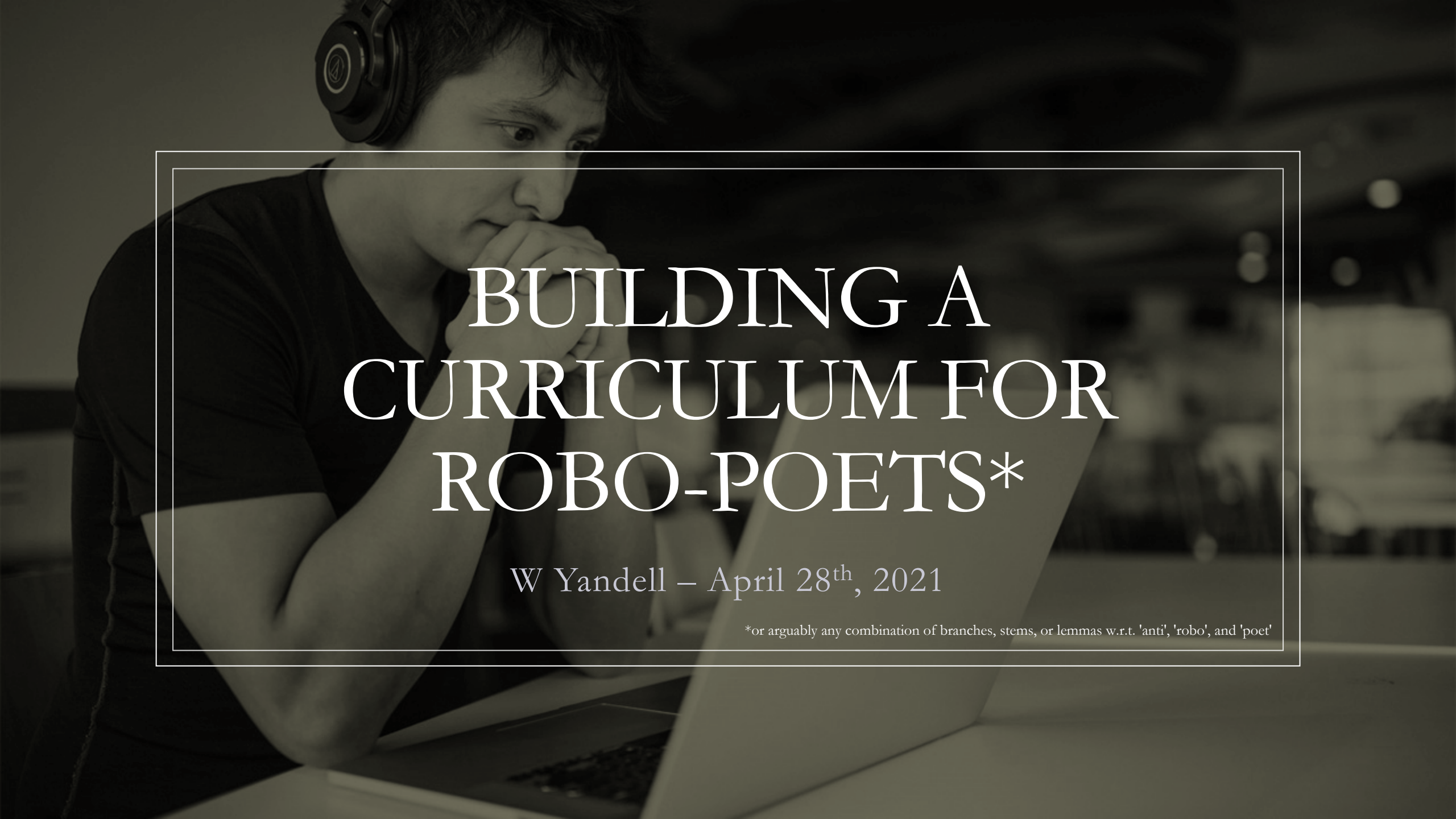


BUILDING A CURRICULUM FOR ROBO-POETS*

W Yandell – April 28th, 2021

*or arguably any combination of branches, stems, or lemmas w.r.t. 'anti', 'robo', and 'poet'

Part 1. Course Outline

Lesson 1.

For simplicity, we shall call a String instance which has been classified as a poem a 'poem string'.

Lesson 2.

Most poem strings are very messy from one point of view or another—you must reform them or they could be incomparable.

Lesson 3.

Even reformed, these still complex structures should be converted into tidier numeric representations.

Lesson 4.

Patterns can be deduced across a set of these vectors wherein repetitions and differences in the construction and meanings may be rendered.

Lesson 5.

These patterns and relationships can be compressed into clusters according to distance such that if you have generated a poem string, it should belong.

Raw Data

- Source: Poetryfoundation.org via Kaggle
  - https://www.kaggle.com/johnhallman/complete-poetryfoundationorg-dataset
- Total Instances: 15652
- Feature: # Unique
  - Author: 3310
  - Title: 14997
  - Content: 15638

	author	title	content
0	Wendy Videlock	!	Dear Writers, I'm compiling the first in what ...
1	Hailey Leithauser	0	Philosophic\nin its complex, ovoid emptiness,\...
2	Jody Gladding	1-800-FEAR	We'd like to talk with you about fear t...
3	Joseph Brodsky	1 January 1965	The Wise Men will unlearn your name.\nAbove yo...
4	Ted Berrigan	3 Pages	For Jack Collom\n10 Things I do Every Day\n\np...
...	...	...	...
15647	Hannah Gamble	Your Invitation to a Modest Breakfast	It's too cold to smoke outside, but if you com...
15648	Eleni Sikelianos	Your Kingdom\n \n \n \n Launch Audio in a N...	if you like let the body feel\nall its own evo...
15649	Susan Elizabeth Howe	"Your Luck Is About To Change"	(A fortune cookie)\nOminous inscrutable Chines...
15650	Andrew Shields	Your Mileage May Vary	1\nOur last night in the house was not our las...
15651	Joseph O. Legaspi	Your Mother Wears a House Dress	If your house\nis a dress\nit'll fit like\nLos...

15652 rows × 3 columns

State of the effort

Step 1-N

Clean & Tokenize

Step 2

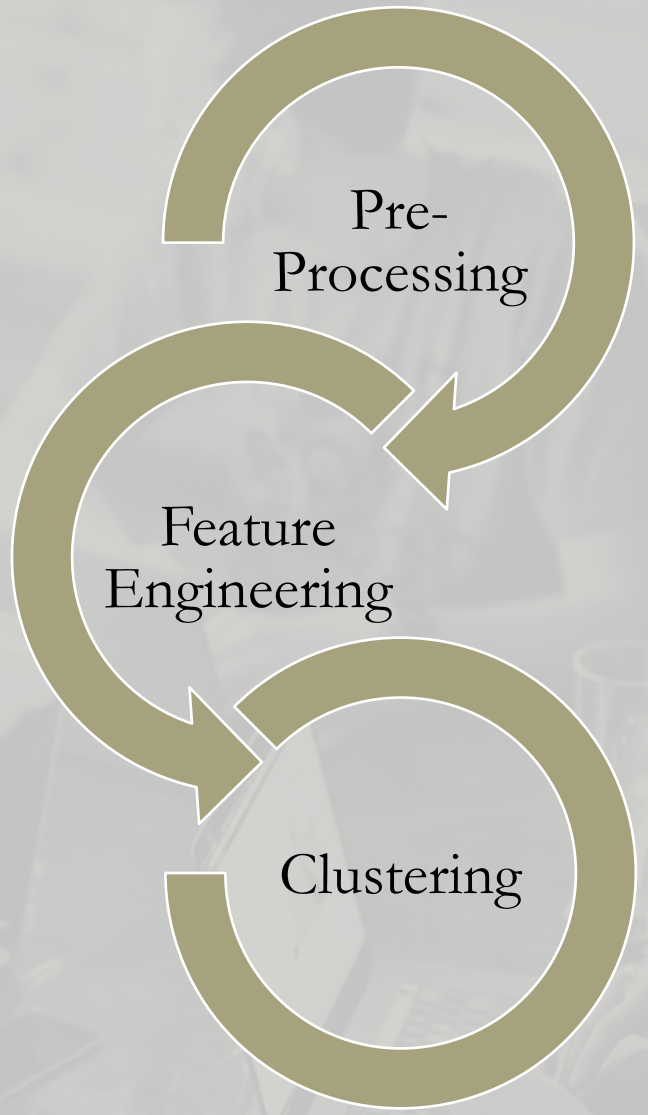
Vectorize with Doc2Vec

Step 1-M

Aggregate & Engineer

Step 3-4

Unsupervised Learning



Processing

Non-standard cleaning:

- Line structure retention
  - Delayed filtering
  - No sentence tokenization
- Grammar/style retention
  - No stemming
  - No lemmatizing
  - No stop-word removal

content	line	words
Philosophic\nin its complex, ovoid emptiness,\...	[philosophic, in its complex ovoid emptiness, ...	[philosophic, in, its, complex, ovoid, emptine...
We'd like to talk with you about fear t...	[we'd like to talk with you about fear they sa...	[we'd, like, to, talk, with, you, about, fear,...
The Wise Men will unlearn your name.\nAbove yo...	[the wise men will unlearn your name, above yo...	[the, wise, men, will, unlearn, your, name, ab...
For Jack Collom\n10 Things I do Every Day\n\np...	[for jack collom, things i do every day, play ...	[for, jack, collom, things, i, do, every, day,...
WINTER\nMore time is spent at the window.\n\nS...	[winter, more time is spent at the window, sum...	[winter, more, time, is, spent, at, the, windo...

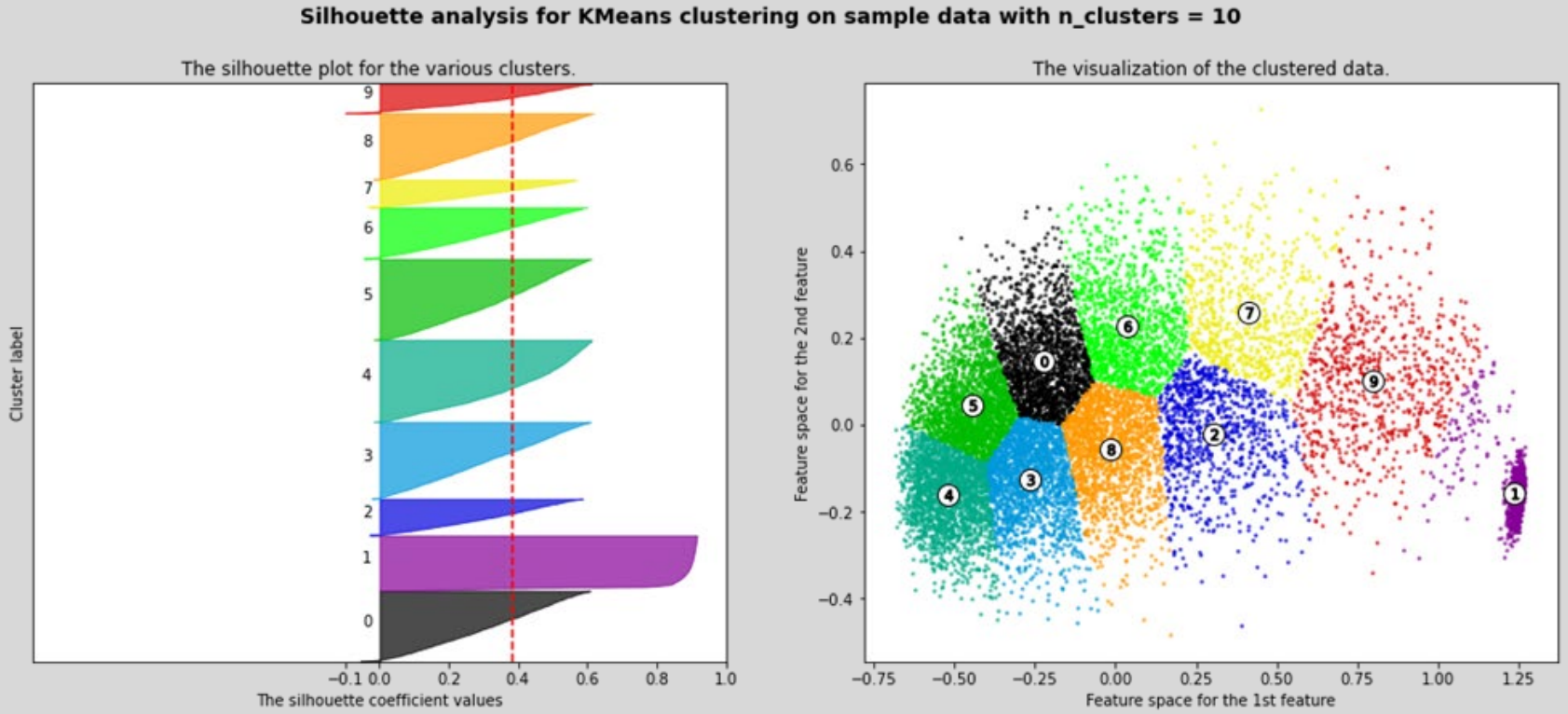
Aggregation/ Outlier Removal

- Lengths – in/of lines, words, characters
- Lexical diversity – unique/total
- Frequency Distributions
  - Word length – Mendenhall
  - Part-of-speech

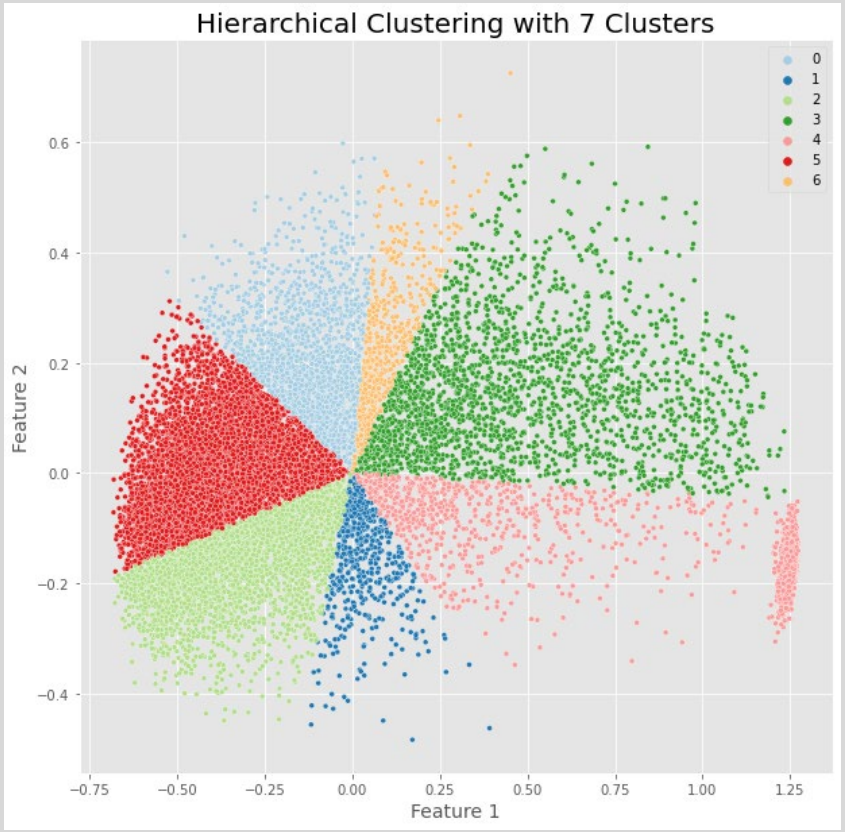
length_in_lines	lexical_diversity	word_lengths	max_word_length	pos_tags
15	0.863636	[11, 2, 3, 7, 5, 9, 1, 8, 6, 6, 2, 2, 1, 4, 2,...	11	[JJ, IN, PRP$, JJ, JJ, NN, DT, JJ, NN, VBD, PR...
11	0.644068	[4, 4, 2, 4, 4, 3, 5, 4, 4, 4, 2, 4, 6, 4, 2, ...	10	[PRP, MD, VB, TO, VB, IN, PRP, IN, NN, PRP, VB...
24	0.679739	[3, 4, 3, 4, 7, 4, 4, 5, 4, 4, 2, 4, 4, 5, 3, ...	12	[DT, JJ, NNS, MD, VB, PRP, ***NN***, ***IN***, ***PRP***, NN, ...
26	0.848101	[3, 4, 6, 6, 1, 2, 5, 3, 4, 5, 5, 4, 5, 3, 4, ...	13	[IN, JJ, NN, NNS, PRP, VBP, DT, NN, VB, NN, VB...
65	0.575843	[6, 4, 4, 2, 5, 2, 3, 6, 6, 3, 2, 5, 4, 3, 2, ...	11	[NN, JJR, NN, VBZ, VBN, IN, DT, NN, NN, PRP, V...

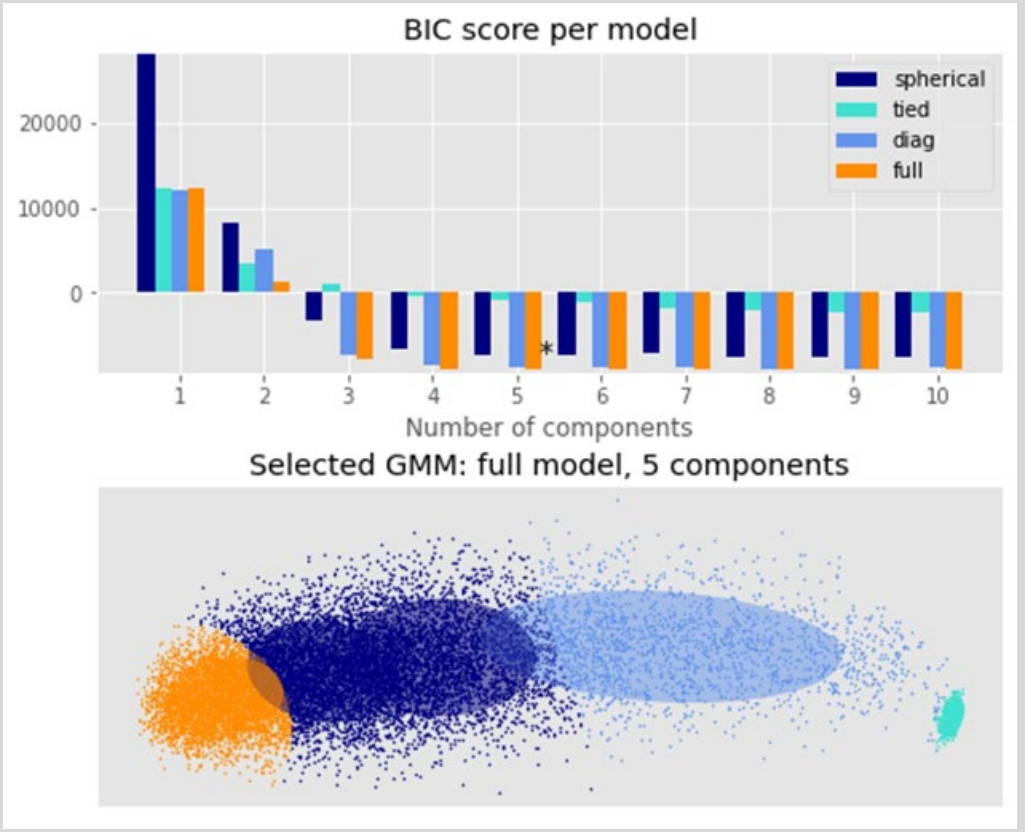
Note: only document level dataset shown (which has different filters) with abridged featureset

Analysis – K-Means Clustering

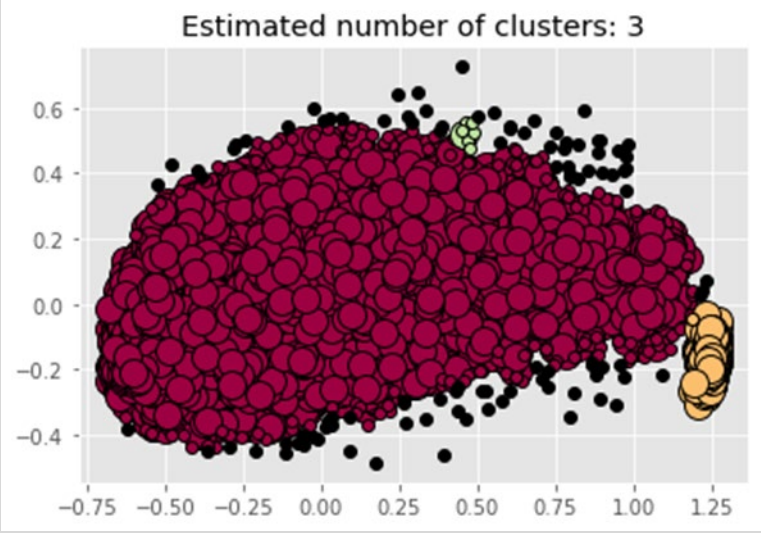


Analysis – Hierarchical (cosine) & GMM

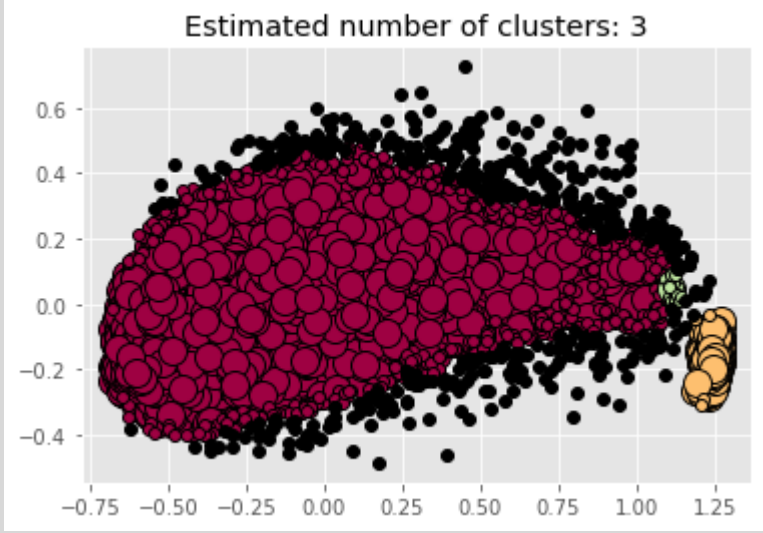




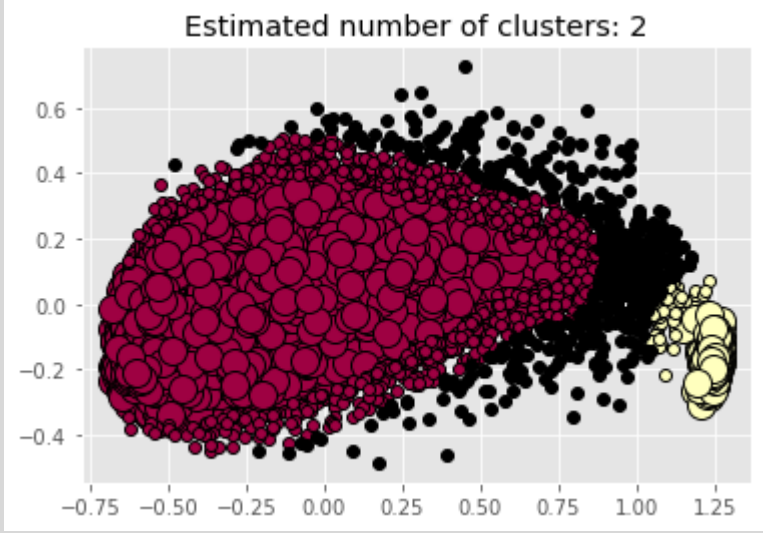
Analysis – DBSCAN



epsilon = 0.05
minimum samples = 10
n_noise = 103
silhouette score = 0.395



epsilon = 0.05
minimum samples = 25
n_noise = 478
silhouette score = 0.44



epsilon = 0.12
minimum samples = 200
n_noise = 542
silhouette score = 0.539

Conclusion

- More processing than anything
- Challenges everywhere
- Still a ways to go
- Stay tuned for multiple DL models in the final


THANK YOU!