DTI5125[EG]: Data Science Applications

Group Assignment 2 Clustering

Group: 6

Contents

1. Introduction

1.1. Problem Formulation

During the 20th century, the industrial and scientific century, there was a significant rise in the effect of philosophy on people accompanied by the emergence of new philosophies. Out of these main philosophies were three primary ones that are prominent until now and made a breakthrough back then: Nihilism, Existentialism and Absurdity.

Each philosophy had its proponents and famous philosophers, like Nietzsche and his contributions in this field.

However, right now when people read for Nietzsche or similar philosophers, they might get confused whether this book/novel advocates for nihilism or existentialism.

Is this for those who get confused, aren't intellectual enough? Or may be the boundary between the three philosphies isn’t sharp or discriminative enough.

For this, we will ask for the aid of machine learning and AI to help us know if we can differentiate between these 3 philosophies; i.e. **cluster** them in 3 different fine and accurate clusters, or we might find something more interesting?

In this assignment we will work with this question in our minds: Is there a significant boundary between the 3 famous philosophies that we sometimes get confused between, or not?

2. Data

2.1. Chosen data

We are working on the Gutenberg digital books along with some download PDFs as we didn’t find all the books needed on Gutenberg. We chose specifically 5 books; 2 books for Nihilism, 2 for existentialism and 1 for absurdity. Since our main focus was on nihilism and existentialism. We believe that they’re close semantically and even confusing to us, so we want to see what AI will tell us about this confusion.

2.2. Data Preparation

For the pdf books we use a python library “pdfplumber” to extract text from the pdf.
Then we proceed as before—in assignment 1, where we take from each book **n** *(e.g. 200)* **random samples**, comprising **m** *(e.g. 100)* **words** each. Each sample is accompanied by its label *(i.e. the philosophical category of the book)* that will be used afterwards for validating the clustering results.

2.3. Data Preprocessing

The main milestones used for preprocessing in NLP were utilized in our case.
Tokenization, lowercasing, removing English stopwords and some custom stopwords that might cause leakage such as the word ‘nihilism’ or ‘existentialism’ in books, also removing punctuations.
In addition to that, we used lemmatization but refrained from accompanying it with stemming as when both methods are integrated together, the word loses its morphology.

The output of this step is clean and neat 200 segments from each book.

2.4. Feature Engineering

Now that we finished cleaning the text itself, we need to make it interpretable by the machine. In other words, we need to convert the text into numbers; vectorization.
Vectorization has different methodologies that we tried some of; as BOW, TF-IDF and Word2vec.
Word2vec is responsible mainly for capturing semantics as known.
In word2vec, we measure the similarity between words from same documents and words from different documents through distance measures as Euclidean and cosine distance to validate the discrimination between words from different philosophies before being fed to any model.

3. Models and Evaluation

For the models' section, there are quite several clustering models and different combinations can be used between various models and feature engineering techniques, we shall summarize the experiments done in the table below (table 1) and then give insights on the most informative ones.

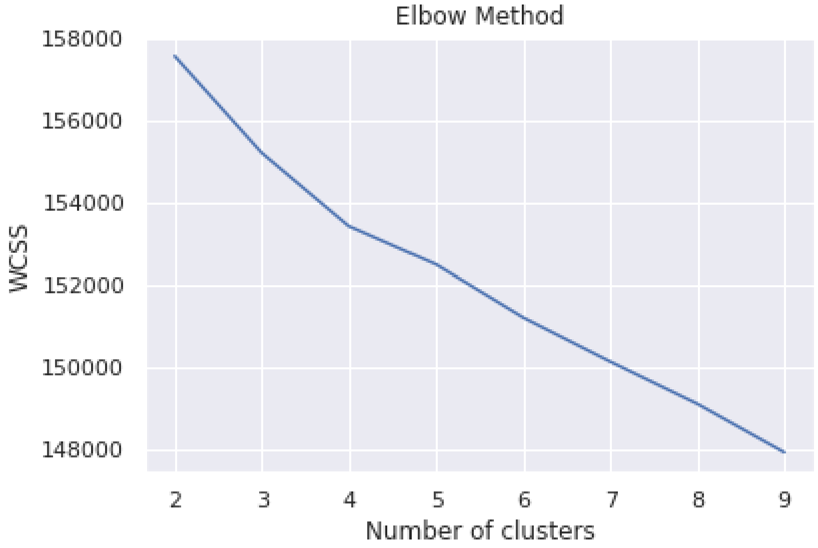
Clustering models that were used: kmeans, Agglomerative clustering, EM clustering and LDA was used as a clustering method (Topic modelling)

For each combination we provide 2 metrics for measuring its performance which are the silhouette score and kappa score.

3.1. Number of clusters

The number of clusters in clustering algorithms is not known a priori, hence it's a common way to use the "Elbow" method to guide us for the proposed number of clusters to be experimented.

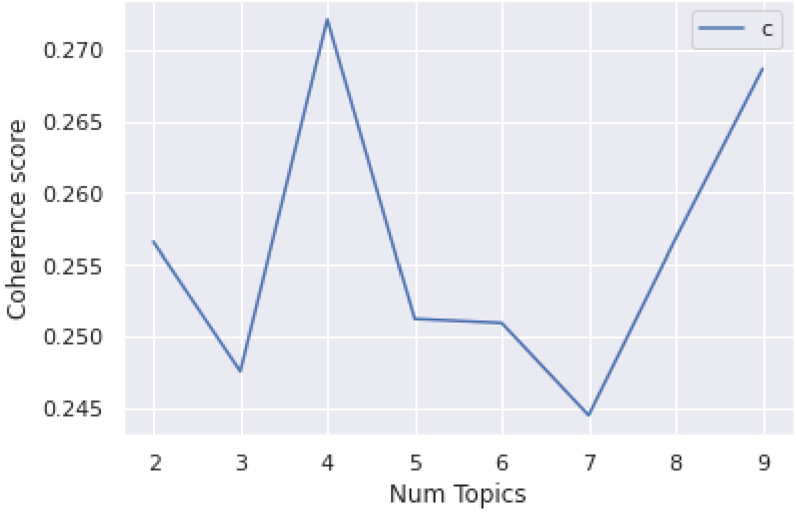
The graph below is an example for so and is obtained from the combination of (BoW + Kmeans)



However, after plotting WCSS for each model with different number of clusters the elbow wasn't very visible enough so we used a function to locate it called "KneeLocator" and the output was 3, so the optimal number is 3 which is plausible because we already have 3 categories.

For LDA, there's a similar method used to inform us with the proposed number of topics (clusters) to use, which is called the coherence score.

The graph below is an example for so and is obtained from the combination of (BoW+LDA)



We take the highest value in coherence as our proposed number of topics, apparently here we should use 4 topics.

Insight:

After investigation, the topics seemed to be: Nihilism, existentialism, absurdity and a 4th topic that holds some records from existentialism and absurdity (a confusion cluster). This means that there's some confusion plane that occurs between existentialism and absurdity.

3.2. Evaluation metrics

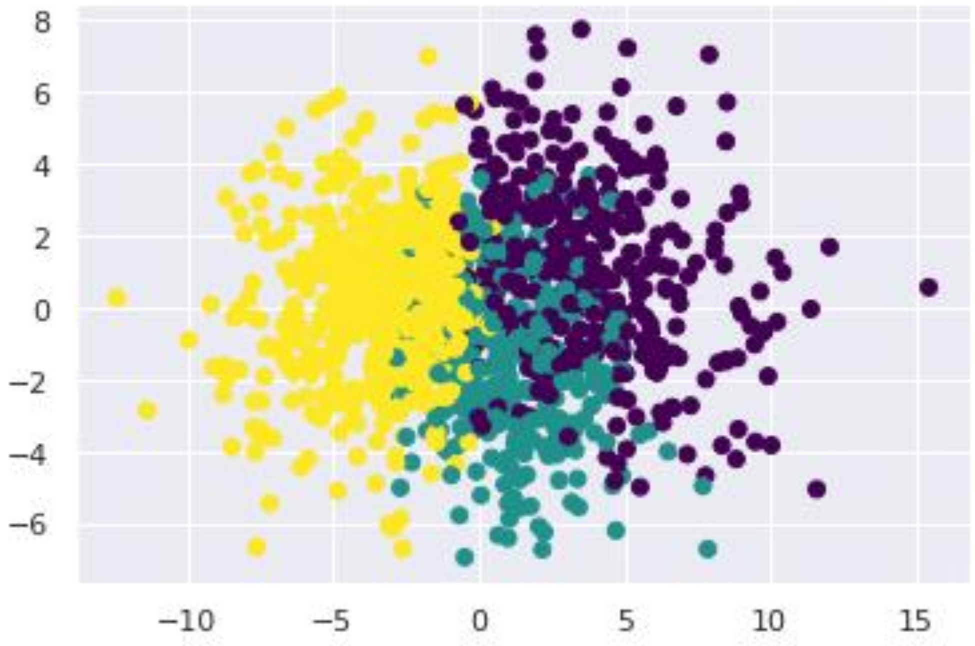
Two main metrics are used: kappa→ measures what the classification results would do corresponding to clustering, as it's a measure of how closely the instances classified by the *algorithm* matched the data labeled as *ground truth [1]*

Silhouette→ measures how small is the inter-class distance (points in the same cluster) and how large is the intra-class distance (distance between clusters), how similar an object is to its own cluster compared to other clusters [2]

3.3. Experiments and Results

# Index	Feature Engineering	Clustering Model	Number of clusters	Silhouette score	Kappa score
Ex_#1	BoW	Kmeans	3	0.0175	0.55
Ex_#2	TF-IDF	Kmeans	3	0.0133	0.72
Ex_#3	Word2Vec	Kmeans	3		
Ex_#4	Word2Vec	EM(GMM)	3		
Ex_#5	TF-IDF	Agglomerative	3		
Ex_#6	BoW	LDA	4		

Example for the clustering the EM (by gaussian mixture model) for 3 clusters is as follows:



There seems to have some overlapping, but quite good discrimination between some of them.

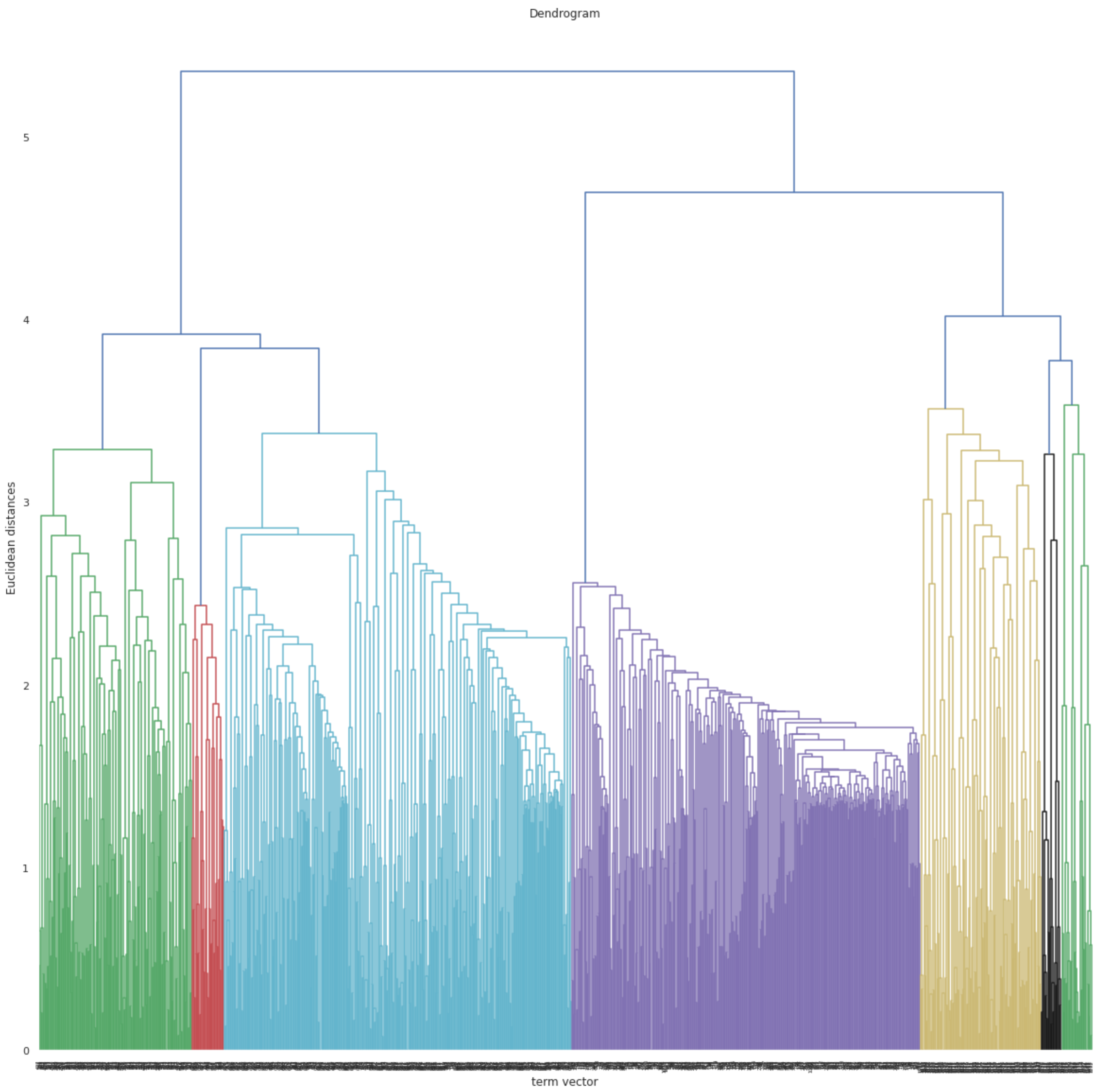
For the Agglomerative clustering: Agglomerative clustering is a type of hierarchical clustering. At first, each point is assigned a separate cluster then in each iteration, the closest pairs of clusters are merged together.

Here, we plotted a dendrogram for our data points to help us decide the number of clusters.

Dendrogram is a diagram that shows the hierarchical relationship between the terms' vectors. The vertical lines in the dendrogram represent the Euclidean distance between the vectors. If the height of the vertical line joining two objects is small, that means these two objects are similar. The number of clusters is determined by drawing a horizontal line through the dendrogram at the desired distance. Each group of observations joined together below the horizontal forms a cluster. Dendrograms don't determine the optimal number of clusters. However, they can be helpful to visualize the similarity between the objects and it will be up to the data scientist to decide where to draw the horizontal line.

For the agglomerative clustering model, we applied the ‘ward’ linkage method with ‘euclidean’ affinity parameter. Linkage method calculates the similarities between objects according to the affinity parameter. Object pairs that minimize the linkage value will be merged together. ‘Ward’ linkage method minimizes the variance of the clusters being merged.

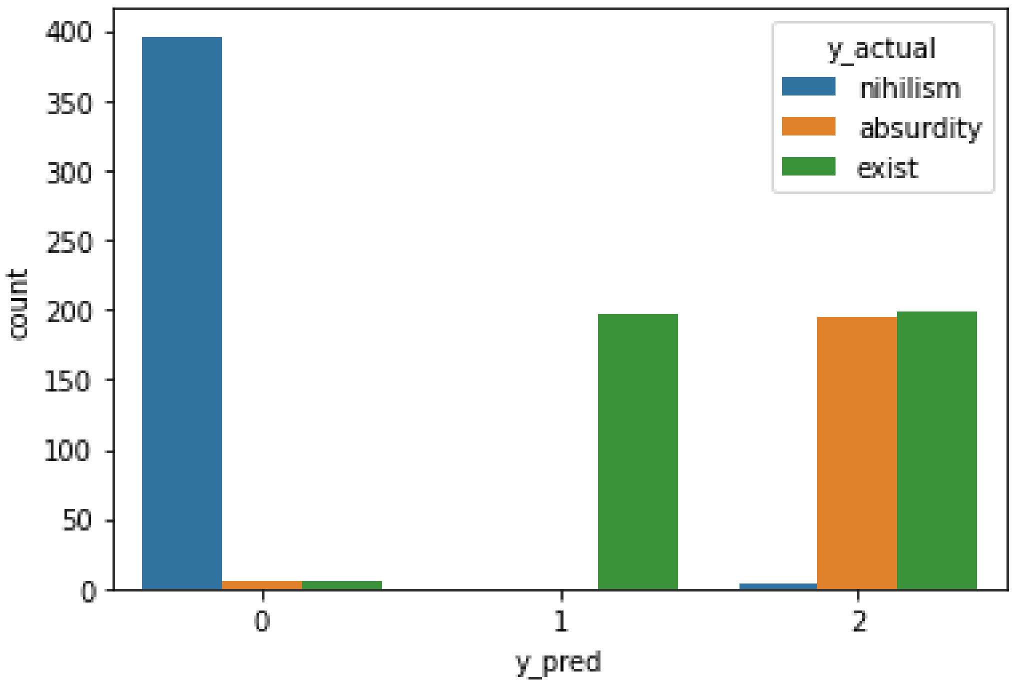
An example for its output is shown below.



3.4. Clusters' constituents

Now we want to know for some candidate experiments, what were the contents of the clusters formed, how many records in each cluster and whether the records are from only one category or there might be some overlapping.

Example for this (TF-IDF + Kmeans)



Insight:

Here we can see that nihilism is easily separated from other clusters, there seem to be a fine line between nihilism and the 2 other philosophies. However, between existentialism and absurdity there seems to be some confusion. While a part of existentialism is represented in one cluster, another part is being confused with absurdity (that's mainly accumulated in one cluster)

From this we may draw a preliminary conclusion, which is that existentialism has some similarities with absurdity semantics and words.

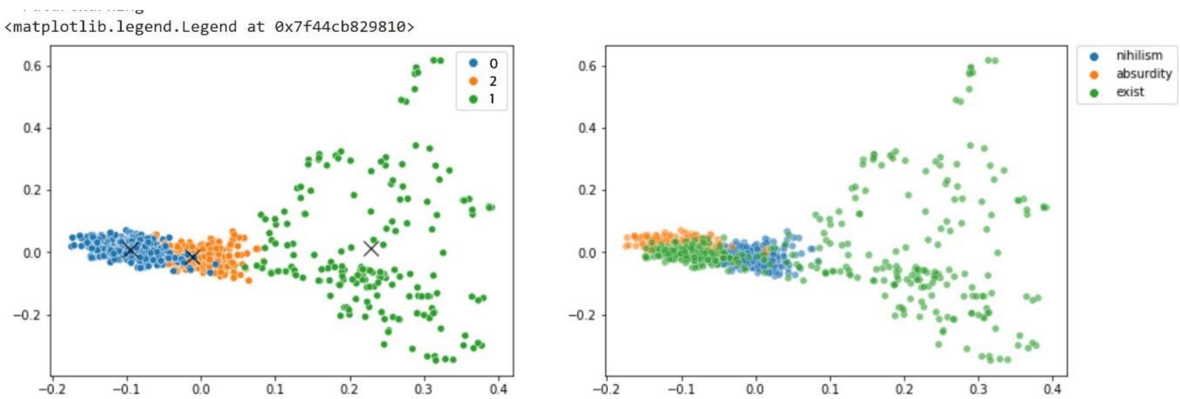
After looking at these, we assume that cluster_0 refers to Nihilism, cluster_1 to existentialism and for simplicity we just refer to cluster_2 as absurdity (the left out category)

3.5. Visualizing Clusters

Firstly, we need to make PCA to reduce the huge dimension of the features *(e.g. word2vec features can be 7500 features and even sometimes 20,000 features on changing the vector sizes, and same goes for TF-IDF)* to be visualized in a 2D image.

Then, we use seaborn to plot the predicted clusters with their centres (in the left hand side) along with a plot for the true categories/clusters and they're represented (right hand side).

Example for the visualization (TF-IDF + Kmeans)



```
<matplotlib.legend.Legend at 0x7f44cb829810>
```

Insight:

From the prediction plot, we can observe that nihilism is predicted to be far from both with little overlapping, however the existentialism is quite spread over the space with some overlapping with the absurdity.

When we look at the real categories, we can see that it’s -more or less- consistent with our prediction, and showing this confusion area even more, where we can see how the existentialism is spread to overlap more and more with the rest 2 clusters.

3.6. Error Analysis

In general, XAI (Explainable AI) has been a field of intensive research, as even though we’ve validated the value of machine learning, it remained as a black box for a long time. Recently, the scientific and research circles have been motivated towards opening this black box and exploring its treasures.

Nevertheless, XAI for clustering and specifically text clustering is still an idle area with not very much contributions.

However, here we provide two error analysis techniques, first one is from a combination of kmeans and TF-IDF (using ELI5) and another one is from LDA clustering (using pyLDA).

3.6.1. ELI5 *(by a trick)*

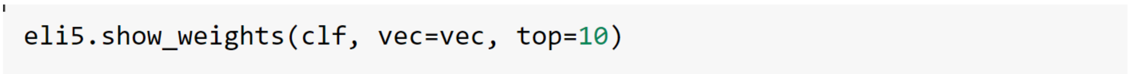
To overcome the scarcity of XAI techniques for text clustering a trick was proposed to use the XAI techniques of text classification in clustering [3].

By providing the prediction of clustering algorithms (kmeans) as the dependent variable of a classification algorithm, along with the training features that were used for clustering beforehand, we get a classification algorithm that can be then fed to XAI techniques.

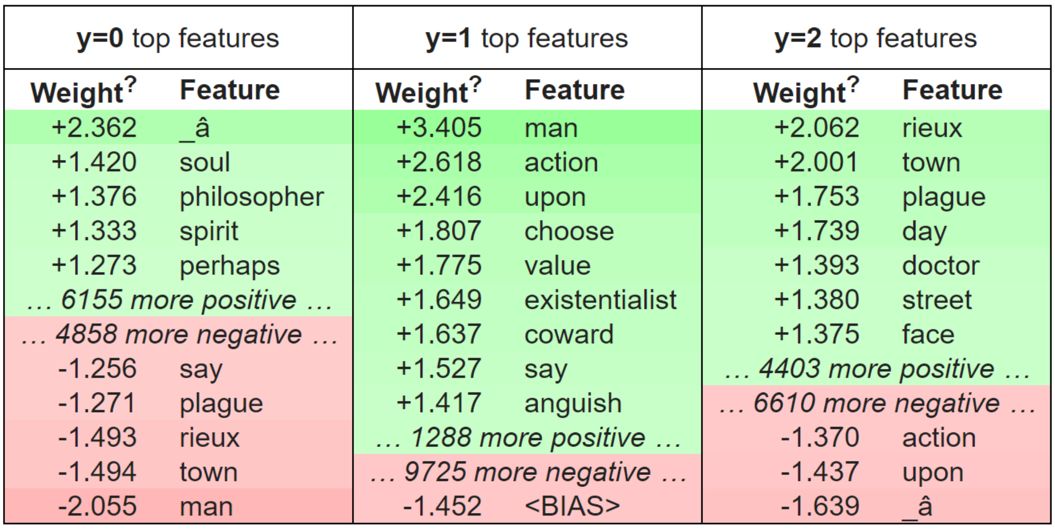
In this context, classification algorithm isn’t used for classification as we’re not using any validation or testing sets, rather it’s used as a ‘mapping function’ that maps these features to that specific cluster (that was already predicted like so by kmeans, regardless of being a true prediction or not)

Therefore, after doing this we can easily use our aforementioned methods as ELI5 and LIME, to interpret the reason why these features were considered to belong to such cluster.

Top words concerning each cluster are obtained:



```
eli5.show_weights(clf, vec=vec, top=10)
```



y=0 top features		**y=1** top features		**y=2** top features	
Weight?	**Feature**	**Weight?**	**Feature**	**Weight?**	**Feature**
+2.362	_â	+3.405	man	+2.062	rieux
+1.420	soul	+2.618	action	+2.001	town
+1.376	philosopher	+2.416	upon	+1.753	plague
+1.333	spirit	+1.807	choose	+1.739	day
+1.273	perhaps	+1.775	value	+1.393	doctor
… 6155 more positive …		+1.649	existentialist	+1.380	street
… 4858 more negative …		+1.637	coward	+1.375	face
-1.256	say	+1.527	say	*… 4403 more positive …*	
-1.271	plague	+1.417	anguish	*… 6610 more negative …*	
-1.493	rieux	*… 1288 more positive …*		-1.370	action
-1.494	town	*… 9725 more negative …*		-1.437	upon
-2.055	man	-1.452	<BIAS>	-1.639	_â

3.6.1.1. Insights

In cluster 1 (existentialism) → we can observe that the highest words are ones like “man”,”action”,”freedom”,”upon” which is consistent with the core of existentialism that focuses on human being, the centralization of the man with his ultimate ‘freedom’ to ’act’ anything upon his choices not constrained or restricted with anything or anyone.

Whilst for cluster 0 (nihilism) we can find words as “soul”, “spirit” are the most frequent words as nihilism focuses on abolishing any meaning for this life which primarily emanates from inside—spiritual aspect. Also, “perhaps” which shows the uncertainty that was prevailed in nihilism, as it was a new, peculiar, bizarre and disappointing philosophy back then, thus it was accompanied by uncertainty.

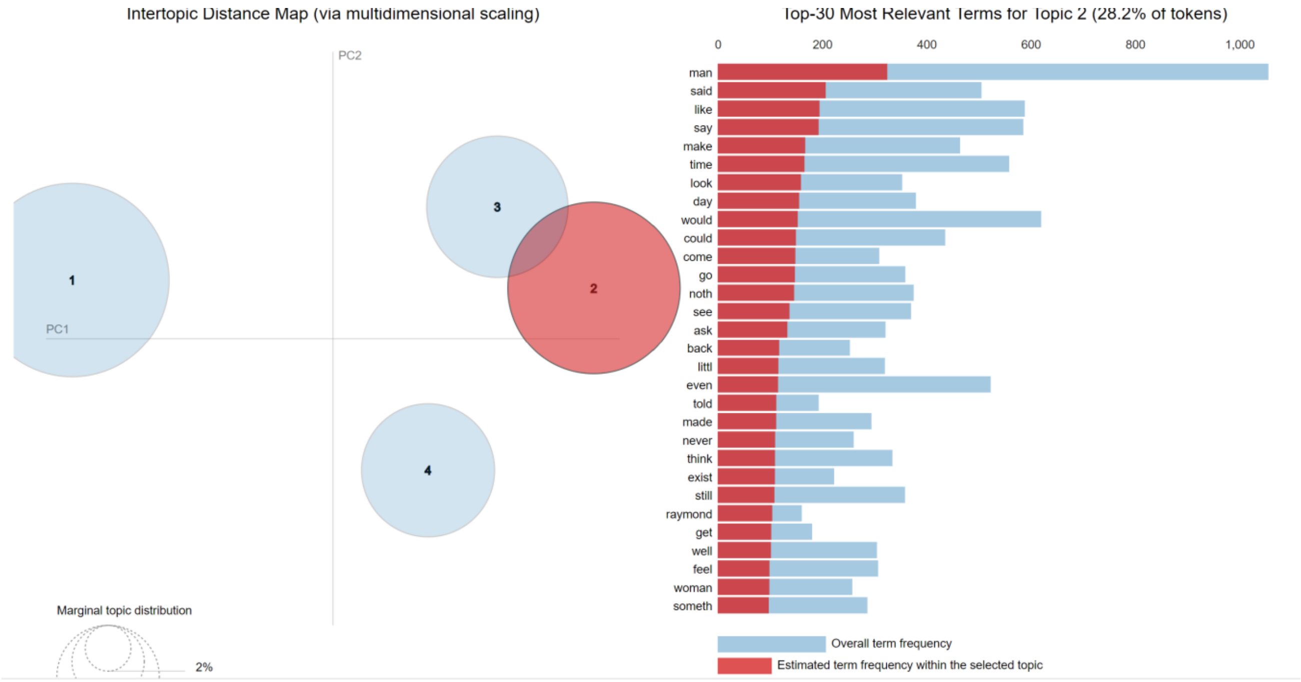
Furthermore, for cluster 2 (absurdity) it seems that there’re more of general words: “doctor”, “day”, “face”, “town”, these words are supposed to be common in usage between absurdity and existentialism since they were clustered together in this. But also there seem to be some words related to absurdity only like “plague”

3.6.2. pyLDAvis

Another method uses a tool above LDA which is pyLDAvis [4] to visualize the clusters, the top words of them along with their weights, as you click on a word you can see how the clusters inflate or deflate based on the contribution/weight of this word on them.

It’s an interactive method, and as we said before in LDA we calculated it for 4 topics. From this interactive visualization we could draw the insight mentioned aforehand about the interpretation of the 4 topics.

As we can see in the following figure, topic 2 encompasses some relevant words from topic 3 (existentialism): [man, said, make] that we could validate their intersection and presence in topic 3 by clicking on it, and also words from topic 4 (absurdity): [day, time, could] and so on.

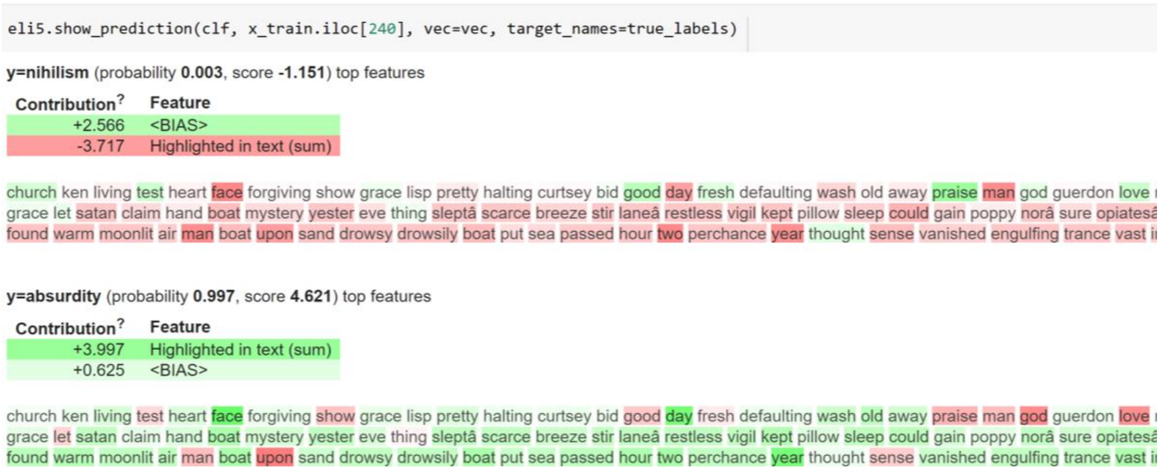


3.7. Misclassification

Two techniques are provided, one of them is still under investigation but seemed quite promising. In each technique we look at the record’s words in conjunction with the top words/meaning of the clusters’ words. ELI5 and LDA sentence coloring.

Firstly, for kmeans and TF-IDF we had about 200 records out of 1000 records that were misclassified. On analyzing one of them we can see the different contribution of words.

Note that: The output is clipped to focus on some words.



```
eli5.show_prediction(clf, x_train.iloc[240], vec=vec, target_names=true_labels)
```

y=nihilism (probability **0.003**, score **-1.151**) top features

Contribution?	Feature
+2.566	<BIAS>
-3.717	Highlighted in text (sum)

church ken living test heart face forgiving show grace lisp pretty halting curtsey bid good day fresh defaulting wash old away praise man god guerdon love
grace let satan claim hand boat mystery yester eve thing sleptâ scarce breeze stir laneâ restless vigil kept pillow sleep could gain poppy norâ sure opiatesâ
found warm moonlit air man boat upon sand drowsy drowsily boat put sea passed hour two perchance year thought sense vanished engulfing trance vast

y=absurdity (probability **0.997**, score **4.621**) top features

Contribution?	Feature
+3.997	Highlighted in text (sum)
+0.625	<BIAS>

church ken living test heart face forgiving show grace lisp pretty halting curtsey bid good day fresh defaulting wash old away praise man god guerdon love
grace let satan claim hand boat mystery yester eve thing sleptâ scarce breeze stir laneâ restless vigil kept pillow sleep could gain poppy norâ sure opiatesâ
found warm moonlit air man boat upon sand drowsy drowsily boat put sea passed hour two perchance year thought sense vanished engulfing trance vast

This record is from a nihilism book, however it was clustered in the absurdity group, can we tell why?

If we look at the 2nd cluster interpretation (the absurdity). The top "positive" contributing and frequent words in this example are words like "face", "day", "year" normal words that are daily used in general that were abundant and frequent in the absurdity cluster. Nonetheless, when we look at the top deterring negative words we can see words like "god", "praise" which are the spiritual and divine words that are frequent in nihilism.

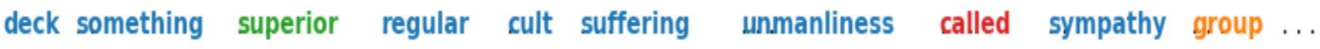
Then, why did it choose Absurdity over Nihilism, the overall number of words –in this record- that account for absurdity are way larger than the ones accounting for nihilism.

Another technique in using LDA, the same color codes are used.

Here what happens is that for the misclassified record, each word is colored with the color of the cluster it accounts for.

This record was misclassified as Nihilism, we can see how several words account for nihilism, which is also plausible (suffering, sympathy and unmanliness), only one word accounted for existentialism (superior),…. And so on.

The red color accounts for the 4th topic that we claimed is a confusion area between existentialism and absurdity.



deck something superior regular cult suffering unmanliness called sympathy group . . .

Meanwhile, this method needs more investigation and work that due to the scarcity of time we couldn't provide, thus we aren't relying on it very much yet.

4. Conclusion and Findings

Machine learning isn’t a rigid science, it can be used to solve our own problems not just major business problems. And this is the beauty of machine learn, the adaptability it encompasses.

In this assignment, we tried to answer this question: Is there a significant boundary between the 3 famous philosophies that we sometimes get confused between? and we reached an answer that maybe considered as an edge response → ‘to some extent’, nihilism seemed to be distinct from the other two philosophies, however absurdity and existentialism have much overlapping and this may induce another question inside us to investigate -afterwards- the distinction between both.

7. References

[1] classification - Cohen's kappa in plain English - Cross Validated (stackexchange.com)

[2] Silhouette (clustering) - Wikipedia

[3] https://arxiv.org/pdf/2003.01670.pdf

[4] bmabey/pyLDAvis: Python library for interactive topic model visualization. Port of the R LDAvis package. (github.com)