         4
William Makepeace Thackeray       0.00      0.00      0.00         4
         Sir Walter Scott       0.00      0.00      0.00         4
      Nathaniel Hawthorne       0.00      0.00      0.00         4
         Charles Kingsley       0.00      0.00      0.00         4
      William Dean Howells       0.00      0.00      0.00         4
           Charles Darwin       0.05      1.00      0.10         4

              avg / total       0.00      0.05      0.00        80
```

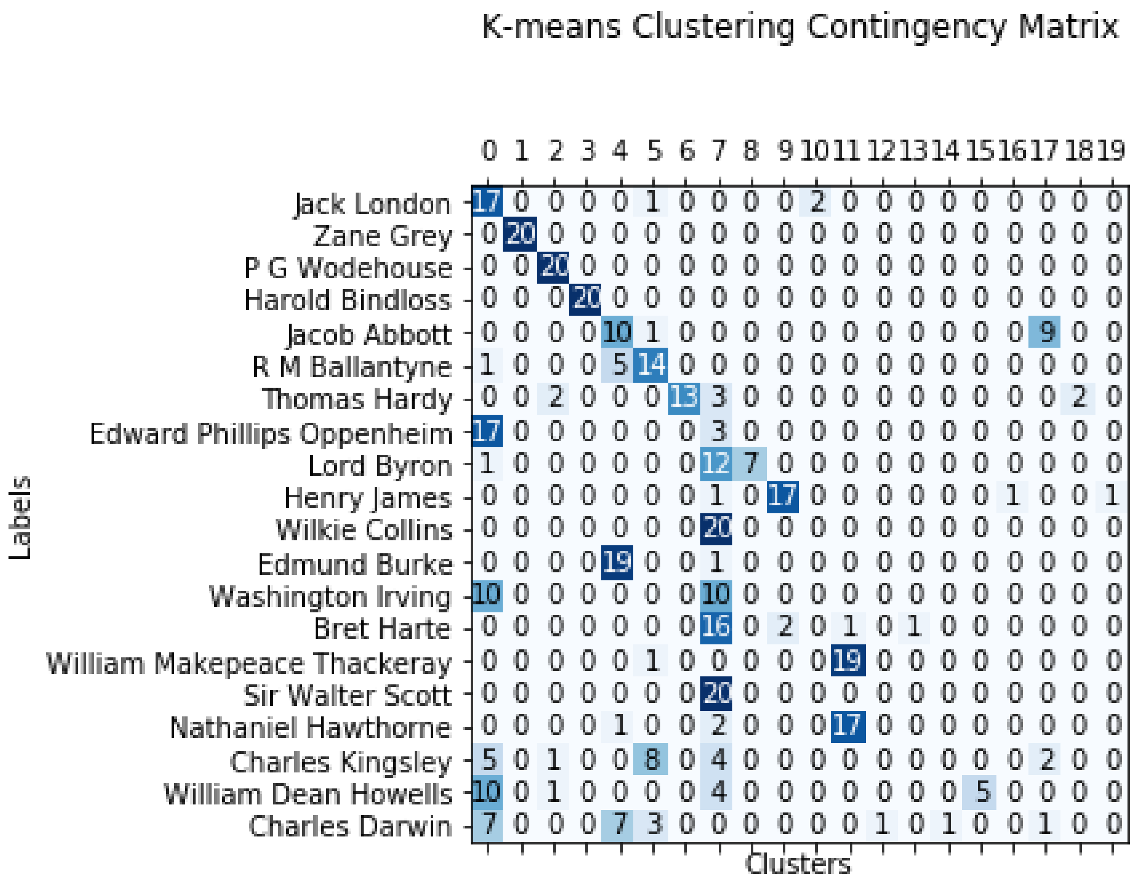
IV) Clustering

a) K-MEANS Clustering

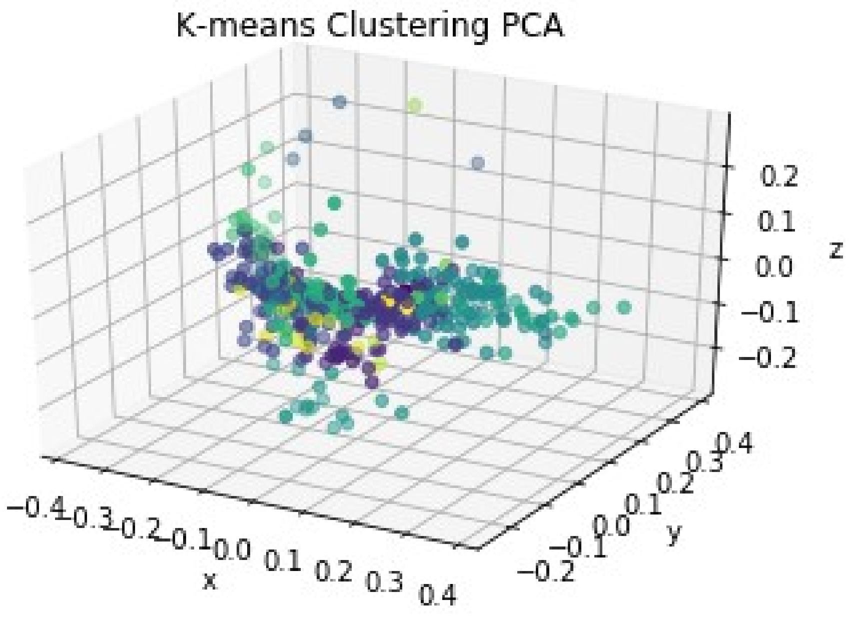
This clustering method was okay. I played around with parameters in search of a better clustering. I searched all of the scores I could find and even invented one myself and calculated the mean of the scores. I experimentally tried grid searching based on that score, but I was unsuccessful. The visualizations look a bit different to the real labels. I plotted a contingency matrix to see what clusters it considers.

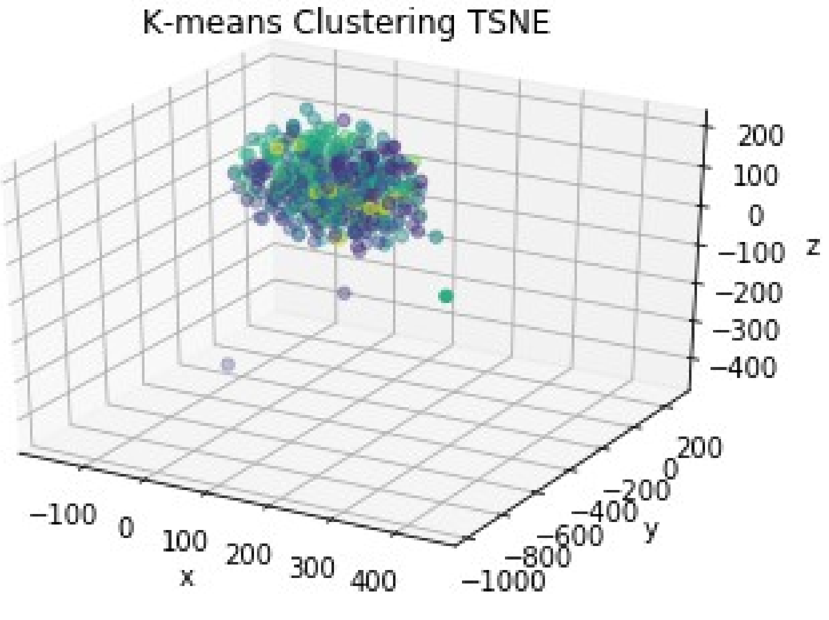


```
 K-MEANS CLUSTERING
Adjusted random score:  0.3313898857705675
Silhouette score:  0.027653979987670906
Homogeneity score:  0.6017368344966921
Completeness score:  0.7606545580556893
V-measure score:  0.6719271252182586
Fowlkes-Mallows score:  0.4198204822089874
Calinski-Harabaz score:  10.412352898355994
Supervised-like Accuracy:  0.36
Mean Score:  0.4533118379625522
```

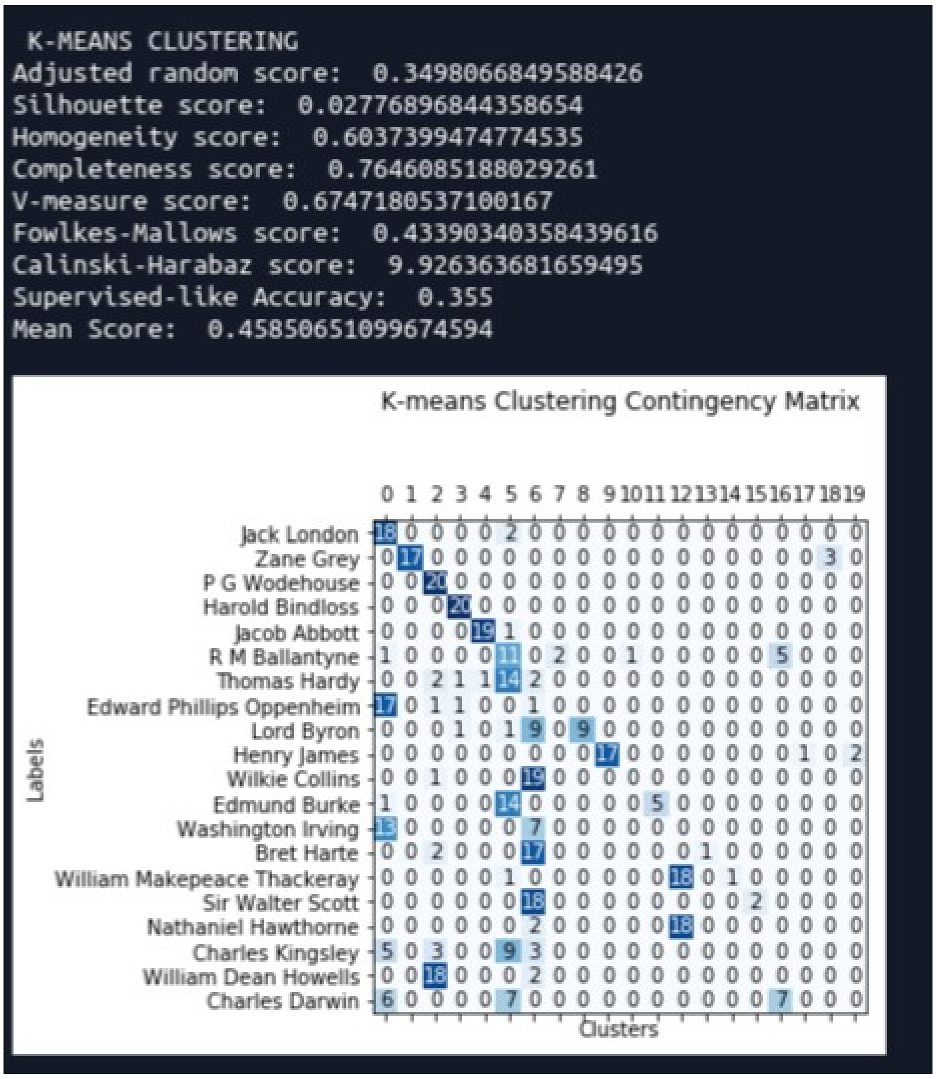


	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Jack London	17	0	0	0	0	1	0	0	0	0	2	0	0	0	0	0	0	0	0	0
Zane Grey	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P G Wodehouse	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Harold Bindloss	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Jacob Abbott	0	0	0	0	10	1	0	0	0	0	0	0	0	0	0	0	0	9	0	0
R M Ballantyne	1	0	0	0	5	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Thomas Hardy	0	0	2	0	0	0	13	3	0	0	0	0	0	0	0	0	0	0	2	0
Edward Phillips Oppenheim	17	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0
Lord Byron	1	0	0	0	0	0	0	12	7	0	0	0	0	0	0	0	0	0	0	0
Henry James	0	0	0	0	0	0	0	1	0	17	0	0	0	0	0	0	1	0	0	1
Wilkie Collins	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0
Edmund Burke	0	0	0	0	19	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Washington Irving	10	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0
Bret Harte	0	0	0	0	0	0	0	16	0	2	0	1	0	1	0	0	0	0	0	0
William Makepeace Thackeray	0	0	0	0	0	1	0	0	0	0	0	19	0	0	0	0	0	0	0	0
Sir Walter Scott	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0
Nathaniel Hawthorne	0	0	0	0	1	0	0	2	0	0	0	17	0	0	0	0	0	0	0	0
Charles Kingsley	5	0	1	0	0	8	0	4	0	0	0	0	0	0	0	0	0	2	0	0
William Dean Howells	10	0	1	0	0	0	0	4	0	0	0	0	0	0	0	5	0	0	0	0
Charles Darwin	7	0	0	0	7	3	0	0	0	0	0	0	1	0	1	0	0	1	0	0


K-means Clustering PCA
0.2
0.1
0.0
−0.1
−0.2
z
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
x
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
y


K-means Clustering TSNE
200
100
0
−100
−200
−300
−400
z
−100
0
100
200
300
400
x
200
0
−200
−400
−600
−800
−1000
y

The scores are slightly lower with a ‘random’ init instead of the default ‘kmeans++’ and the clusters are a bit more uneven,

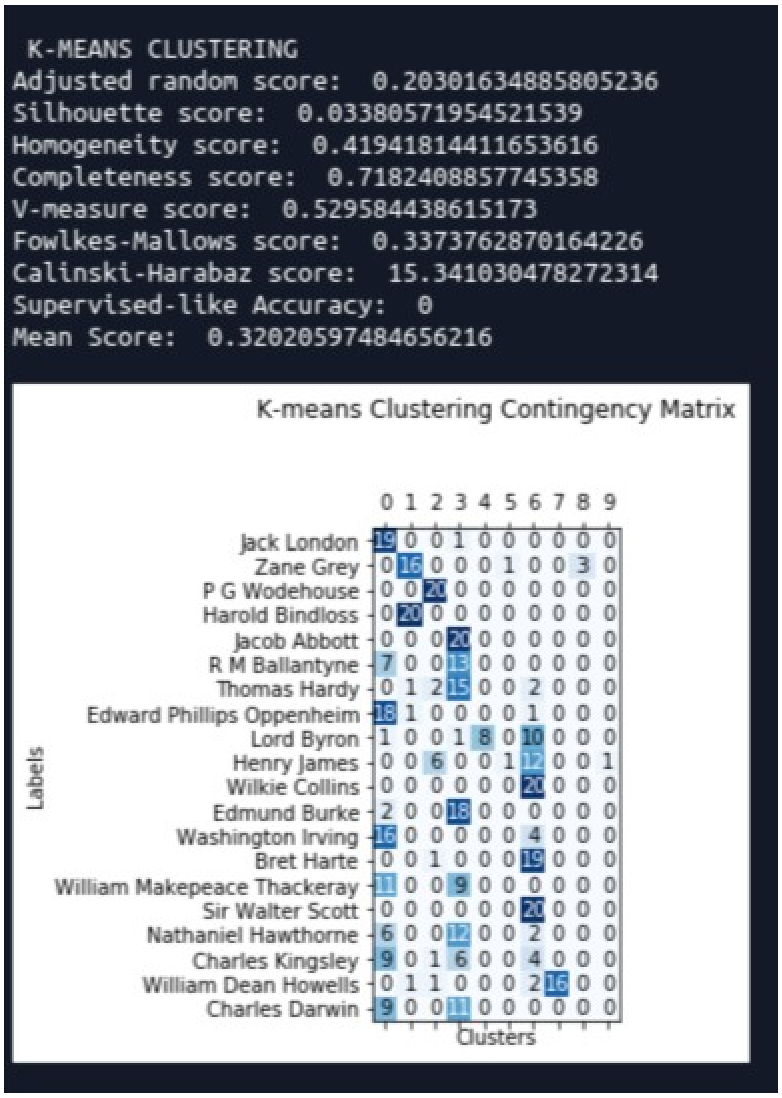


```
 K-MEANS CLUSTERING
Adjusted random score:  0.3498066849588426
Silhouette score:  0.02776896844358654
Homogeneity score:  0.6037399474774535
Completeness score:  0.7646085188029261
V-measure score:  0.6747180537100167
Fowlkes-Mallows score:  0.43390340358439616
Calinski-Harabaz score:  9.926363681659495
Supervised-like Accuracy:  0.355
Mean Score:  0.45850651099674594
```

K-means Clustering Contingency Matrix

Labels \ Clusters	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Jack London	18	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Zane Grey	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0
P G Wodehouse	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Harold Bindloss	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Jacob Abbott	0	0	0	0	19	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R M Ballantyne	1	0	0	0	0	11	0	2	0	0	1	0	0	0	0	0	5	0	0	0
Thomas Hardy	0	0	2	1	1	14	2	0	0	0	0	0	0	0	0	0	0	0	0	0
Edward Phillips Oppenheim	17	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Lord Byron	0	0	0	1	0	1	9	0	9	0	0	0	0	0	0	0	0	0	0	0
Henry James	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0	0	0	1	0	2
Wilkie Collins	0	0	1	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0
Edmund Burke	1	0	0	0	0	14	0	0	0	0	0	5	0	0	0	0	0	0	0	0
Washington Irving	13	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0
Bret Harte	0	0	2	0	0	0	17	0	0	0	0	0	0	1	0	0	0	0	0	0
William Makepeace Thackeray	0	0	0	0	0	1	0	0	0	0	0	0	18	0	1	0	0	0	0	0
Sir Walter Scott	0	0	0	0	0	0	18	0	0	0	0	0	0	0	0	2	0	0	0	0
Nathaniel Hawthorne	0	0	0	0	0	0	2	0	0	0	0	0	18	0	0	0	0	0	0	0
Charles Kingsley	5	0	3	0	0	9	3	0	0	0	0	0	0	0	0	0	0	0	0	0
William Dean Howells	0	0	18	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
Charles Darwin	6	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	7	0	0	0

Also, these are the clusters it considers for n_components=10:



```
 K-MEANS CLUSTERING
Adjusted random score:  0.20301634885805236
Silhouette score:  0.03380571954521539
Homogeneity score:  0.41941814411653616
Completeness score:  0.7182408857745358
V-measure score:  0.529584438615173
Fowlkes-Mallows score:  0.3373762870164226
Calinski-Harabaz score:  15.341030478272314
Supervised-like Accuracy:  0
Mean Score:  0.32020597484656216
```

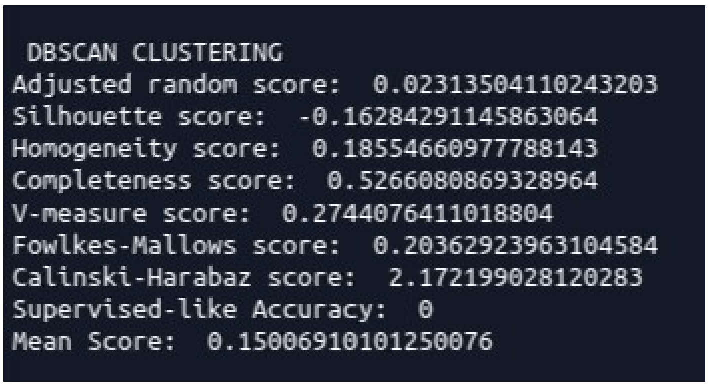
K-means Clustering Contingency Matrix

Labels \ Clusters	0	1	2	3	4	5	6	7	8	9
Jack London	19	0	0	1	0	0	0	0	0	0
Zane Grey	0	16	0	0	0	1	0	0	3	0
P G Wodehouse	0	0	20	0	0	0	0	0	0	0
Harold Bindloss	0	20	0	0	0	0	0	0	0	0
Jacob Abbott	0	0	0	20	0	0	0	0	0	0
R M Ballantyne	7	0	0	13	0	0	0	0	0	0
Thomas Hardy	0	1	2	15	0	0	2	0	0	0
Edward Phillips Oppenheim	18	1	0	0	0	0	1	0	0	0
Lord Byron	1	0	0	1	8	0	10	0	0	0
Henry James	0	0	6	0	0	1	12	0	0	1
Wilkie Collins	0	0	0	0	0	0	20	0	0	0
Edmund Burke	2	0	0	18	0	0	0	0	0	0
Washington Irving	16	0	0	0	0	0	4	0	0	0
Bret Harte	0	0	1	0	0	0	19	0	0	0
William Makepeace Thackeray	11	0	0	9	0	0	0	0	0	0
Sir Walter Scott	0	0	0	0	0	0	20	0	0	0
Nathaniel Hawthorne	6	0	0	12	0	0	2	0	0	0
Charles Kingsley	9	0	1	6	0	0	4	0	0	0
William Dean Howells	0	1	1	0	0	0	2	16	0	0
Charles Darwin	9	0	0	11	0	0	0	0	0	0

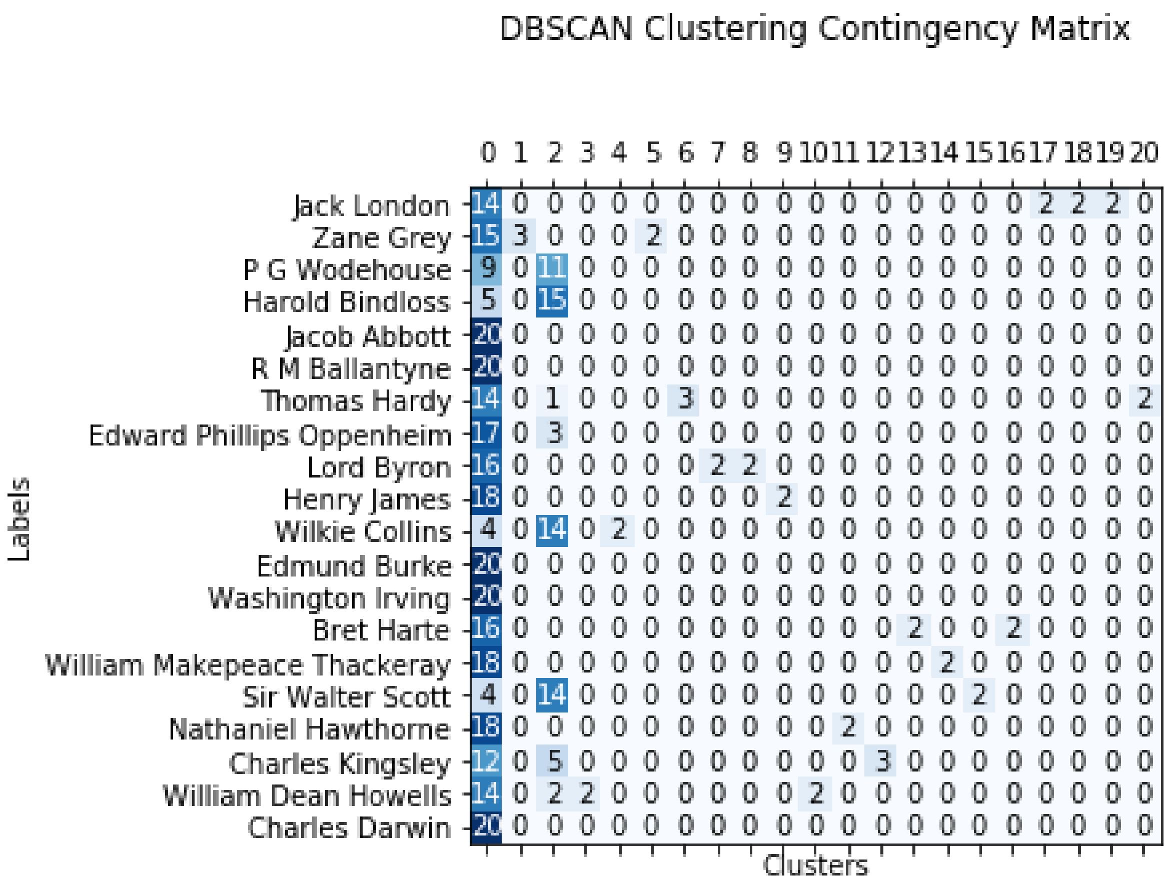
This clustering was very uneven and had low scores.

b)DBSCAN

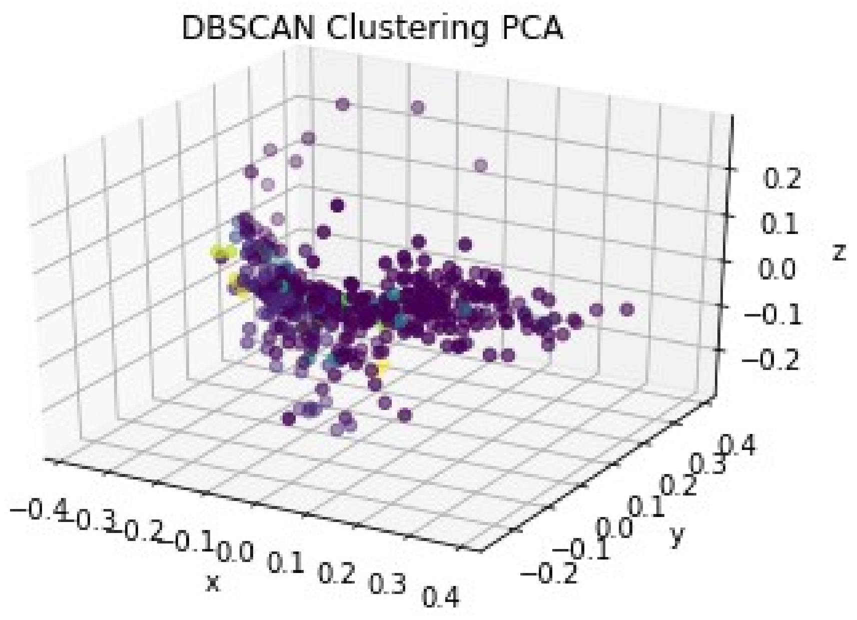
This was my worst clustering. For this task, DBSCAN is not reliable. After an endless parameter search, I’ve found eps=0.22 and min_samples=2 to at least cluster the data into 21 clusters. It’s scores are the worst and clusterings are very uneven. It picks a little difference in writing styles (some authors have 2 works making up whole clusters). Visualizations also show data dominated by only one color.

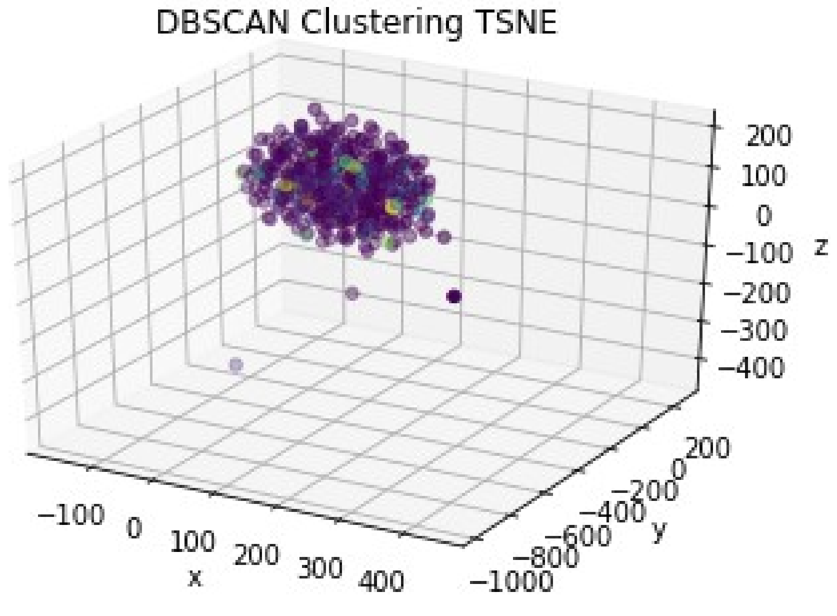


```
 DBSCAN CLUSTERING
Adjusted random score:  0.02313504110243203
Silhouette score:  -0.16284291145863064
Homogeneity score:  0.18554660977788143
Completeness score:  0.5266080869328964
V-measure score:  0.2744076411018804
Fowlkes-Mallows score:  0.20362923963104584
Calinski-Harabaz score:  2.172199028120283
Supervised-like Accuracy:  0
Mean Score:  0.15006910101250076
```

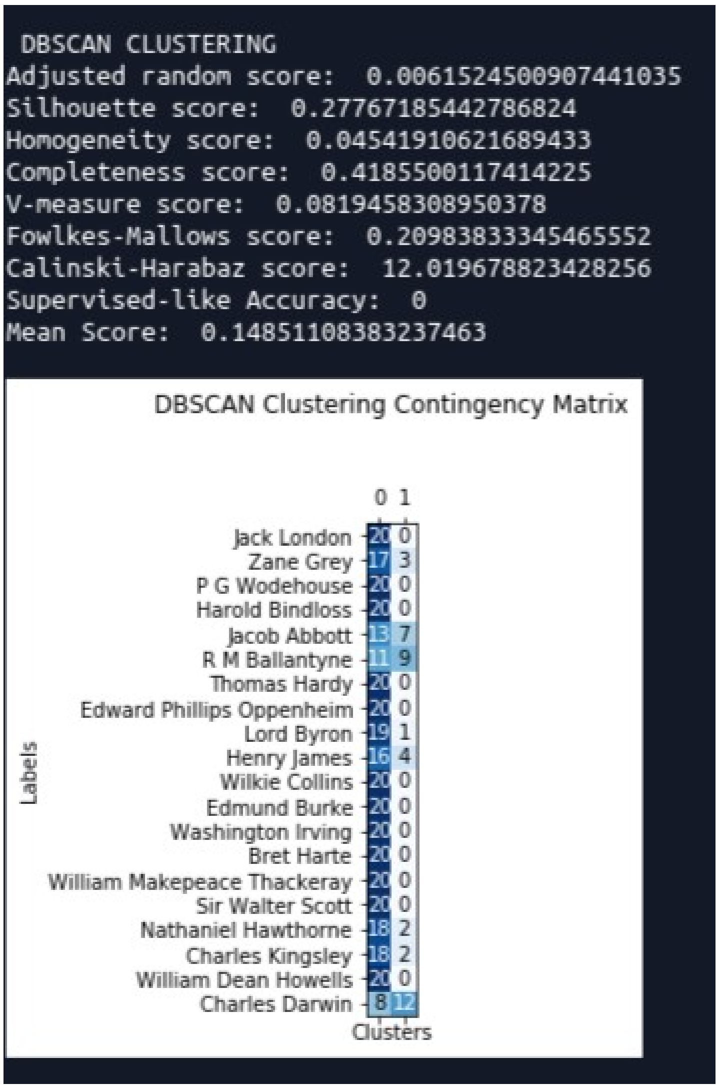


	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Jack London	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	2	0
Zane Grey	15	3	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P G Wodehouse	9	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Harold Bindloss	5	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Jacob Abbott	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R M Ballantyne	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Thomas Hardy	14	0	1	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Edward Phillips Oppenheim	17	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Lord Byron	16	0	0	0	0	0	0	2	2	0	0	0	0	0	0	0	0	0	0	0	0
Henry James	18	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
Wilkie Collins	4	0	14	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Edmund Burke	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Washington Irving	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bret Harte	16	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	2	0	0	0	0
William Makepeace Thackeray	18	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
Sir Walter Scott	4	0	14	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
Nathaniel Hawthorne	18	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
Charles Kingsley	12	0	5	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0
William Dean Howells	14	0	2	2	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
Charles Darwin	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0


DBSCAN Clustering PCA
x
y
z
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
0.2
0.1
0.0
−0.1
−0.2


DBSCAN Clustering TSNE
x
y
z
−100
0
100
200
300
400
−1000
−800
−600
−400
−200
0
200
200
100
0
−100
−200
−300
−400

This is what the default DBSCAN looks like. Still very bad.



```
 DBSCAN CLUSTERING
Adjusted random score:  0.0061524500907441035
Silhouette score:  0.27767185442786824
Homogeneity score:  0.04541910621689433
Completeness score:  0.4185500117414225
V-measure score:  0.0819458308950378
Fowlkes-Mallows score:  0.20983833345465552
Calinski-Harabaz score:  12.019678823428256
Supervised-like Accuracy:  0
Mean Score:  0.14851108383237463
```

DBSCAN Clustering Contingency Matrix

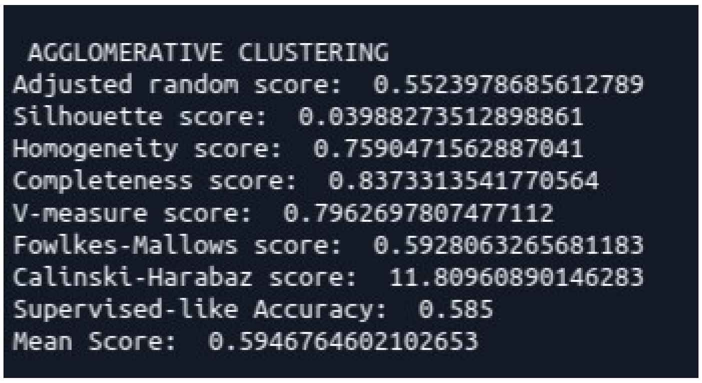
Labels	0	1
Jack London	20	0
Zane Grey	17	3
P G Wodehouse	20	0
Harold Bindloss	20	0
Jacob Abbott	13	7
R M Ballantyne	11	9
Thomas Hardy	20	0
Edward Phillips Oppenheim	20	0
Lord Byron	19	1
Henry James	16	4
Wilkie Collins	20	0
Edmund Burke	20	0
Washington Irving	20	0
Bret Harte	20	0
William Makepeace Thackeray	20	0
Sir Walter Scott	20	0
Nathaniel Hawthorne	18	2
Charles Kingsley	18	2
William Dean Howells	20	0
Charles Darwin	8	12

Clusters

And this is what I imagine grid searching cluster parameters will result in.

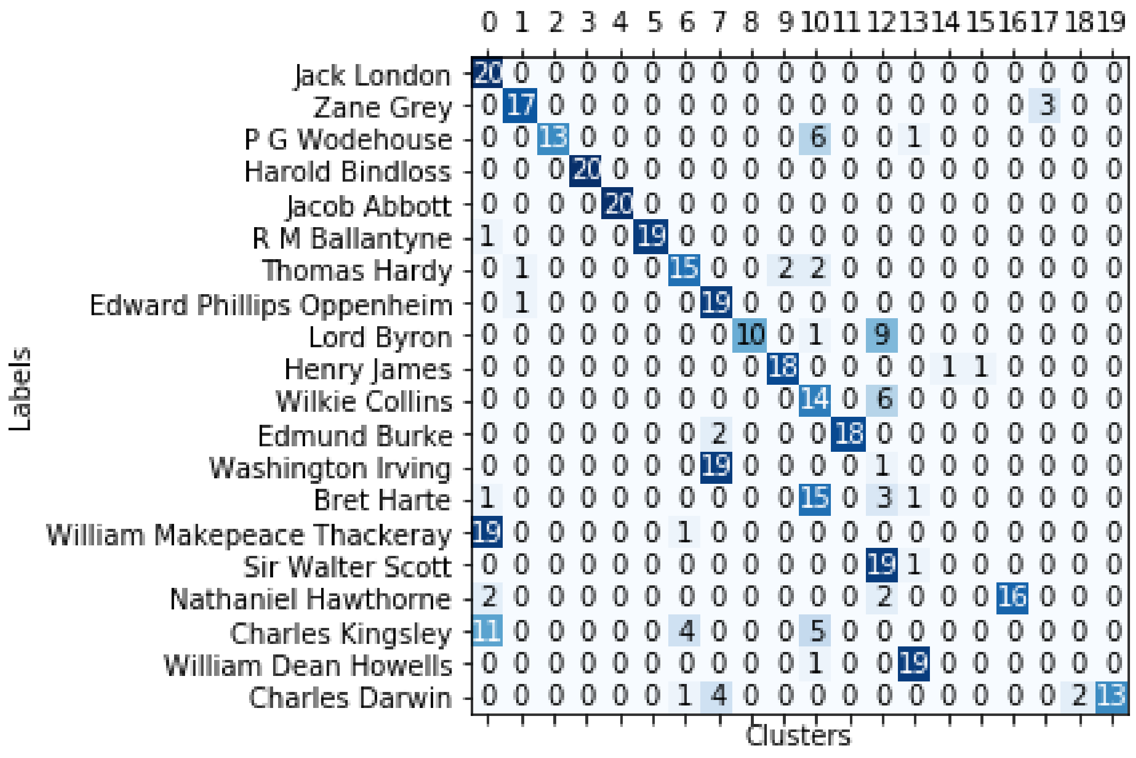
c) Agglomerative Clustering

This is the best clustering. The scores are the highest, the visualizations are the closest to the labels and the contingency matrix looks almost like a confusion matrix.



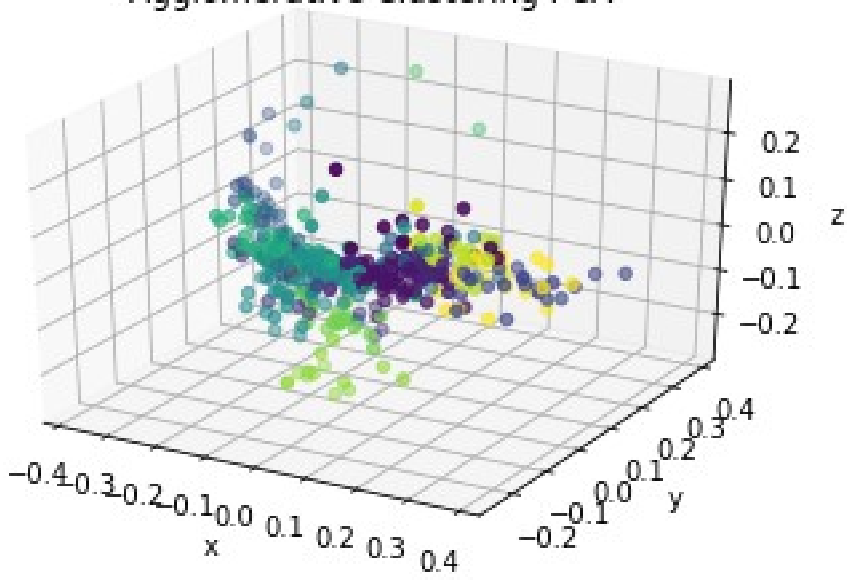
```
 AGGLOMERATIVE CLUSTERING
Adjusted random score:  0.5523978685612789
Silhouette score:  0.03988273512898861
Homogeneity score:  0.7590471562887041
Completeness score:  0.8373313541770564
V-measure score:  0.7962697807477112
Fowlkes-Mallows score:  0.5928063265681183
Calinski-Harabaz score:  11.80960890146283
Supervised-like Accuracy:  0.585
Mean Score:  0.5946764602102653
```

Agglomerative Clustering Contingency Matrix

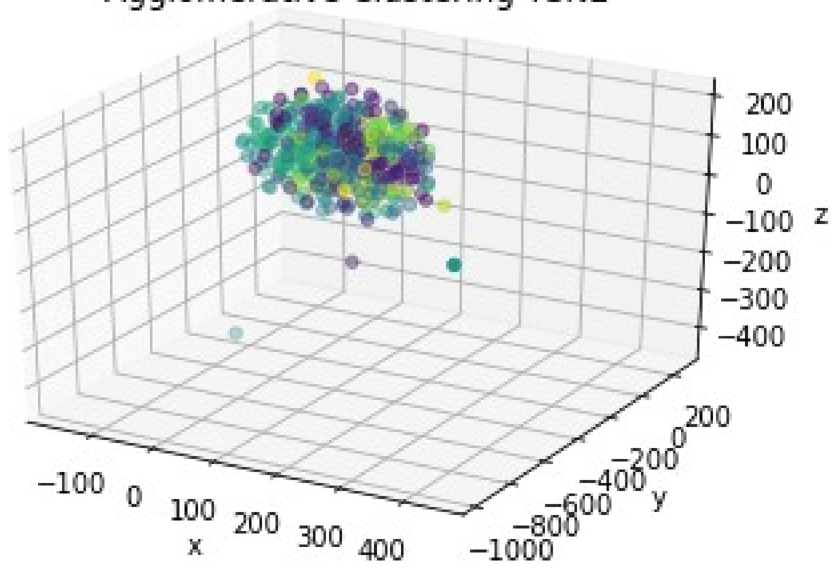


	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Jack London	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Zane Grey	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0
P G Wodehouse	0	0	13	0	0	0	0	0	0	0	6	0	0	1	0	0	0	0	0	0
Harold Bindloss	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Jacob Abbott	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R M Ballantyne	1	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Thomas Hardy	0	1	0	0	0	0	15	0	0	2	2	0	0	0	0	0	0	0	0	0
Edward Phillips Oppenheim	0	1	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0
Lord Byron	0	0	0	0	0	0	0	0	10	0	1	0	9	0	0	0	0	0	0	0
Henry James	0	0	0	0	0	0	0	0	0	18	0	0	0	0	1	1	0	0	0	0
Wilkie Collins	0	0	0	0	0	0	0	0	0	0	14	0	6	0	0	0	0	0	0	0
Edmund Burke	0	0	0	0	0	0	0	2	0	0	0	18	0	0	0	0	0	0	0	0
Washington Irving	0	0	0	0	0	0	0	19	0	0	0	0	1	0	0	0	0	0	0	0
Bret Harte	1	0	0	0	0	0	0	0	0	0	15	0	3	1	0	0	0	0	0	0
William Makepeace Thackeray	19	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Sir Walter Scott	0	0	0	0	0	0	0	0	0	0	0	0	19	1	0	0	0	0	0	0
Nathaniel Hawthorne	2	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	16	0	0	0
Charles Kingsley	11	0	0	0	0	0	4	0	0	0	5	0	0	0	0	0	0	0	0	0
William Dean Howells	0	0	0	0	0	0	0	0	0	0	1	0	0	19	0	0	0	0	0	0
Charles Darwin	0	0	0	0	0	0	1	4	0	0	0	0	0	0	0	0	0	0	2	13

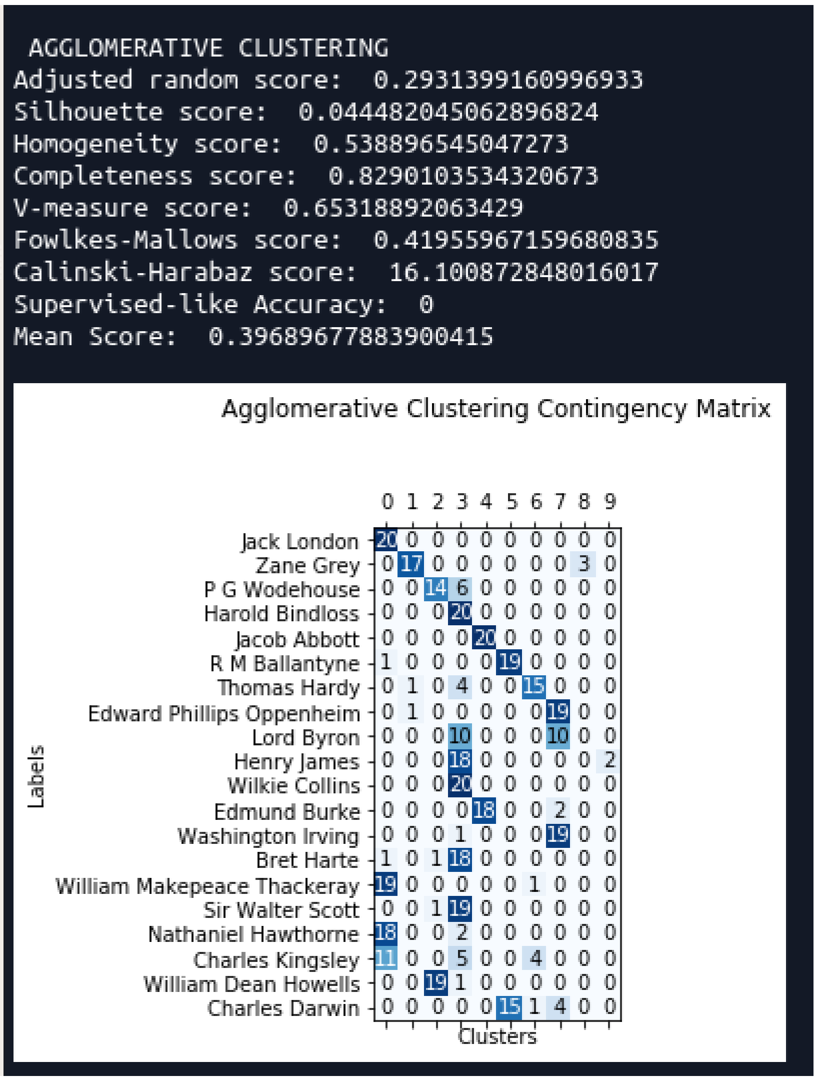
Agglomerative Clustering PCA



Agglomerative Clustering TSNE



This is how it clusters with n_components=10. Only 8 labels are meaningful, the last 2 are very small. Not very good, but still comparable to kmeans.



```
 AGGLOMERATIVE CLUSTERING
Adjusted random score:  0.2931399160996933
Silhouette score:  0.044482045062896824
Homogeneity score:  0.538896545047273
Completeness score:  0.8290103534320673
V-measure score:  0.65318892063429
Fowlkes-Mallows score:  0.41955967159680835
Calinski-Harabaz score:  16.100872848016017
Supervised-like Accuracy:  0
Mean Score:  0.39689677883900415
```

Agglomerative Clustering Contingency Matrix

Labels	0	1	2	3	4	5	6	7	8	9
Jack London	20	0	0	0	0	0	0	0	0	0
Zane Grey	0	17	0	0	0	0	0	0	3	0
P G Wodehouse	0	0	14	6	0	0	0	0	0	0
Harold Bindloss	0	0	0	20	0	0	0	0	0	0
Jacob Abbott	0	0	0	0	20	0	0	0	0	0
R M Ballantyne	1	0	0	0	0	19	0	0	0	0
Thomas Hardy	0	1	0	4	0	0	15	0	0	0
Edward Phillips Oppenheim	0	1	0	0	0	0	0	19	0	0
Lord Byron	0	0	0	10	0	0	0	10	0	0
Henry James	0	0	0	18	0	0	0	0	0	2
Wilkie Collins	0	0	0	20	0	0	0	0	0	0
Edmund Burke	0	0	0	0	18	0	0	2	0	0
Washington Irving	0	0	0	1	0	0	0	19	0	0
Bret Harte	1	0	1	18	0	0	0	0	0	0
William Makepeace Thackeray	19	0	0	0	0	0	1	0	0	0
Sir Walter Scott	0	0	1	19	0	0	0	0	0	0
Nathaniel Hawthorne	18	0	0	2	0	0	0	0	0	0
Charles Kingsley	11	0	0	5	0	0	4	0	0	0
William Dean Howells	0	0	19	1	0	0	0	0	0	0
Charles Darwin	0	0	0	0	0	15	1	4	0	0

Needed to run my code:
https://pypi.org/project/stemming/1.0/
https://pypi.org/project/gensim/