Text Clustering

- Data preparation and pre-processing
- Data Transformation
- Text Clustering
- Model Evaluation
- Error Analysis

Data Preparation & Pre-Processing

Data Pre-Processing

Data Pre-Processing

- Removing Stop Words
- Converting words to lower case
- Performing Lemmatization
- Partition Every book to 200 Partitions
- Every Partition have 150 words

Data Transformation

- Bag Of Words(BOW)
- TF-IDF
- LDA
- Word2Vec

Bag of Words(BOW) Transformation

A bag of words is a representation of text that describes the occurrence of words within a document.

	aaron	abandon	abandoned	abandoning	abandonment	abated	abb	abbe	abbey	abbott	...	zodiacal	zonal	zone	zool	zoologique	zoologist	zoology	zur	zwischen	zygomatic
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
2	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
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
995	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
996	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
997	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
998	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
999	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

1000 rows × 16048 columns

TF-IDF

Term frequency (TF) vectors show how important words are to documents. They are computed by using:

$$tf(term, document) = \frac{number\,of\,times\,the\,term\,occurs\,in\,the\,document}{total\,number\,of\,terms\,in\,the\,document}$$

	aaron	abandon	abandoned	abandoning	abandonment	abated	abb	abbe	abbey	abbott	...	zodiacal	zonal	zone	zool	zoologique	zoologist	zoology	zur	zwischen	zygomatic
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
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
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
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
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
995	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
996	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
997	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
998	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
999	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

000 rows × 16048 columns

LDA

LDA is used to classify text in a document to a particular topic. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions. Each document is modeled as a multinomial distribution of topics and each topic is modeled as a multinomial distribution of words.

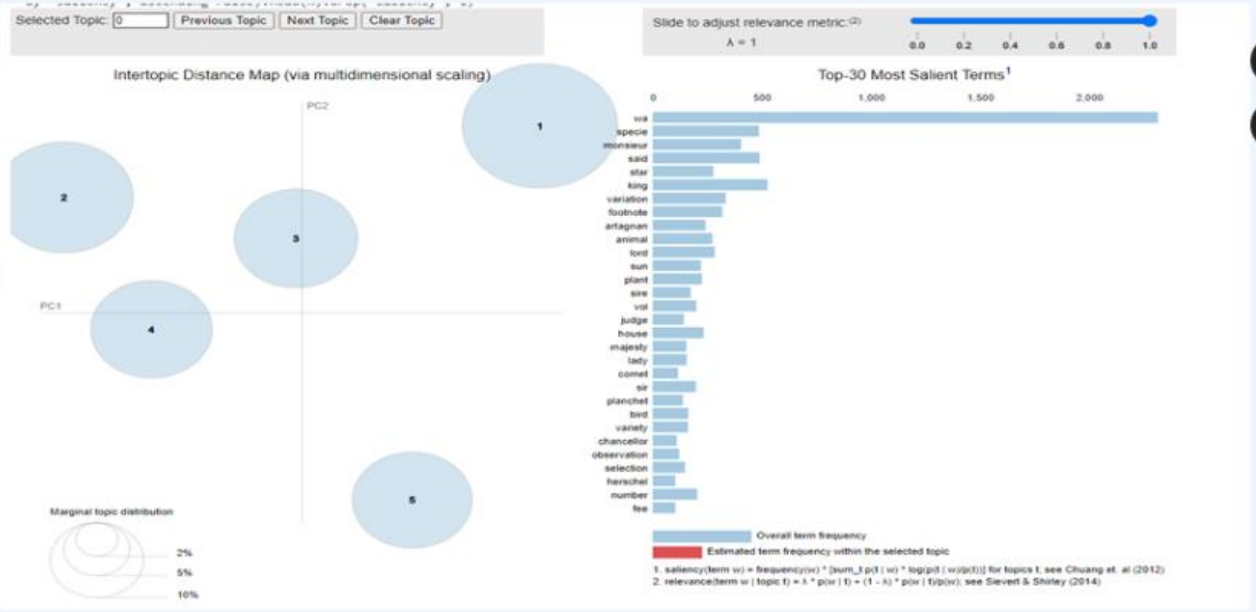
	1	2	3	4	5	res
0	40.195923	0.179589	86.287994	17.819510	6.548666	3
1	6.840732	0.179266	0.174638	9.724366	134.112686	5
2	42.947739	86.382599	3.532179	5.912141	12.257037	2
3	13.155953	1.853795	0.175965	131.157608	4.688347	4
4	11.478949	5.613846	0.174887	15.622274	118.141701	5

```
(array([[ 40.195923  ,   0.17958926,  86.287994  ,  17.81951   ,
           6.548666  ],
        [  6.840732  ,   0.1792659 ,   0.17463806,   9.724366  ,
         134.11269   ],
        [ 42.94774   ,  86.3826    ,   3.5321789 ,   5.9121413 ,
          12.257037  ],
        ...,
        [ 35.19193   , 115.3203    ,   0.17444904,   0.20968111,
           0.13532138],
        [ 20.348946  , 113.14746   ,   7.5316253 ,   0.20784229,
           9.795807  ],
        [  0.37172854,   0.17994398,   4.8293304 ,  26.066813  ,
         119.58385   ]], dtype=float32), None)
```

LDA as a Topic Modeling

LDA as a Topic Modeling

LDA is an example of topic model and is used to classify text in a document to a particular topic. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions.



Word2Vec

Word2Vec consists of models for generating word embedding.

```
[ 6.39032647e-02 -1.00076301e-02 -2.47444082e-02  2.49532200e-02
 -1.64452597e-01 -4.05765250e-02  2.44001463e-01  5.67615628e-02
  2.46330827e-01  1.70743361e-01 -2.93254405e-02  1.92961097e-02
  1.25935256e-01  1.31534727e-03  1.90286748e-02  2.25926861e-02
 -9.74692628e-02  1.18261680e-01  6.58749230e-03  6.97476864e-02
  3.08375317e-03  7.08437189e-02  3.83944809e-02 -5.04423641e-02
 -4.23085783e-03 -7.40122944e-02  1.63575495e-03  2.11675793e-01
 -2.21240018e-02 -2.99621429e-02  1.15621099e-02 -1.41864195e-01
  2.80737784e-02 -5.88704869e-02  3.37919854e-02 -1.00738637e-01
  7.35347271e-02 -6.23905435e-02 -9.10108723e-03  6.61205128e-02
 -1.37935922e-01  2.97885109e-02 -7.19079301e-02  2.71166041e-02
 -5.13751842e-02 -5.43224849e-02 -9.41491723e-02 -1.49340108e-01
 -3.66348475e-02  1.78339824e-01 -6.72631431e-03  3.38511840e-02
  9.75831002e-02  1.75848529e-01  1.65832154e-02  2.20664948e-01
 -1.99374110e-01 -1.83159238e-04  9.93847549e-02  1.08866366e-02
  7.73861483e-02 -2.10308209e-01 -6.40131012e-02  2.25411534e-01
  2.77508423e-02 -1.01526216e-01 -1.68809928e-02  2.53384203e-01
 -1.07851893e-01 -8.49062856e-03 -2.08907742e-02 -2.18841985e-01
  6.51334375e-02 -4.35428321e-02  9.99771878e-02  1.06946655e-01
 -8.23654160e-02 -3.00993957e-02  2.81349123e-01  5.98165877e-02
  1.28292799e-01 -1.07078373e-01  1.92346856e-01  3.58668827e-02
  1.78771645e-01 -4.08566408e-02 -1.65989958e-02  2.05088761e-02
 -1.60516784e-01 -9.39027220e-02  1.24719597e-01 -6.23902828e-02
  1.01312868e-01  1.01647533e-01 -1.16523758e-01  1.58404782e-02
  2.49551699e-01 -2.11534634e-01 -7.85620362e-02 -2.11563744e-02
  1.24335773e-01  1.13423526e-01  8.49671811e-02 -2.49626804e-02
  2.84476101e-01 -2.10208058e-01 -1.88032840e-03  9.00942460e-02
  8.39605778e-02  2.24193856e-01 -4.03151363e-02 -1.30077943e-01
  4.24622511e-03 -2.51844257e-01 -3.11834086e-02 -3.11199576e-02
  2.07286865e-01  2.36877650e-01  1.44839182e-01 -1.97042711e-02
 -1.57025203e-01 -1.11813262e-01 -5.71856424e-02 -1.29472002e-01
  2.70347148e-01  1.84511244e-01  6.89389929e-02 -1.04417324e-01
  1.01622865e-01  2.80982628e-02  1.89747680e-02 -2.16896329e-02
  1.43130109e-01 -9.49415490e-02 -2.38705799e-01  9.39103402e-03
 -3.66527140e-02 -1.06075712e-01 -7.29509890e-02  1.47223040e-01
  5.28361695e-03 -1.24949686e-01  1.47070944e-01 -1.62038818e-01
 -1.51137924e-02  1.67974338e-01 -1.37995034e-01 -1.45792872e-01
  6.50867296e-04  1.07391723e-01]
```

Text Clustering Algorithms

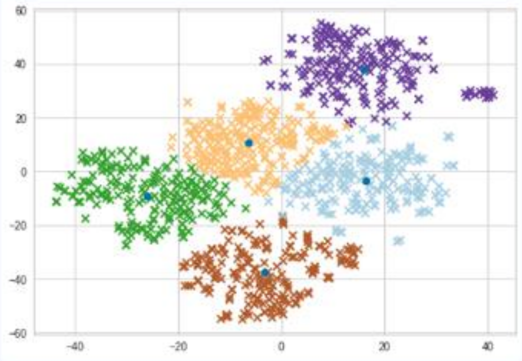
K-Means Clustering

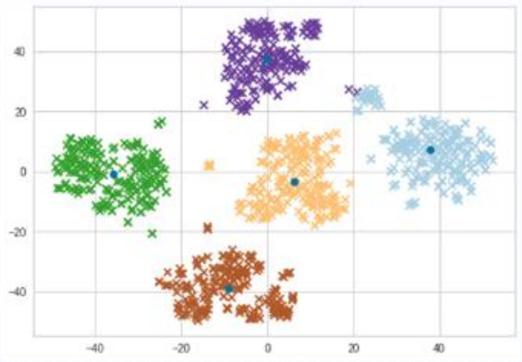
Expectation Maximization

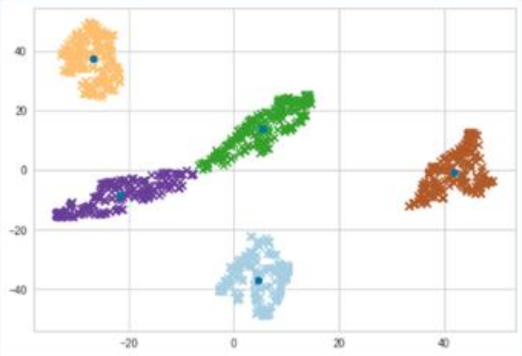
Hierarchical Clustering

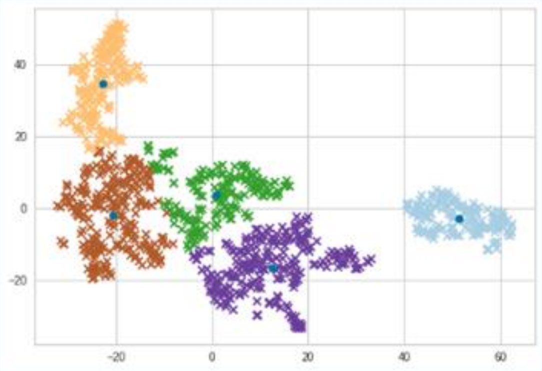
K-Means Clustering

K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. In this topic, we will learn what is K-means clustering algorithm, how the algorithm works, along with the Python implementation of k-means clustering.



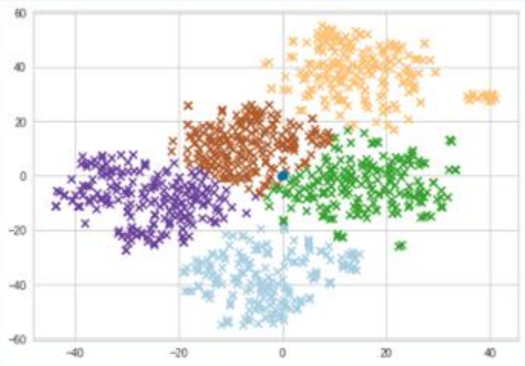


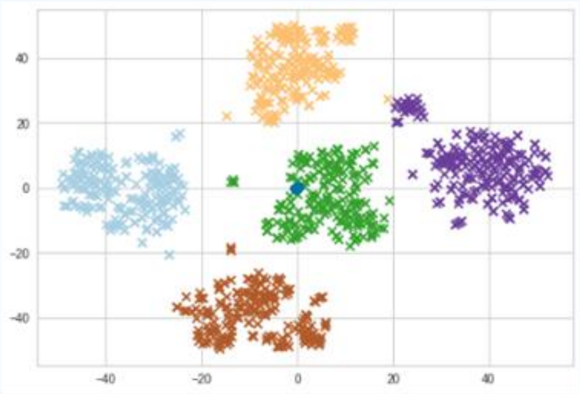


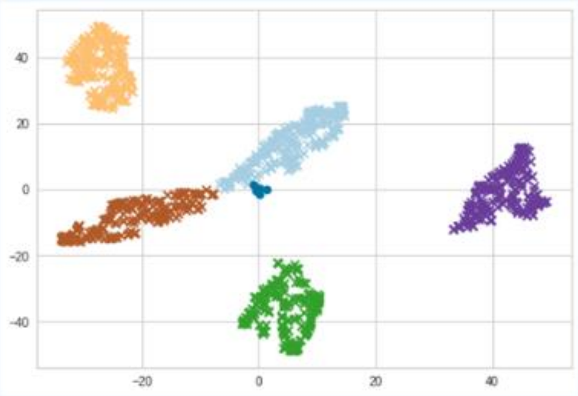


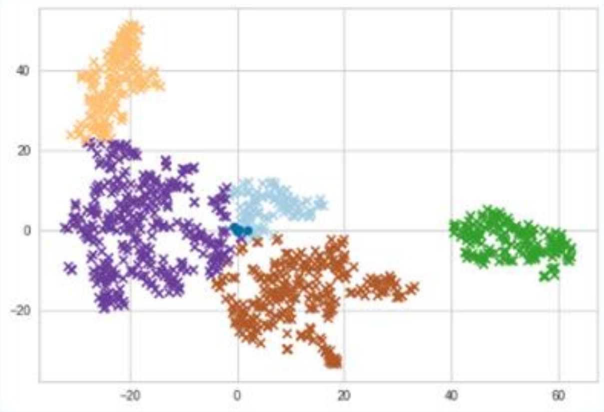
Expectation Maximization

Expectation-Maximization algorithm can be used for the latent variables (variables that are not directly observable and are actually inferred from the values of the other observed variables) too in order to predict their values with the condition that the general form of probability distribution governing those latent variables is known to us. This algorithm is actually at the base of many unsupervised clustering algorithms in the field of machine learning.



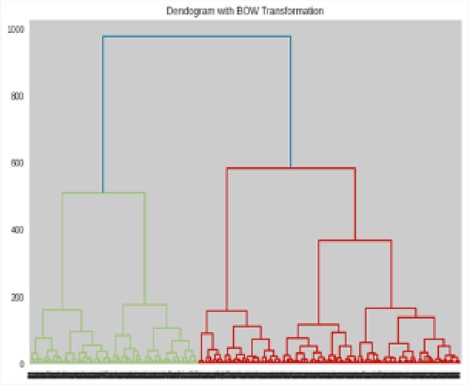


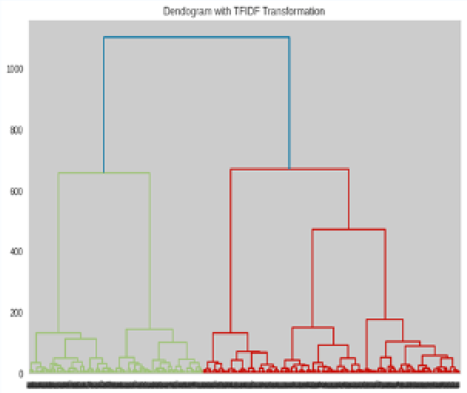


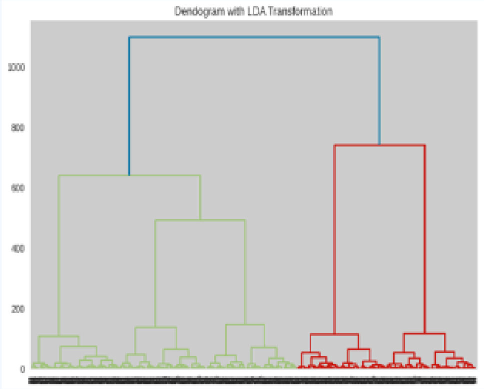


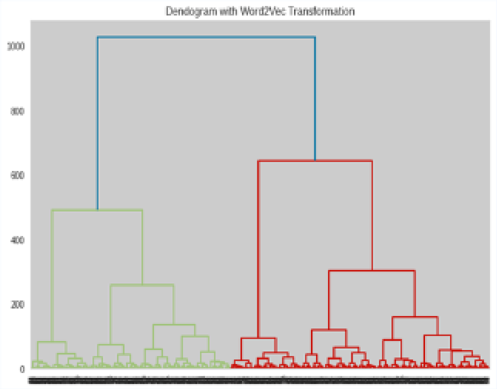
Hierarchical Clustering

Hierarchical clustering, also known as hierarchical cluster analysis, is an algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.









Prezi

Model Evaluation

- Model Evaluation using Kappa
- Model Evaluation using consistency with V-Score
- Model Evaluation using Coherence
- Model Evaluation using Silhouette Score

Model Evaluation using Kappa

Cohen's kappa: a statistic that measures inter-annotator agreement.

Kappa's Score with K-Means clustering algorithm			
K-means with BOW	K-means with TF-IDF	K-means with LDA	K-means with Word2Vec
0.92125	0.96375	0.98625	0.73375
Kappa Score of LDA as Topic Modeling		0.27795	
Kappa's Score with Expectation Maximization(EM) algorithm clustering algorithm			
EM with BOW	EM with TF-IDF	EM with LDA	EM with Word2Vec
0.92750	0.96250	0.98625	0.70125
Kappa's Score with Hierarchal clustering algorithm			
Hierarchal clustering with BOW	Hierarchal clustering with TF-IDF	Hierarchal clustering with LDA	Hierarchal clustering with Word2Vec
0.91000	0.96125	0.98625	0.72875

Model Evaluation using consistency with V-Score

V-measure cluster labeling given a ground truth.

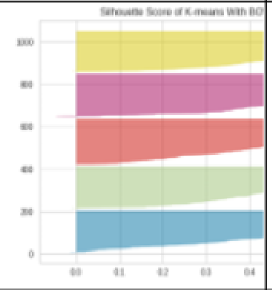
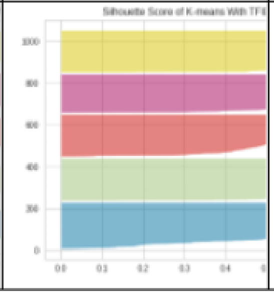
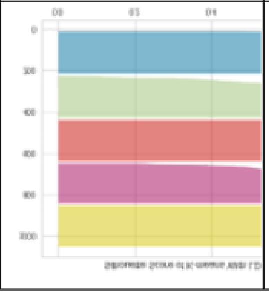
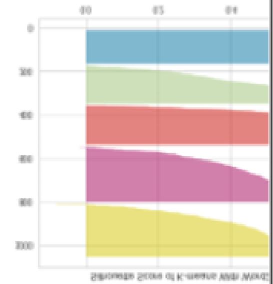
V-Score with K-Means clustering algorithm			
K-means with BOW	K-means with TF-IDF	K-means with LDA	K-means with Word2Vec
0.8466708764209049	0.9227255819276063	0.9604186793621567	0.604108703376123
V-Score with Expectation Maximization(EM) algorithm clustering algorithm			
EM with BOW	EM with TF-IDF	EM with LDA	EM with Word2Vec
0.8523648741765133	0.9212490869283759	0.9604186793621567	0.6168610207344851
V-Score with Hierarchal clustering algorithm			
Hierarchal clustering with BOW	Hierarchal clustering with TF-IDF	Hierarchal clustering with LDA	Hierarchal clustering with Word2Vec
0.8460145486690573	0.9197807272127881	0.9604186793621567	0.627040876468714

Model Evaluation using Coherence

coherence is used to measure how well the topics are extracted.
Coherence score With LDA using c_v: 0.4363
Higher value is better.
Coherence score With LDA using u_mass: -8.0567
Lower value is better.
t that describes the occurrence of words within a document.

Model Evaluation using Silhouette Score

Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other. The Silhouette score is calculated for each sample of different clusters. Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.

Silhouette Score with K-Means clustering algorithm			
K-means with BOW	K-means with TF-IDF	K-means with LDA	K-means with Word2Vec
0.47652	0.62407	0.72321	0.51574
			
Silhouette Score with Expectation Maximization(EM) algorithm clustering algorithm			
EM with BOW	EM with TF-IDF	EM with LDA	EM with Word2Vec
0.47449	0.62412	0.72321	0.48069
Silhouette Score with Hierarchal clustering algorithm			
Hierarchal clustering with BOW	Hierarchal clustering with TF-IDF	Hierarchal clustering with LDA	Hierarchal clustering with Word2Vec
0.46485	0.62326	0.72321	0.46807

Error Analysis

The Model Misclassified 64 rows

The Most common collocations with their repeat count in all records that were labelled uncorrectly according to the human label

[(('cost', 'almost'), 6), (('anyone', 'anywhere'), 6), (('english', 'character'), 6), (('character', 'set'), 6), (('almost', 'restriction'), 6), (('distributed', 'proofreading'), 3), (('copy', 'give'), 3), (('anywhere', 'cost'), 3), (('imp', 'tersbourg'), 1), (('louise', 'valliere'), 1), (('license', 'included'), 1), (('language', 'english'), 1), (('encoding', 'iso'), 1), (('give', 'away'), 1), (('de', 'sci'), 1), (('clearly', 'reader'), 1), (('extent', 'independently'), 1), (('date', 'january'), 1), (('excess', 'defect'), 1), (('heart', 'mind'), 1)]

The most common words that threw the machine off

	PartitionsList	Label_of_Book	index	clustersOutput
177	vol rosenberger calculated though lived lynn o...	e	4	0
3	may aware project gutenberg ha involved writin...	d	3	0
189	dim perception already arrived perhaps observa...	a	0	4
71	body much exception part agree large marked in...	c	2	4
81	considerable number spread varying distance si...	c	2	4
...	...	...	...	...
174	beneficial le beneficial le size body would be...	c	2	4
25	wa thus recognised domain far reaching specula...	e	4	1
50	trans vol footnote ibid vol cvii footnote bull...	e	4	0
175	dissipation extinction footnote footnote allge...	e	4	0
115	undoubtedly wa accads clear authentic insight ...	a	0	2

64 rows × 4 columns

The most common words that threw the machine off