Unsupervised Capstone:

Classify authors based on the style of text

By Karen McGee

Problem Statement

- Build an unsupervised model that will classify authors based on the style of writing using natural language processing

Research Questions:

- Are authors consistently grouped into the same cluster?
- Does your clustering on those members perform as you'd expect?
- Have your clusters remained stable or changed dramatically?
- Does our model provide a consistent performance?

Solution statement

- Import data from ten different authors of various writing styles
- Clean, tokenize and lemmatize the data.
- Generate features using TFIDF
- Generate clusters (i.e. K-mean, MeanShift…etc)
- Evaluate the performance of the clusters
- Generate models (i.e. Random Forest, logistic…etc)
- Evaluate the performance of the models

Evaluation metrics

- Utilize the RI adjusted score to evaluate the performance of each clusters.
- Use confusion matrix, classification report and accuracy score to evaluate the performance of each model.

Analysis – input of raw text file

Example of raw text file – Macbeth

"[The Tragedie of Macbeth by William Shakespeare 1603]\n\n\nActus Primus. Scoena Prima.\n\n
Thunder and Lightning. Enter three Witches.\n\n 1. When shall we three meet againe?\nIn Th
under, Lightning, or in Raine?\n 2. When the Hurley-burley's done,\nWhen the Battaile's lo
st, and wonne\n\n   3. That will be ere the set of Sunne\n\n   1. Where the place?\n 2. Vp
on the Heath\n\n   3. There to meet with Macbeth\n\n   1. I come, Gray-Malkin\n\n   All. Pa
dock calls anon: faire is foule, and foule is faire,\nHouer through the fogge and filthie a
yre.\n\nExeunt.\n\n\nScena Secunda.\n\nAlarum within. Enter King Malcome, Donalbaine, Lenox
, with\nattendants,\nmeeting a bleeding Captaine.\n\n King. What bloody man is that? he ca
n report,\nAs seemeth by his plight, of the Reuolt\nThe newest state\n\n   Mal. This is the
Serieant,\nWho like a good and hardie Souldier fought\n'Gainst my Captiuitie: Haile braue f
riend;\nSay to the King, the knowledge of the Broyle,\nAs thou didst leaue it\n\n   Cap. Do
ubtfull it stood,\nAs two spent Swimmers, that doe cling together,\nAnd choake their Art: T
he mercilesse Macdonwald\n(Worthie to be a Rebell, for to that\nThe multiplying Villanies o
f Nature\nDoe swarme vpon him) from the Westerne Isles\nOf Kernes and Gallowgrosses is supp
ly'd,\nAnd Fortune on his damned Quarry smiling,\nShew'd like a Rebells Whore: but all's to
o weake:\nFor braue Macbeth (well hee deserues that Name)\nDisdayning Fortune, with his bra
ndisht Steele,\nWhich smoak'd with bloody execution\n(Like Valours Minion) caru'd out his p
assage,\nTill hee fac'd the Slaue:\nWhich neu'r shooke hands, nor bad farwell to him,\nTill
he vnseam'd him from the Naue toth' Chops,\nAnd fix'd his Head vpon our Battlements\n\n   K

Analysis – Text file cleaned, lemmatized and tokenized

Example of a cleaned, lemmatized and tokenized file - Macbeth

['Actus Primus Scoena Prima Thunder and Lightning Enter three Witches 1 When shall we three meet againe In Thunder Lightning or in Raine 2 When the Hurleyburleys done When the Battailes lost and wonne 3 That will be ere the set of Sunne 1 Where the place 2 Vpon the Heath 3 There to meet with Macbeth 1 I come GrayMalkin All Padock call anon faire is foule and foule is faire Houer through the fogge and filthie ayre Exeunt Scena Secunda Alarum within Enter King Malcome Donalbaine Lenox with attendant meeting a bleeding Captaine King What bloody man is that he can report As seemeth by his plight of the Reuolt The newest state Mal This is the Serieant Who like a good and hardie Souldier fought Gainst my Captiuitie Haile braue friend Say to the King the knowledge of the Broyle As thou didst leaue it Cap Doubtfull it stood As two spent Swimmers that doe cling together And choake their Art The mercilesse Macdonwald Worthie to be a Rebell for to that The multiplying Villanies of Nature Doe swarme vpon him from the Westerne Isles Of Kernes and Gallowgrosses is supplyd And Fortune on his damned Quarry smiling Shewd like a Rebells Whore but alls too weake For braue Macbeth well hee deserues that Name Disdayning Fortune with his brandisht Steele Which smoakd with bloody execution Like Valours Minion carud out his passage Till hee facd the Slaue Which neur shooke hand nor bad farwell to him Till he vnseamd him from the Naue toth Chops And fixd his Head vpon our Battlements King O valiant Cousin worthy Gentleman Cap As whence the Sunne gin

Analysis – Feature generation of text file

Generate features using TFIDF

```
1 X_tfidf.head()
```

	10	11	12	13	14	15	16	17	18	1839	19	20	21	30	40	45	50	90	_had_	_he_	_page	_so_	_the	_was_	_you_	aback	abandon	aband
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

Analysis – Top terms identified for each cluster

Top terms per cluster:

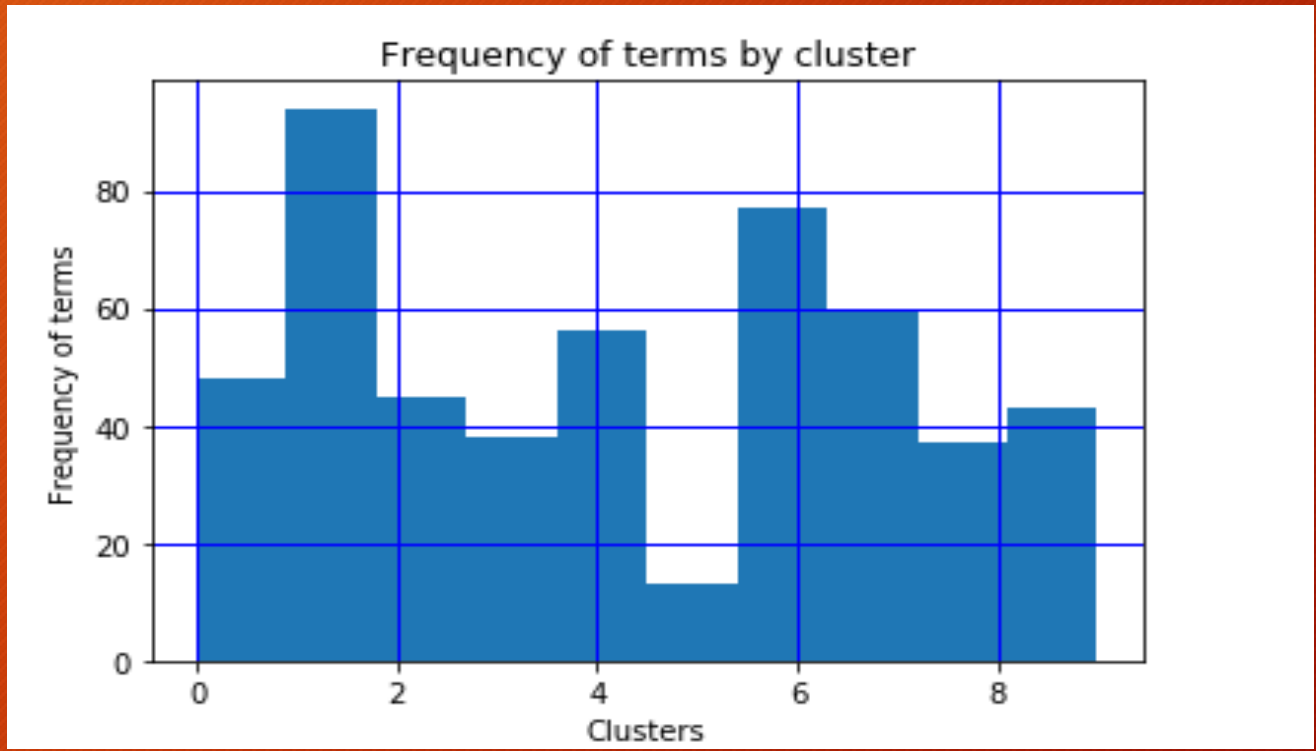
Cluster 0:	Cluster 1:	Cluster 2:	Cluster 3:	Cluster 4:	Cluster 5:	Cluster 6:	Cluster 7:
whaling	michael	love	haue	whale	turnbull	ahab	buster
whale	cross	soul	ham	boat	macian	ye	joe
mouse	brahmin	thee	lord	sperm	evan	whale	bear
queequeg	ebook	shall	enter	sea	quite	ship	browns
whaler	project	song	thou	ship	wall	captain	farmer
oil	lucifer	earth	caesar	leviathan	sword	stubb	otter
ship	king	thy	macb	water	mean	queequeg	blacky
whalemen	girl	woman	king	line	garden	boat	pool
voyage	gutenberg	land	brutus	though	god	thou	trout
dutch	robert	city	bru	whales	really	starbuck	boy

Cluster 8:	Cluster 9:
brown	syme
father	gregory
flambeau	professor
priest	bull
door	sunday
garden	marquis
dont	dr
house	secretary
looked	anarchist
rather	colonel

Top Terms identified for each cluster

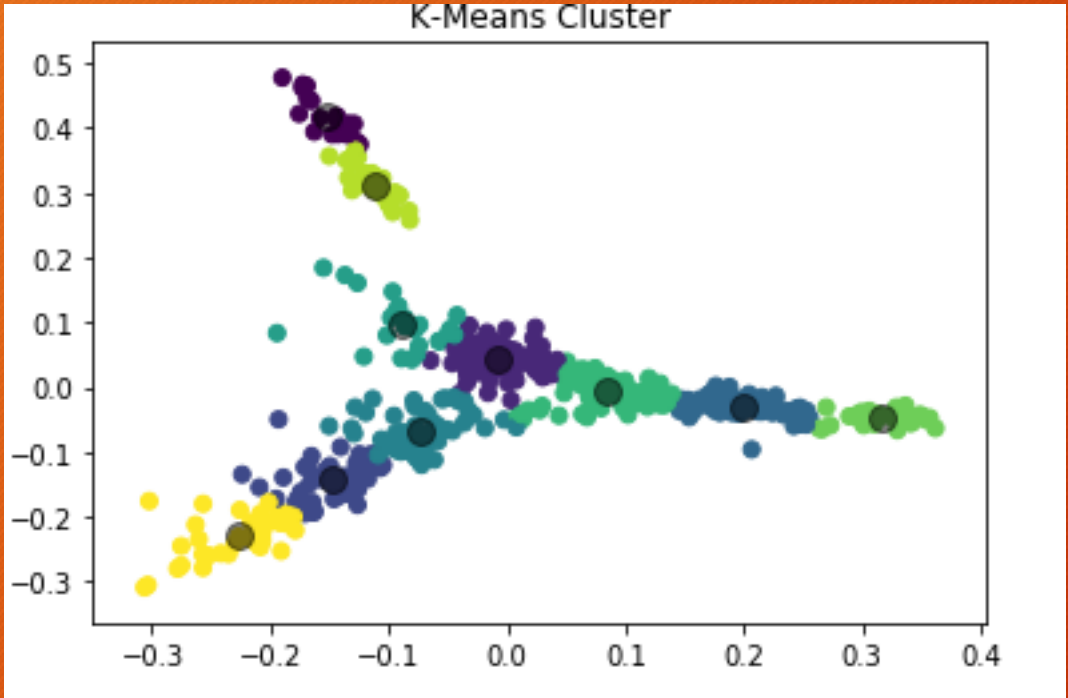
Analysis – Distribution of top terms for each cluster

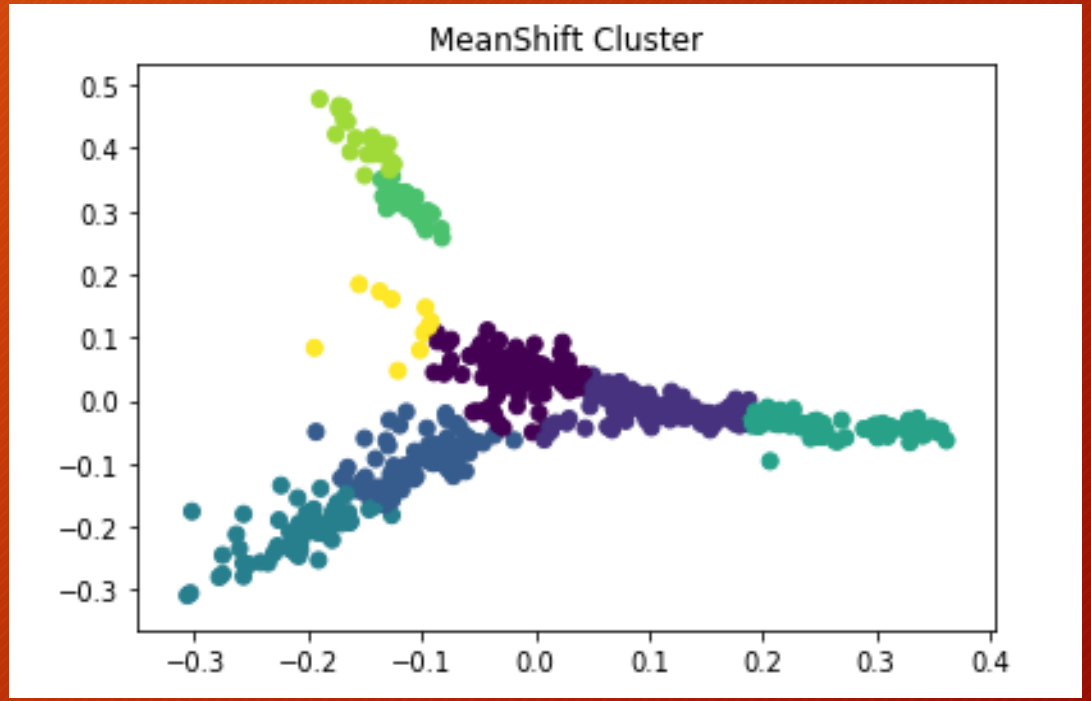
Frequency of terms by clusters



Analysis – Cluster visualizations

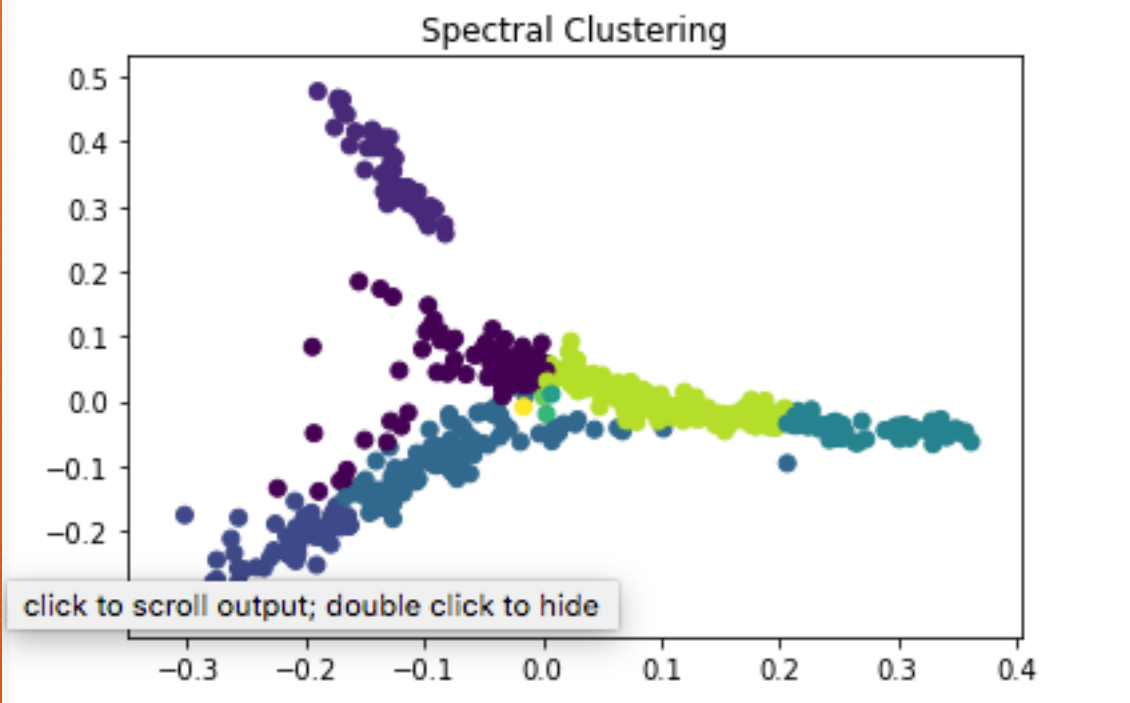
Graphically visualization of each cluster

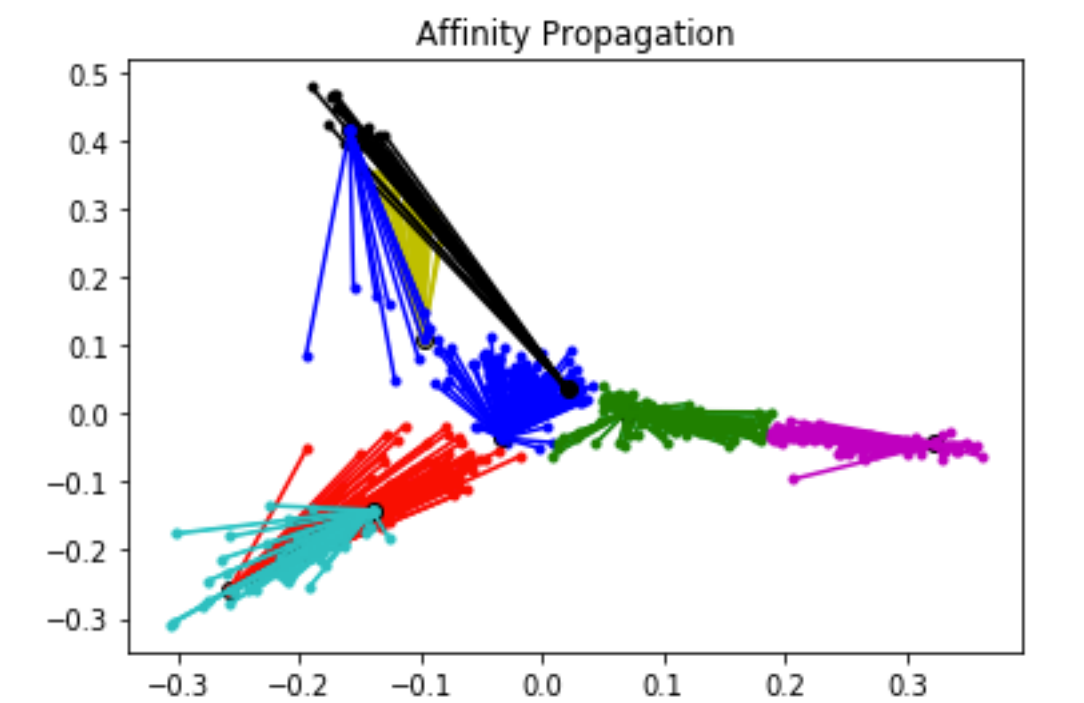




Analysis – Cluster visualizations

Graphically visualization of each cluster





Cluster evaluation results

Overall Spectral Clustering performed the best based on the RI Adjusted score.

	Cluster	Number of clusters	RI Score	RI adjusted score
0	K-Means	10	0.0114633	0.419496
1	MeanShift	8	0.00883209	0.465032
2	SpectralClustering	10	0.034522	0.495028
3	AffinityPropagation	14	0.0382379	0.424641

Model Performance - Knn

```
KNN Training mean set score: 0.9786117517268057
KNN Testing mean set score: 0.936743799727022

 KNN Confustion Matrix
 [[15  0  0  0  0  0  0  0  0  0]
 [ 0 18  0  0  0  0  0  0  0  0]
 [ 0  1 14  0  0  0  1  0  0  0]
 [ 0  0  0  3  0  0  0  0  0  0]
 [ 0  0  0  0  7  0  0  0  0  0]
 [ 0  0  0  0  0  9  0  0  0  0]
 [ 0  0  0  0  0  0 22  0  0  0]
 [ 0  0  0  0  0  0  0  5  0  0]
 [ 0  0  0  0  0  0  0  0 56  0]
 [ 0  0  0  0  0  0  0  0  0 20]]

 KNN Classification Report
                precision    recall  f1-score   support

          ball       1.00      1.00      1.00        15
         brown       0.95      1.00      0.97        18
        bryant       1.00      0.88      0.93        16
  busterbrown       1.00      1.00      1.00         3
        caesar       1.00      1.00      1.00         7
        hamlet       1.00      1.00      1.00         9
        leaves       0.96      1.00      0.98        22
       macbeth       1.00      1.00      1.00         5
    moby_dick       1.00      1.00      1.00        56
      thursday       1.00      1.00      1.00        20

    micro avg       0.99      0.99      0.99       171
    macro avg       0.99      0.99      0.99       171
 weighted avg       0.99      0.99      0.99       171


 KNN accuracy score: 0.988304093567515
```

Model Performance – Support Vector

```
SVC Training mean set score: 0.9942846872753414
SVC Testing mean set score: 0.982729673646481 9

 Support vector cufusion matrix
 [[15  0  0  0  0  0  0  0  0  0]
 [ 0 18  0  0  0  0  0  0  0  0]
 [ 0  0 15  0  0  0  0  0  1  0]
 [ 0  0  0  3  0  0  0  0  0  0]
 [ 0  0  0  0  7  0  0  0  0  0]
 [ 0  0  0  0  0  9  0  0  0  0]
 [ 0  0  0  0  0  0 22  0  0  0]
 [ 0  0  0  0  0  0  0  5  0  0]
 [ 0  0  0  0  0  0  0  0 56  0]
 [ 0  0  0  0  0  0  0  0  0 20]]

 Support vector classification report
                precision    recall  f1-score   support

          ball       1.00      1.00      1.00        15
         brown       1.00      1.00      1.00        18
        bryant       1.00      0.94      0.97        16
  busterbrown       1.00      1.00      1.00         3
        caesar       1.00      1.00      1.00         7
        hamlet       1.00      1.00      1.00         9
        leaves       1.00      1.00      1.00        22
       macbeth       1.00      1.00      1.00         5
    moby_dick       0.98      1.00      0.99        56
      thursday       1.00      1.00      1.00        20

    micro avg       0.99      0.99      0.99       171
    macro avg       1.00      0.99      1.00       171
 weighted avg       0.99      0.99      0.99       171


 Support vector accuracy score: 0.9941520467836257
```

Model Performance – Random Forest

```
RFC Training mean set score: 0.7951986611818358
RFC Testing mean set score: 0.6710493719276053

 Random Forest confusion matrix
 [[14  0  0  0  0  0  1  0  0  0]
 [ 1 14  0  0  0  0  0  0  1  2]
 [ 1  1 11  0  0  0  1  0  2  0]
 [ 1  1  0  1  0  0  0  0  0  0]
 [ 0  0  0  0  6  1  0  0  0  0]
 [ 0  0  0  0  0  9  0  0  0  0]
 [ 0  0  0  0  0  0 22  0  0  0]
 [ 0  0  1  0  2  0  0  2  0  0]
 [ 0  0  0  0  0  0  1  0 55  0]
 [ 4  0  0  0  0  0  1  0  5 10]]

 Random Forest classification report
                precision    recall  f1-score   support

          ball       0.67      0.93      0.78        15
         brown       0.88      0.78      0.82        18
        bryant       0.92      0.69      0.79        16
  busterbrown       1.00      0.33      0.50         3
        caesar       0.75      0.86      0.80         7
        hamlet       0.90      1.00      0.95         9
        leaves       0.85      1.00      0.92        22
       macbeth       1.00      0.40      0.57         5
     moby_dick       0.87      0.98      0.92        56
      thursday       0.83      0.50      0.62        20

     micro avg       0.84      0.84      0.84       171
     macro avg       0.87      0.75      0.77       171
  weighted avg       0.85      0.84      0.83       171

Random Forest accuracy score: 0.8421052631578947
```

Model Performance – Logistic regression

```
LR Training mean set score: 0.9240892056625226
LR Testing mean set score: 0.5687751756050468

 Logistic regression confusion matrix
 [[15  0  0  0  0  0  0  0  0  0]
 [ 0 18  0  0  0  0  0  0  0  0]
 [ 0  0  5  0  0  0  4  0  7  0]
 [ 0  0  0  3  0  0  0  0  0  0]
 [ 0  0  0  0  7  0  0  0  0  0]
 [ 0  0  0  0  0  9  0  0  0  0]
 [ 0  0  0  0  0  0 22  0  0  0]
 [ 0  0  0  0  0  0  0  5  0  0]
 [ 0  0  0  0  0  0  0  0 56  0]
 [ 0  0  0  0  0  0  1  0  0 19]]

 Logistic classification report
                precision    recall  f1-score   support

          ball       1.00      1.00      1.00        15
         brown       1.00      1.00      1.00        18
        bryant       1.00      0.31      0.48        16
 busterbrown       1.00      1.00      1.00         3
        caesar       1.00      1.00      1.00         7
        hamlet       1.00      1.00      1.00         9
        leaves       0.81      1.00      0.90        22
       macbeth       1.00      1.00      1.00         5
     moby_dick       0.89      1.00      0.94        56
      thursday       1.00      0.95      0.97        20

    micro avg       0.93      0.93      0.93       171
    macro avg       0.97      0.93      0.93       171
weighted avg       0.94      0.93      0.92       171

Logistic accuracy score:   0.9298245614035088
```

Model Performance – Gradient Boosting

```
Gradient Training mean set score: 0.9433328314678034
Gradient Testing mean set score: 0.8710421590820822

 Gradient Boosting confusion matrix
 [[15  0  0  0  0  0  0  0  0  0]
 [ 0 18  0  0  0  0  0  0  0  0]
 [ 0  0 12  0  0  0  2  0  2  0]
 [ 0  0  0  3  0  0  0  0  0  0]
 [ 0  0  0  0  7  0  0  0  0  0]
 [ 0  0  0  0  0  8  0  0  1  0]
 [ 0  1  0  0  0  0 21  0  0  0]
 [ 0  0  0  0  0  0  0  5  0  0]
 [ 0  0  0  0  0  1  0  0 55  0]
 [ 0  0  0  0  0  0  0  0  0 20]]

 Gradient Boosting classification report
                precision    recall  f1-score   support

          ball       1.00      1.00      1.00        15
         brown       0.95      1.00      0.97        18
        bryant       1.00      0.75      0.86        16
  busterbrown       1.00      1.00      1.00         3
        caesar       1.00      1.00      1.00         7
        hamlet       0.89      0.89      0.89         9
        leaves       0.91      0.95      0.93        22
       macbeth       1.00      1.00      1.00         5
     moby_dick       0.95      0.98      0.96        56
      thursday       1.00      1.00      1.00        20

     micro avg       0.96      0.96      0.96       171
     macro avg       0.97      0.96      0.96       171
  weighted avg       0.96      0.96      0.96       171

Gradient Boosting accuracy score: 0.959064327485380
```

Summary results: Cluster and model performance

- Are authors consistently grouped into the same cluster? –Need help with determining this answer
- Does your clustering on those members perform as you'd expect? Need help with determining this answer

- Overall the clusters remained stable for every type of cluster (i.e. Kmeans, meanshift, spectural…etc)
- Overall Model performance:
  - KNN and SVM where consistent with each other with as their accuracy scores range from 98 and 99%
  - Logistic regression, Random forest and Gradient boosting were not consistent as their accuracy scores range from 84, 92 and 96%