“

(Hello Everyone!)

NLP Analysis of Latin Literature

Presented by Amy Zhou

Overview

Goal: Identify patterns in Latin literature for text classification

Data Collection:

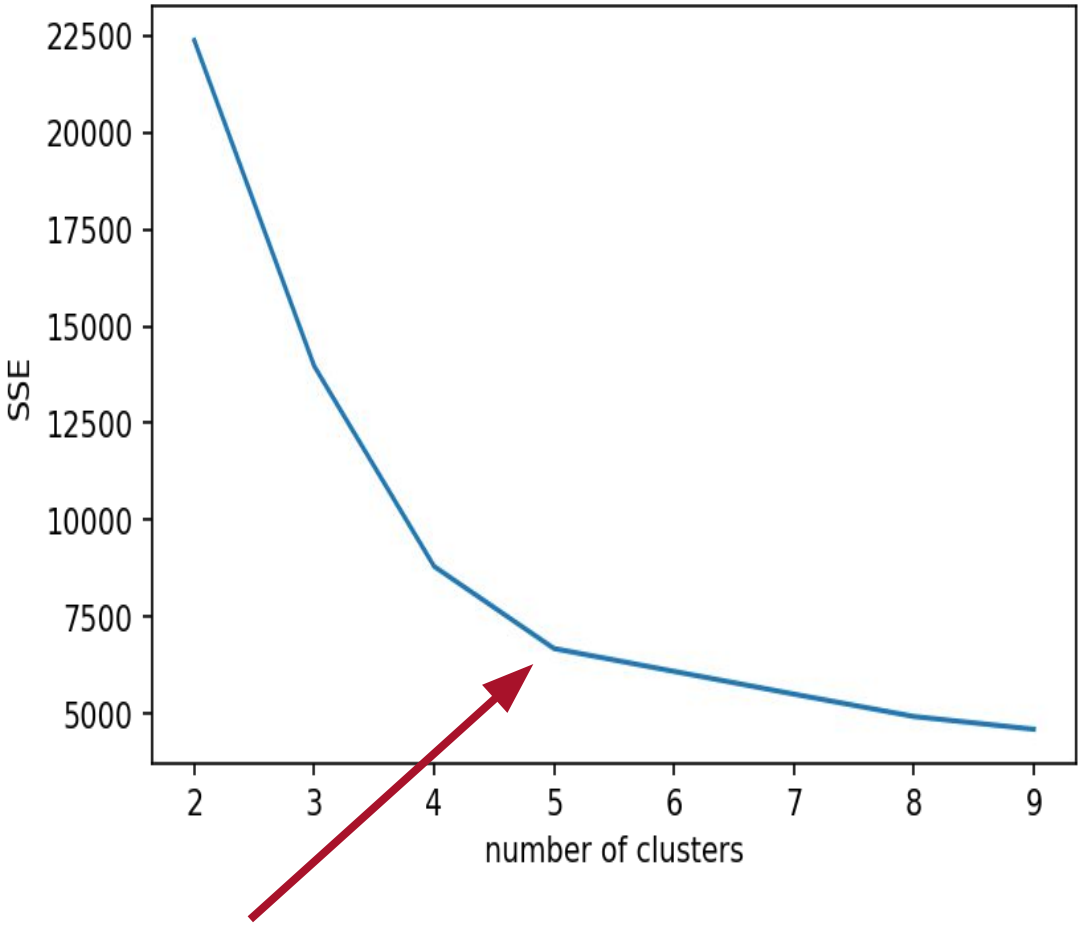
- Open source texts from Perseus
- Classical, medieval, neo-Latin
- Split texts into 5 sentences each

Methodology:

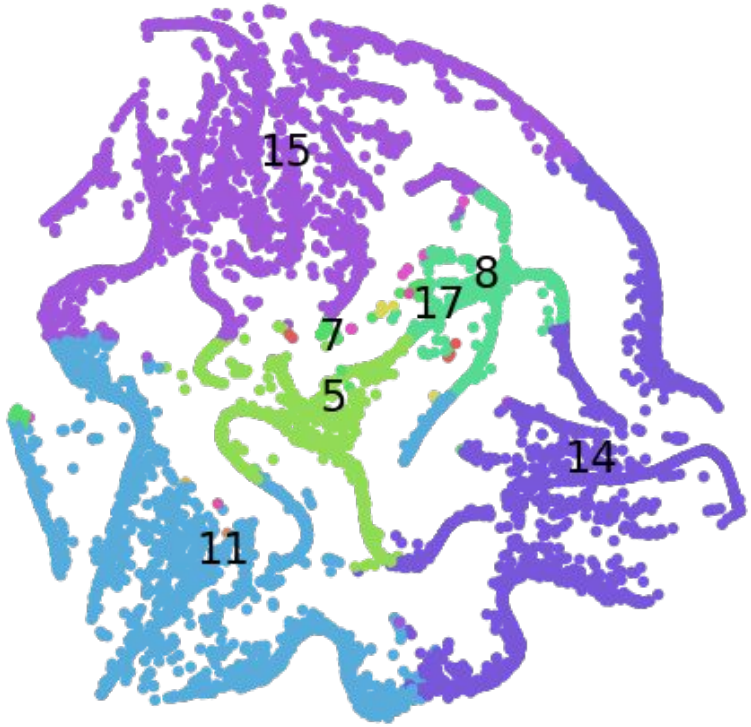
- Parse with Classical Language Toolkit
- Feature extraction **(Td-idf)**
- Dimensionality reduction with topic modelers **(LDA)**
- Clustering **(K-means)**

Example Topics:

K-means inertia

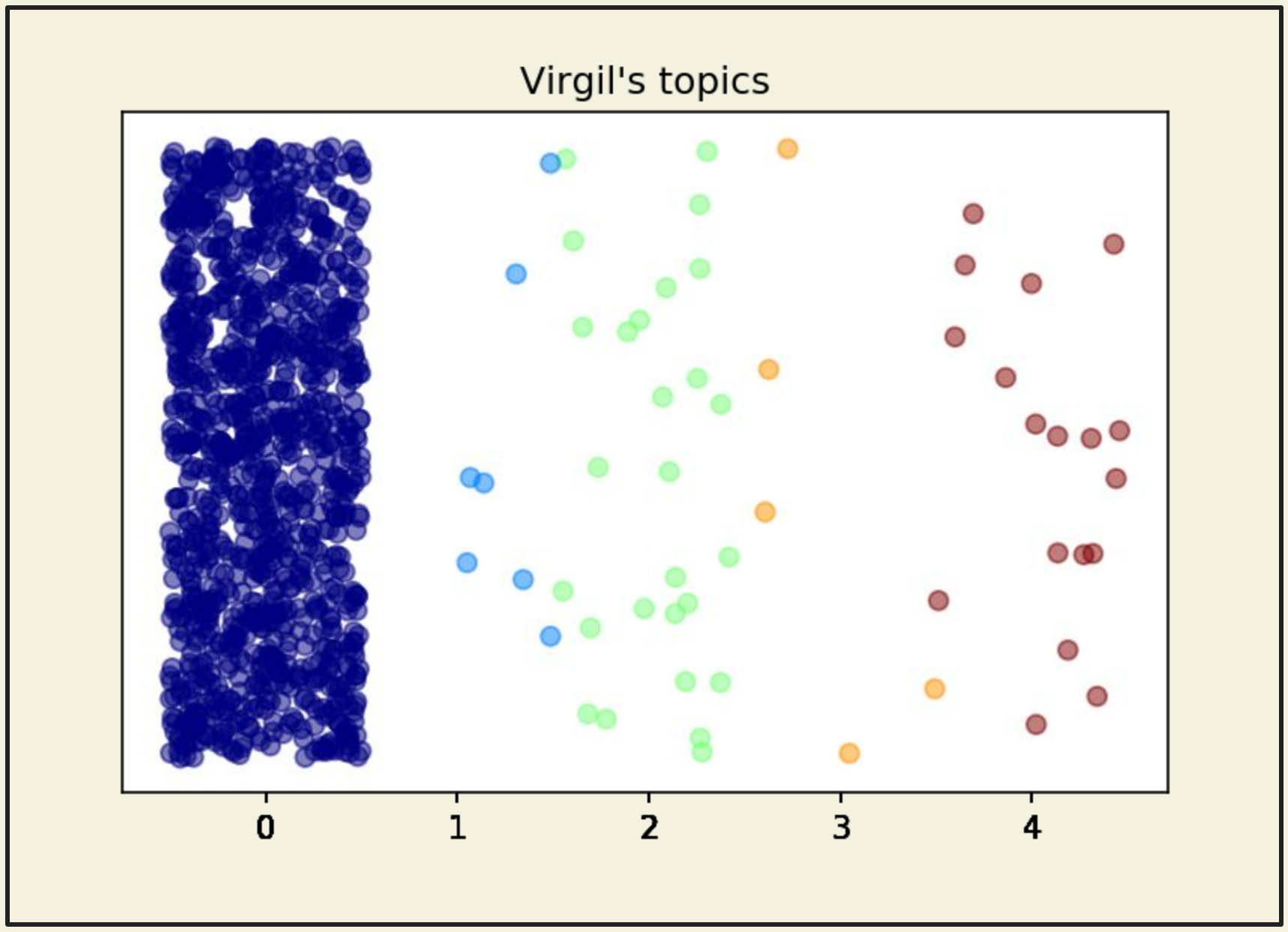


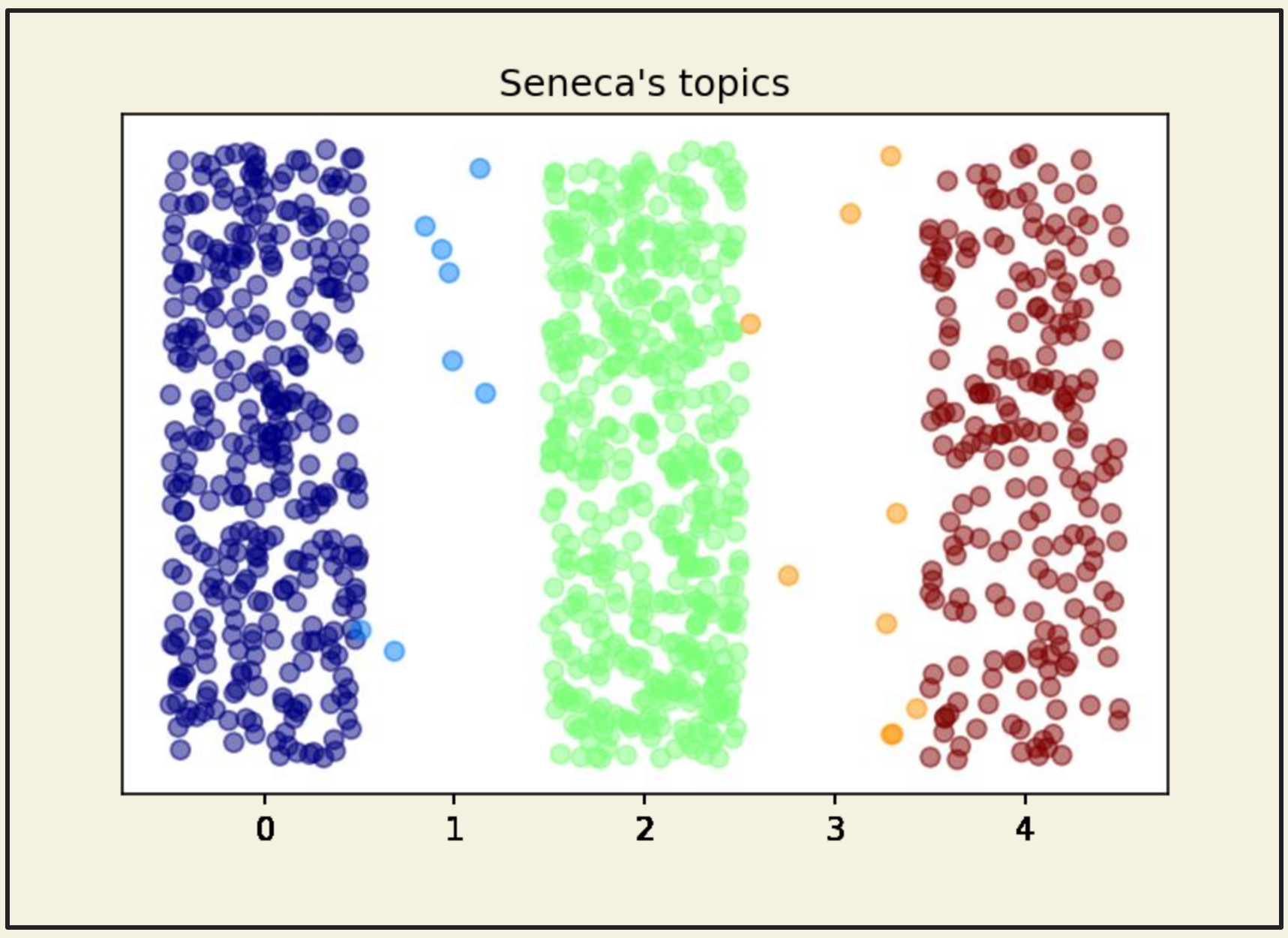
t-SNE clusters

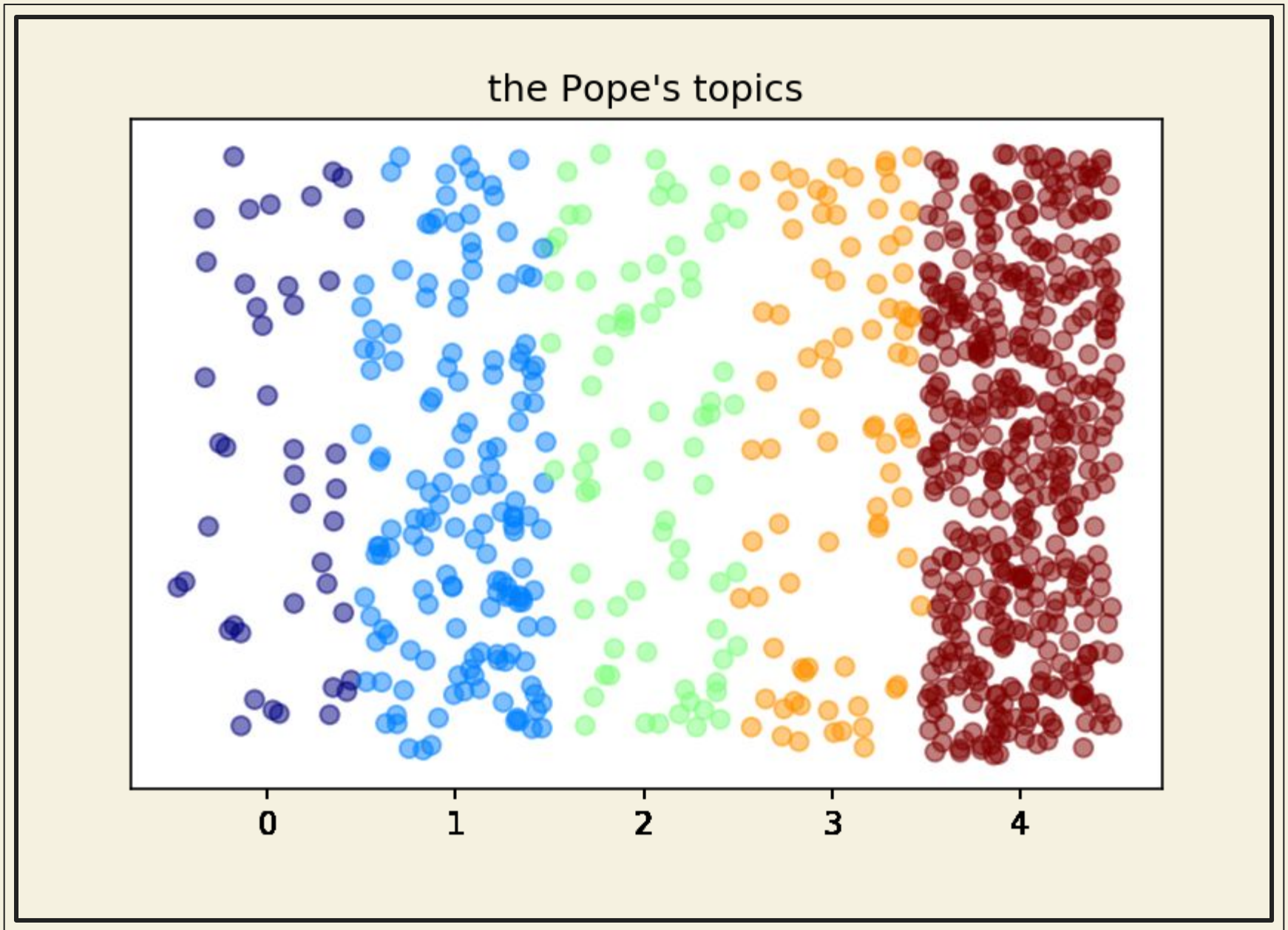


Selected Model Pipeline:

Tf-idf, LDA, K-means (5 clusters)

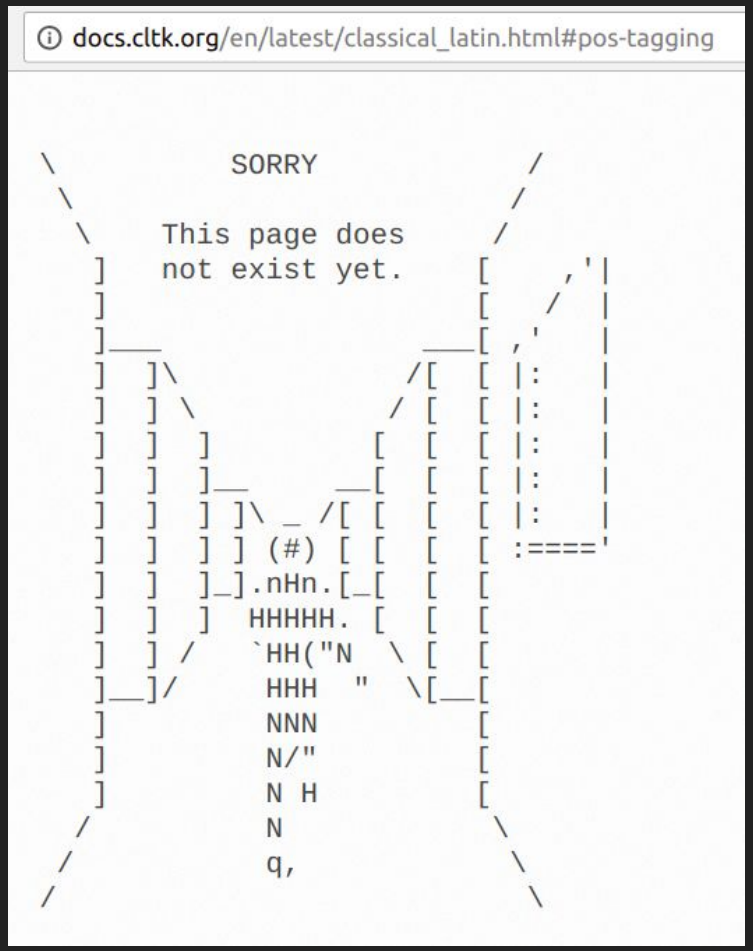

Virgil's topics
0
1
2
3
4


Seneca's topics
0
1
2
3
4


the Pope's topics
0
1
2
3
4

Limitations → Next Steps

- Stem/Lemma
  - Irregular forms (poetry)
  - Typos in transcription
- Prose vs poetry
  - Using rhythm identifier
- POS style vs words
  - Order matters
  - Highly referential writing
- Sophisticated text generator
  - Markov + Word2Vec



Gratias tuis ago! (Thank you!)