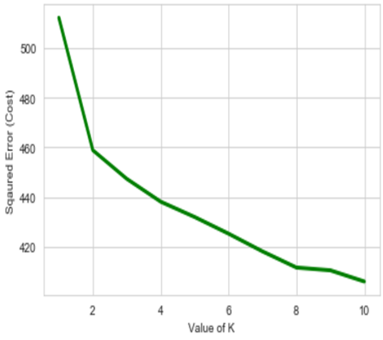
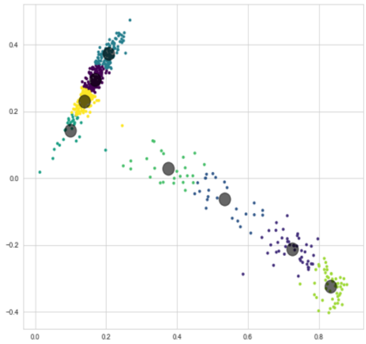
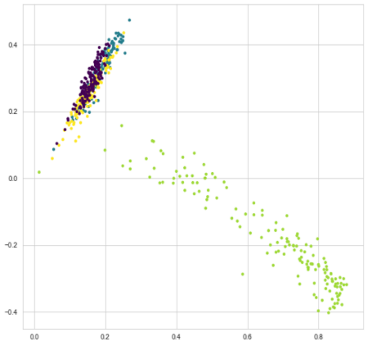
Unsupervised ML-Clustering Analysis

2D Clustering Analysis of Scatter Plot

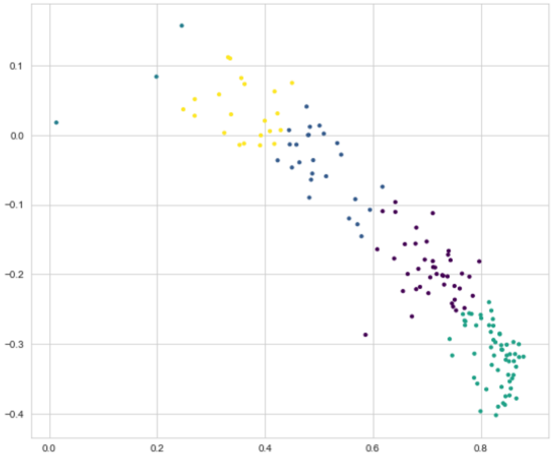
Hook from Mentor's session

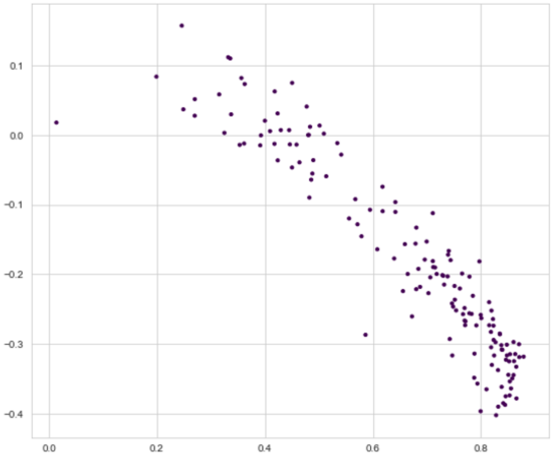


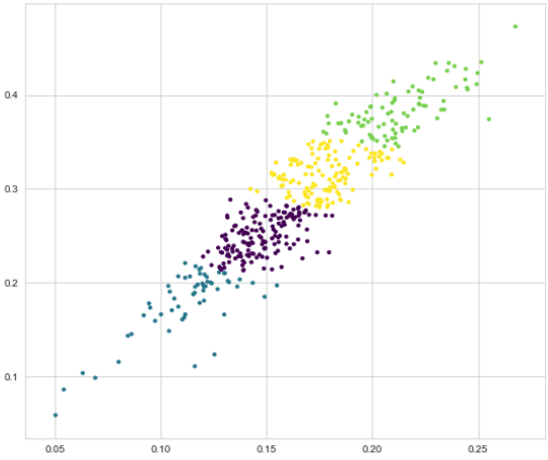


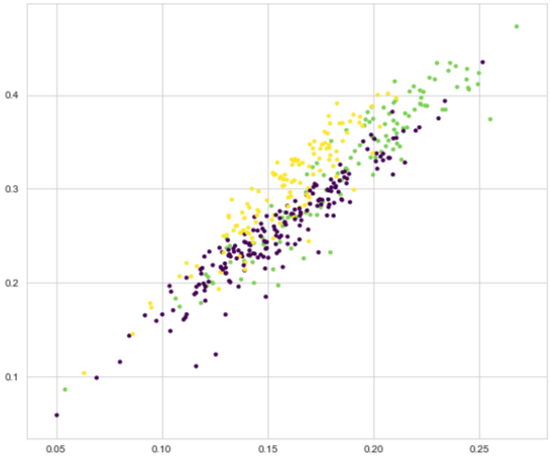


Pick out the special one



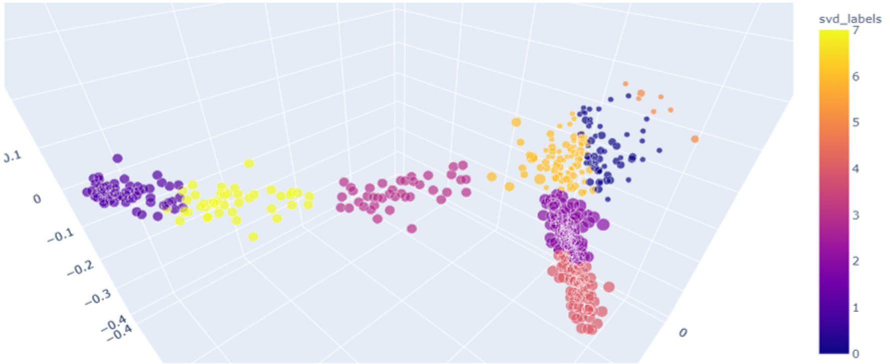


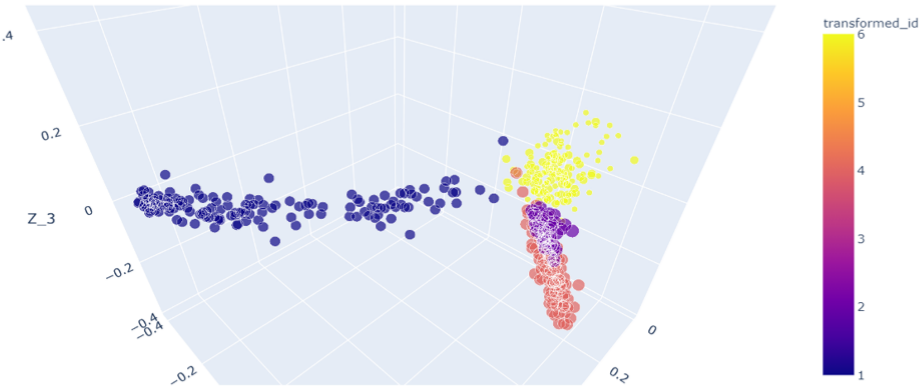


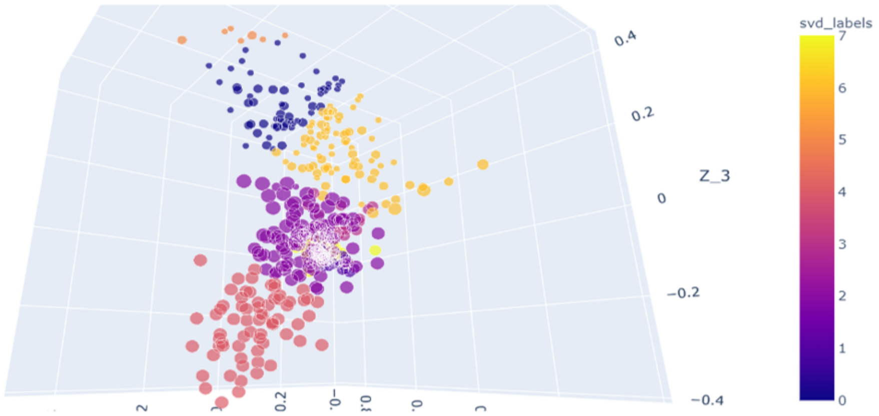


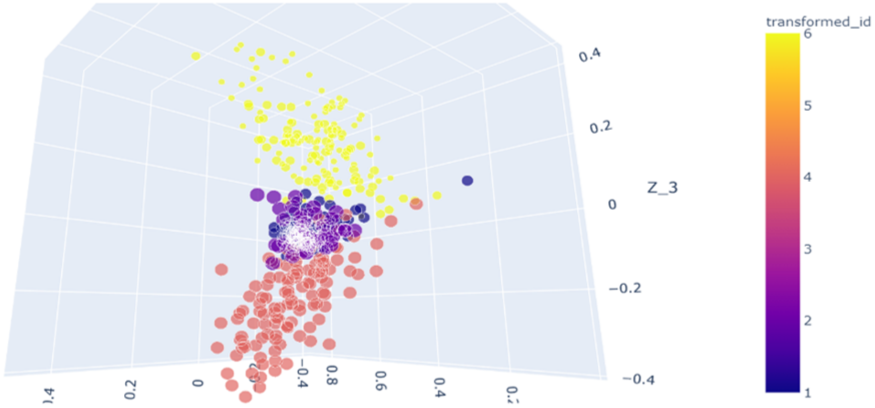
3D Clustering Analysis of Scatter Plot

Plotly Method



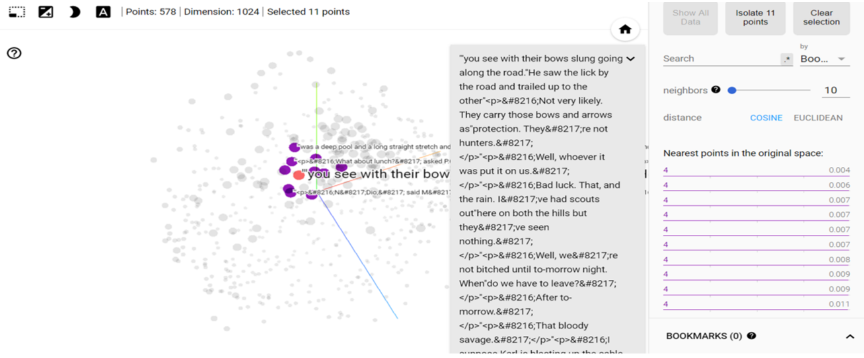






Embedding Projector Method

Find more about the fourth book --- Green Hills of Africa (1935):



| Points: 578 | Dimension: 1024 | Selected 11 points

"you see with their bows slung going along the road."He saw the lick by the road and trailed up to the other"<p>&#8216;Not very likely. They carry those bows and arrows as"protection. They&#8217;re not hunters.&#8217;</p>"<p>&#8216;Well, whoever it was put it on us.&#8217;</p>"<p>&#8216;Bad luck. That, and the rain. I&#8217;ve had scouts out"here on both the hills but they&#8217;ve seen nothing.&#8217;</p>"<p>&#8216;Well, we&#8217;re not bitched until to-morrow night. When"do we have to leave?&#8217;</p>"<p>&#8216;After to-morrow.&#8217;</p>"<p>&#8216;That bloody savage.&#8217;</p>"<p>&#8216;I

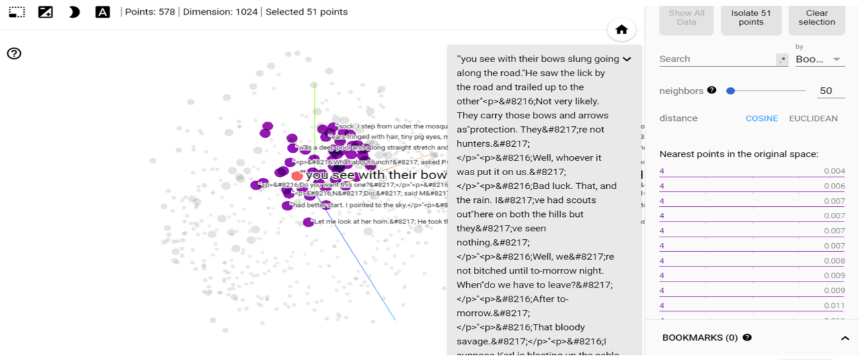
neighbors 10

distance COSINE EUCLIDEAN

Nearest points in the original space:

4	0.004
4	0.006
4	0.007
4	0.007
4	0.007
4	0.007
4	0.008
4	0.009
4	0.009
4	0.011

Show All Data
Isolate 11 points
Clear selection
Search
by
Boo...
"was a deep pool and a long straight stretch and
<p>&#8216;What about lunch?&#8217; asked P
"you see with their bow
<p>&#8216;N&#8217;Dio.&#8217; said M&#821
BOOKMARKS (0)



| Points: 578 | Dimension: 1024 | Selected 51 points

"you see with their bows slung going along the road."He saw the lick by the road and trailed up to the other"<p>&#8216;Not very likely. They carry those bows and arrows as"protection. They&#8217;re not hunters.&#8217;</p>"<p>&#8216;Well, whoever it was put it on us.&#8217;</p>"<p>&#8216;Bad luck. That, and the rain. I&#8217;ve had scouts out"here on both the hills but they&#8217;ve seen nothing.&#8217;</p>"<p>&#8216;Well, we&#8217;re not bitched until to-morrow night. When"do we have to leave?&#8217;</p>"<p>&#8216;After to-morrow.&#8217;</p>"<p>&#8216;That bloody savage.&#8217;</p>"<p>&#8216;I

neighbors 50

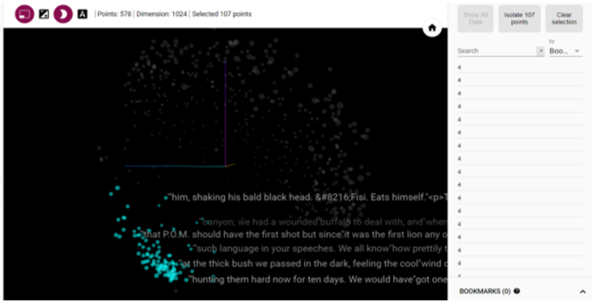
distance COSINE EUCLIDEAN

Nearest points in the original space:

4	0.004
4	0.006
4	0.007
4	0.007
4	0.007
4	0.007
4	0.008
4	0.009
4	0.009
4	0.011

Show All Data
Isolate 51 points
Clear selection
Search
by
Boo...
"sock. I step from under the mosqu
"ears fringed with hair, tiny pig eyes, m
"was a deep pool and a long straight stretch and
<p>&#8216;What about lunch?&#8217; asked P
"you see with their bow
<p>&#8216;Do you want this one?&#8217;</p>"<p>&#8216
<p>&#8216;N&#8217;Dio.&#8217; said M&#821
"had better start. I pointed to the sky.</p>"<p>&#8
"Let me look at her horn.&#8217; He took th
BOOKMARKS (0)

Randomly select points and analyze



| Points: 578 | Dimension: 1024 | Selected 107 points

"him, shaking his bald black head. &#8216;Fisi. Eats himself."<p>

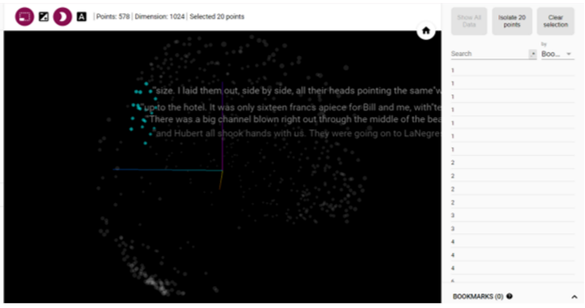
"canyon; we had a wounded"buffalo to deal with, and"when

"that P.O.M. should have the first shot but since"it was the first lion any o

"such language in your speeches. We all know"how prettily t

"at the thick bush we passed in the dark, feeling the cool"wind o

"hunting them hard now for ten days. We would have"got one



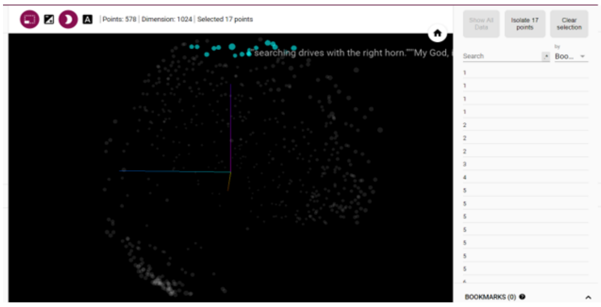
| Points: 578 | Dimension: 1024 | Selected 20 points

"size. I laid them out, side by side, all their heads pointing the same"w

"up to the hotel. It was only sixteen francs apiece for Bill and me, with"te

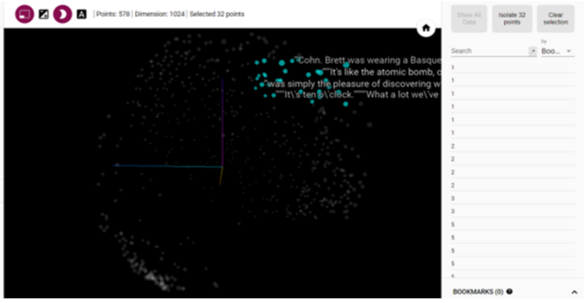
"There was a big channel blown right out through the middle of the bea

"and Hubert all shook hands with us. They were going on to LaNegre



| Points: 578 | Dimension: 1024 | Selected 17 points

"searching drives with the right horn.""My God,



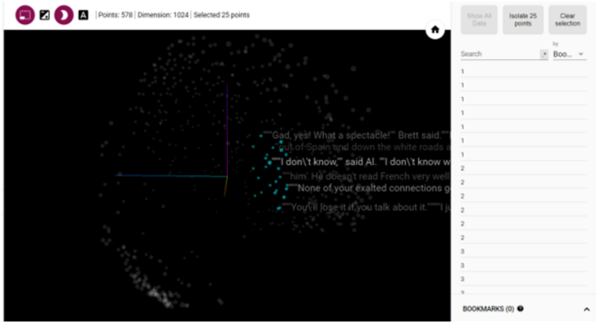
| Points: 578 | Dimension: 1024 | Selected 32 points

"Cohn. Brett was wearing a Basque

""It's like the atomic bomb, c

"was simply the pleasure of discovering w

""It\'s ten o\'clock.""What a lot we\'ve



| Points: 578 | Dimension: 1024 | Selected 25 points

""Gad, yes! What a spectacle!" Brett said.""

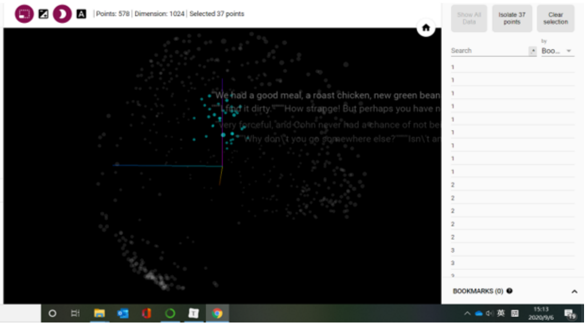
"out of Spain and down the white roads a

""I don\'t know," said Al. "I don\'t know w

"him. He doesn\'t read French very well

""None of your exalted connections g

""You\'ll lose it if you talk about it.""I ju



| Points: 578 | Dimension: 1024 | Selected 37 points

"We had a good meal, a roast chicken, new green bean

"I find it dirty.""How strange! But perhaps you have n

"very forceful, and Cohn never had a chance of not be

""Why don\'t you go somewhere else?""Isn\'t an

Summary of Clustering Analysis above

I will introduce and explain all the plots from top to bottom, left to right.

Basically, I want to make a 3D plot in Plotly of the Unsupervised Machine Learning technique, here I determine to apply clustering analysis for texts of six books from Hemingway.

The plots above start from the method in mentor session. After we gaining needed data by TextBlob package, Tf-idf method in sklearn, we determine to set number of clusters in K-means method equal to eight since we can clearly see an 'elbow' in the first graph. I use surjection to get associated id and transform_id, to avoid overlapping.

According to the 2D plots, the left one clearly shows eight clusters with separated centers, which means the machine can learn to pick similar sentences into one cluster no matter which book it belongs to. However, if we look at the right graph, it only has four clusters which means, one transformed_id contains four different clusters and others are combined and hard to be separated.

After plotting each book, I find the fourth book, Green Hills of Africa(id = 3), is the special one, and the texts of it dominate the whole right-bottom part. As for other five books, we find they contain some texts corresponding to the left graph. Additionally, It also illustrates that only four books are clear enough and rational to use K-means Clustering method.

As I get high dimensional text vectors initially, I consider to use PCA method to transform it to three-dimensional data to avoid losing more information. Hence, 3D plots here are more precise and meaningful. For the whole shape, if we consider the projection of that, we find it kind of similar to 2D plots above. However, the key feature is that it can be divided into six clusters with respect to different transformed_id rather than four. Although each cluster does not match the correct book wholly, it contains parts of texts belonging to the true book.

In addition, to make the 3D plot more accurate and meaningful, I consider the embeddings in the text, since recognizing certain relationship between grams contributes to the Clustering Analysis. Here I use Embedding Projector, a tool in Tensorflow, to get the dynamic 3D plot. I provide some representative screenshots to better understand the Unsupervised ML to text relations among these six books.

From light style, I focus on how bonded the texts of the fourth book are. The two graphs show that no matter 10 nearest points or 50 nearest points from the text of Green Hills of Africa, most of them also belong to this book.

From dark style, we can see the clear two opposite parts of the whole 'ball' shape. Unsurprisingly, as I select several points in the first graph, we see nearly all of them belong to the fourth book. However, as I select different small parts of the other side, other five pictures all show that the highlight points are labeled for several books, not one or two books. I can also claim that, even considering embeddings, the cluster results do not change a lot. Overall, except Green Hills of Africa, ML-generated clusters are not totally corresponding to the clusters of real books.