ETDs (Electronic Theses and Dissertations) Recommendation System

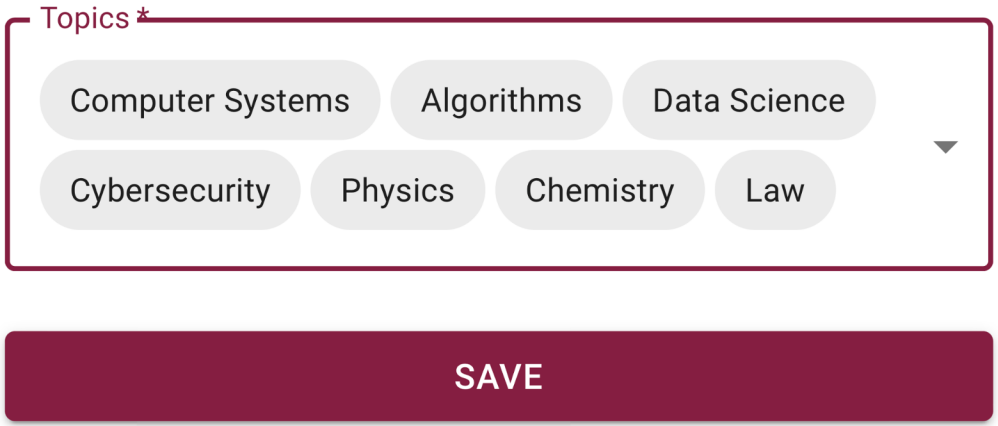
Talia Blakemore and Long Phan

Outline

1. Overview of Project
2. Timeline
3. General Approach
4. Deliverables
5. Testing & Evaluation
6. Acknowledgements
7. References

Overview

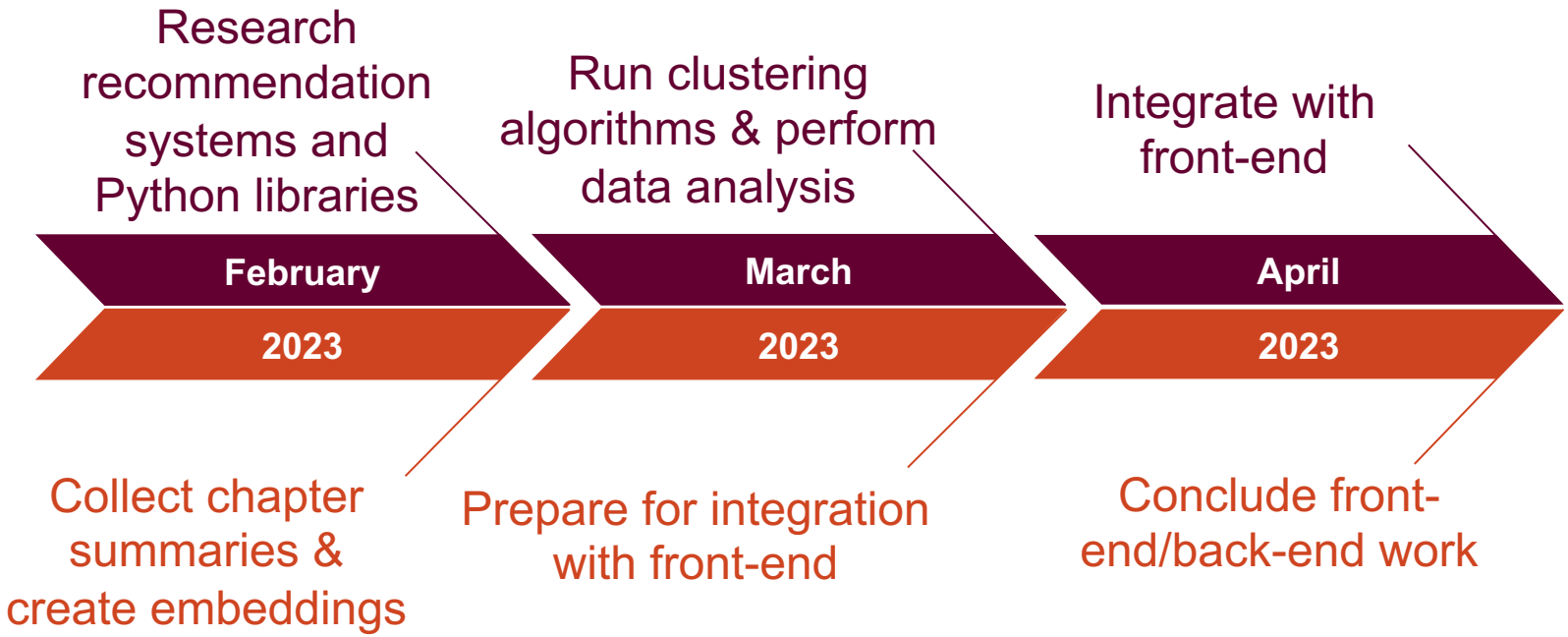
- Client: Satvik Chekuri
- Current System:
  - Recommendation system is built upon interests supplied by the user
  - Uses document titles and abstracts to evaluate similarity between documents



Overview

- We want to use **chapter summaries** to evaluate similarity
  - More efficient and accurate
  - Summaries are based on narrower topics
- Previous capstone group has created chapter summaries for over 5000 ETDS, resulting in over ~20,000 chapter summaries
  - We are using 5000 summaries

Timeline


Research recommendation systems and Python libraries
February
2023
Collect chapter summaries & create embeddings
Run clustering algorithms & perform data analysis
March
2023
Prepare for integration with front-end
Integrate with front-end
April
2023
Conclude front-end/back-end work

General Approach

Retrieve chapter summaries from summarization API

General Approach

Create embeddings using Sentence-Transformers Package

General Approach

Perform dimensionality reduction

General Approach

User performs a search, specifying

- Clustering algorithm
- Number of documents to cluster, *n*
- Number of clusters/Minimum Cluster size
- Whether to use PCA

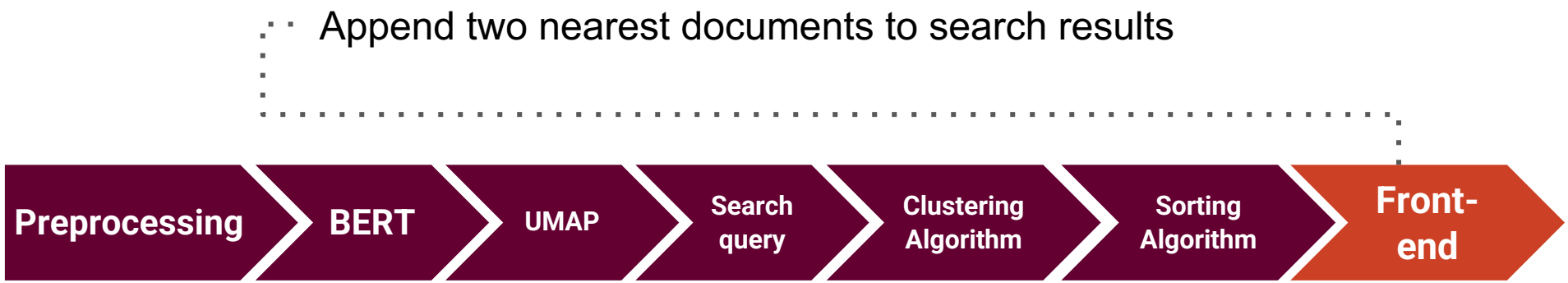
General Approach

KMeans, DBScan or HDBScan using user specifications

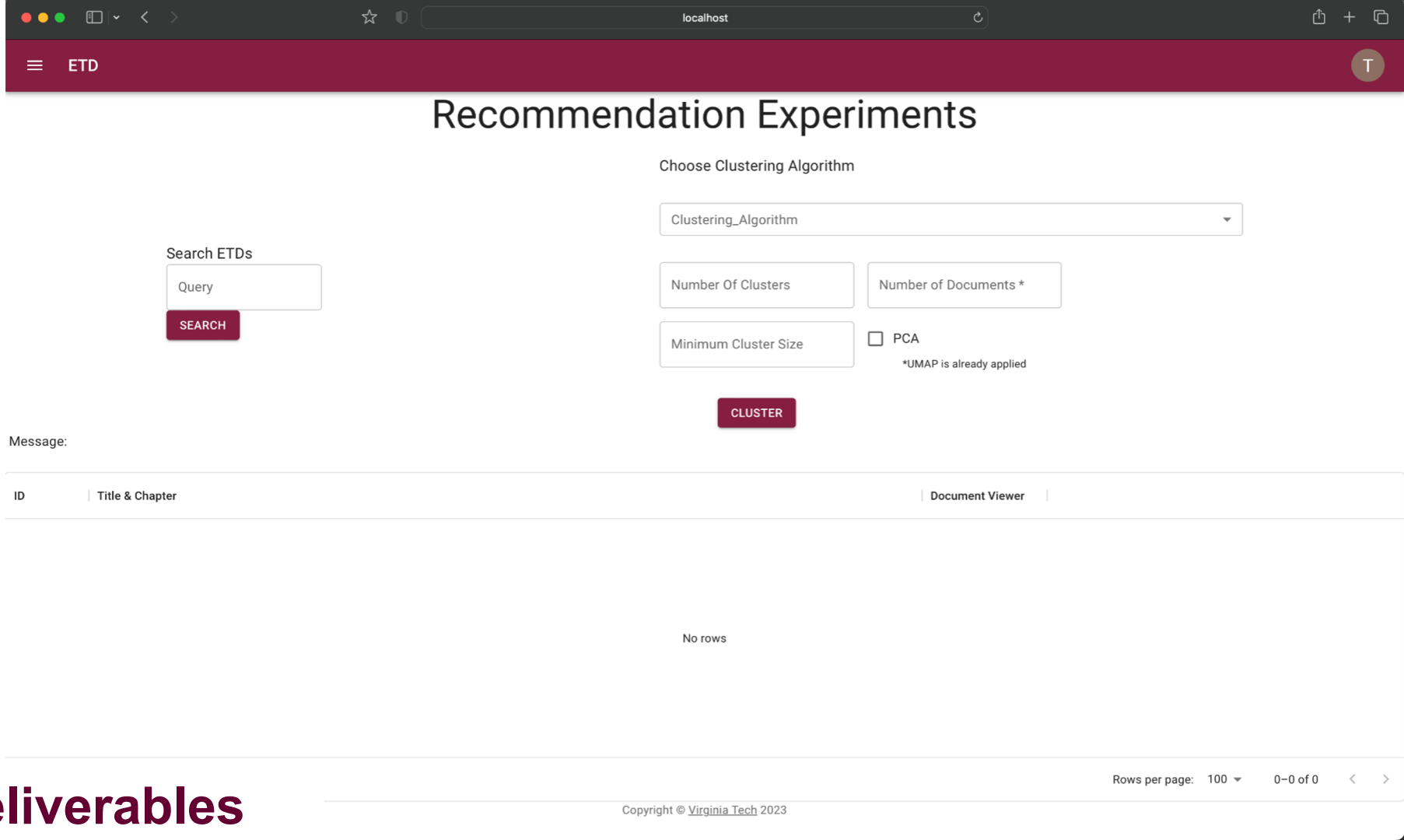
General Approach

Sort output of each cluster to retrieve two nearest results to first *n* chapters

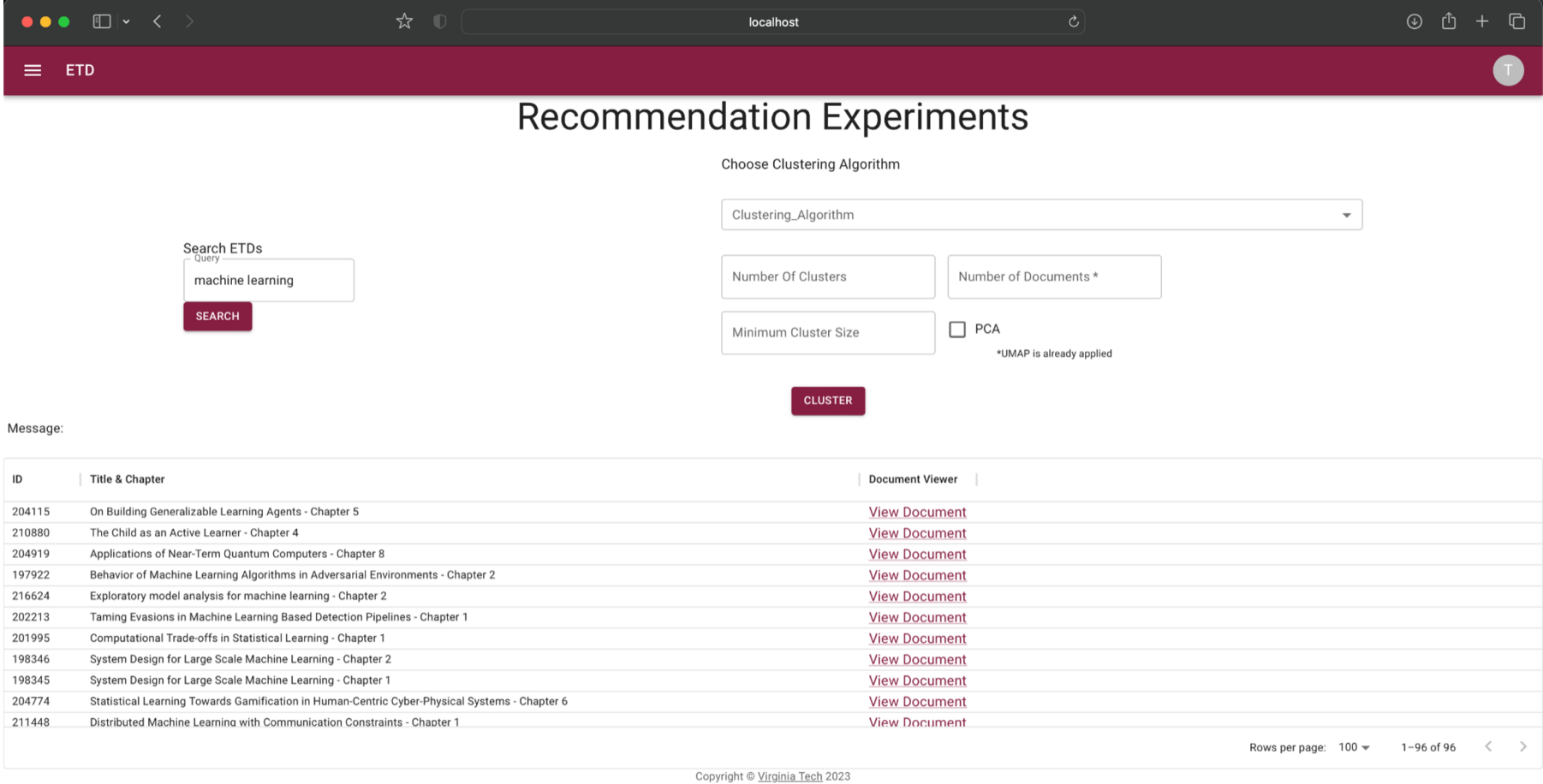
General Approach



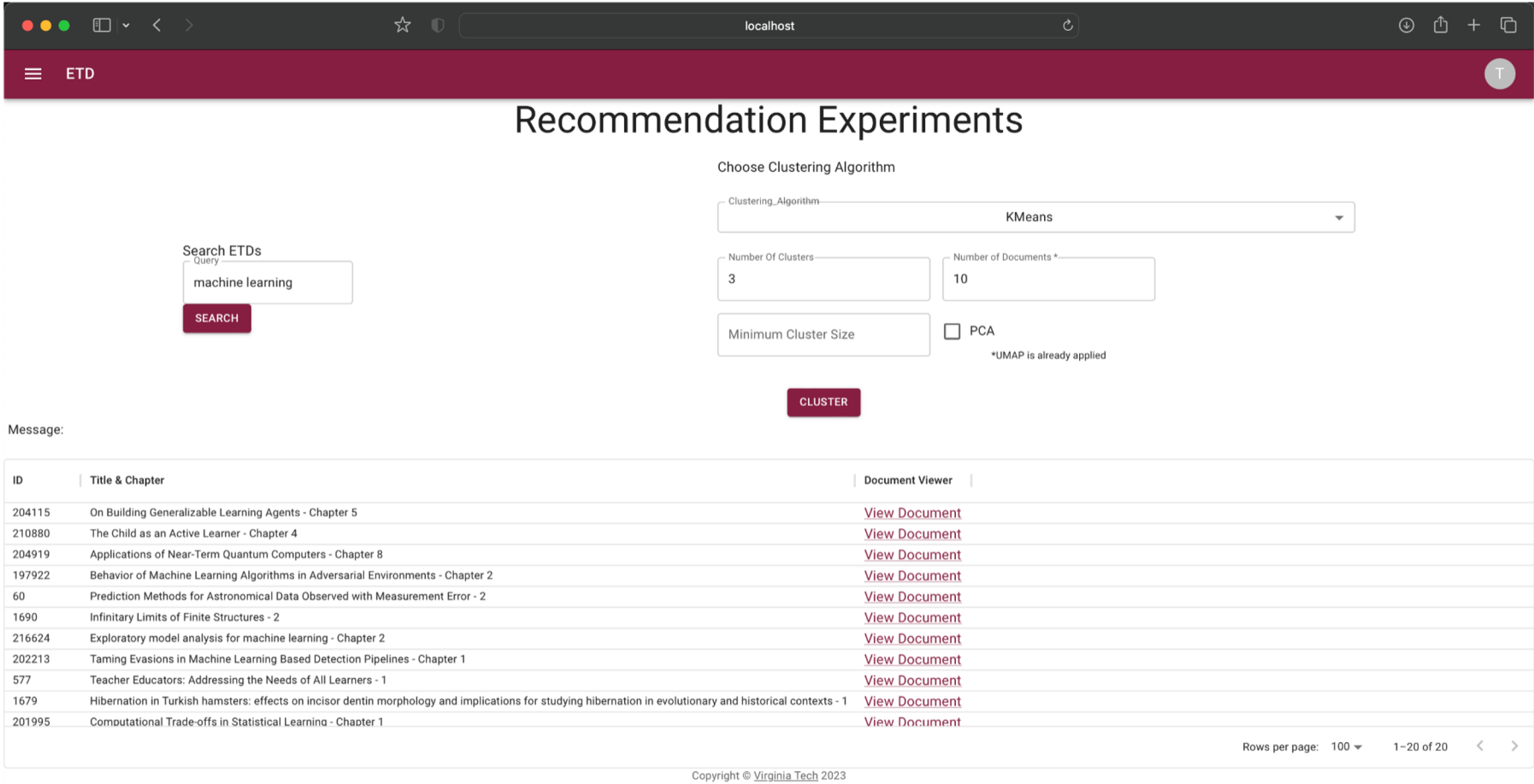
General Approach



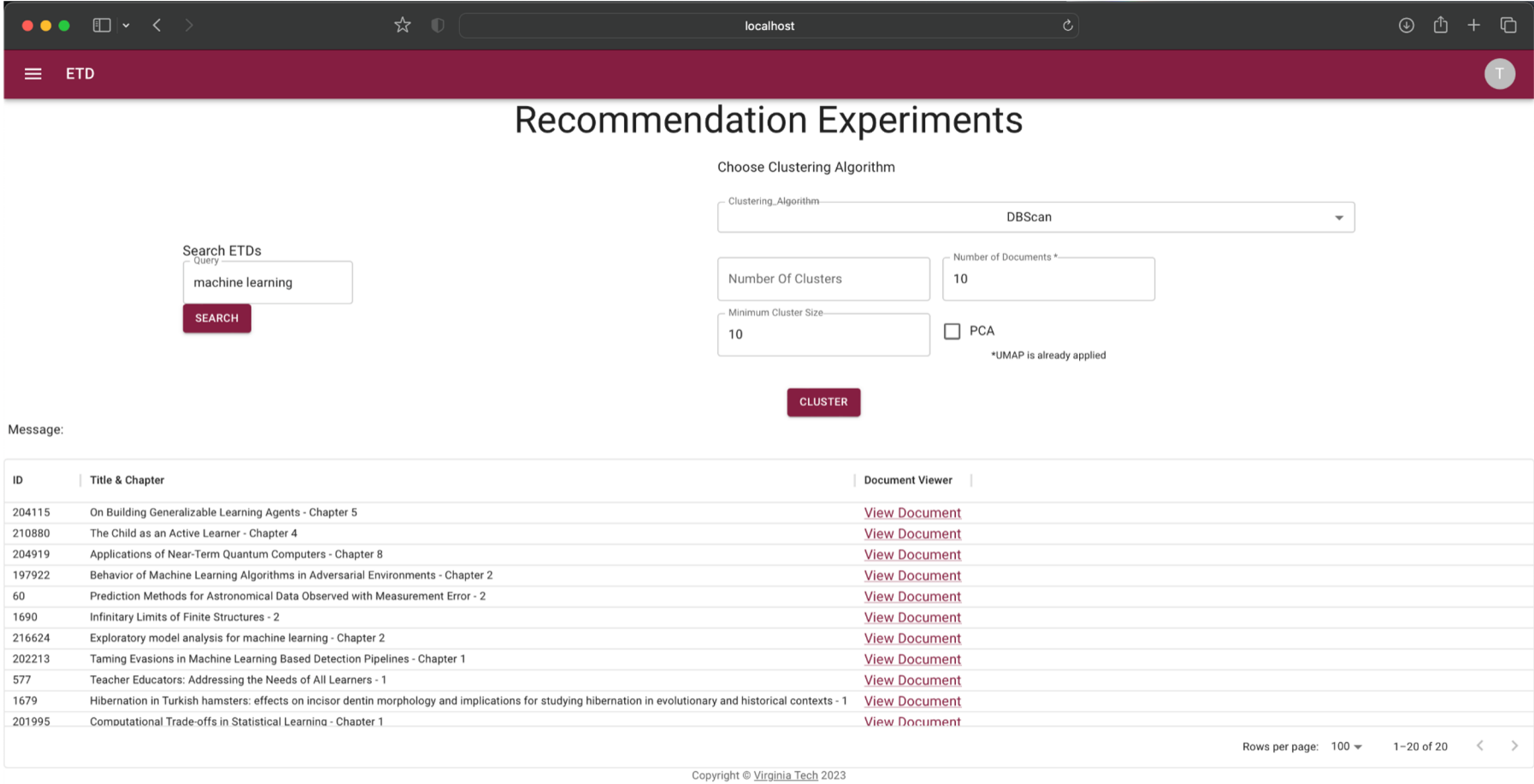
Deliverables



Deliverables

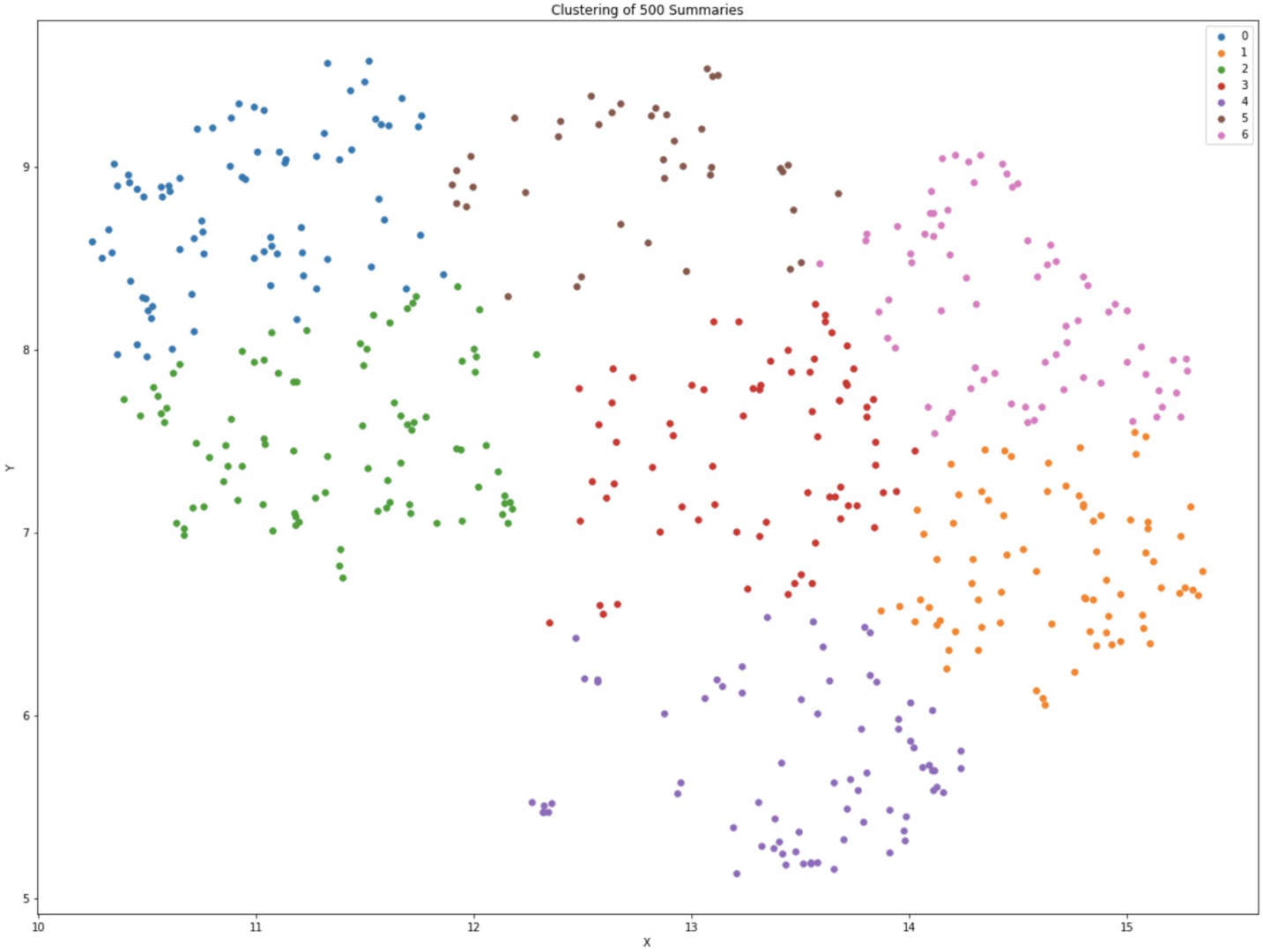


Deliverables - KMeans

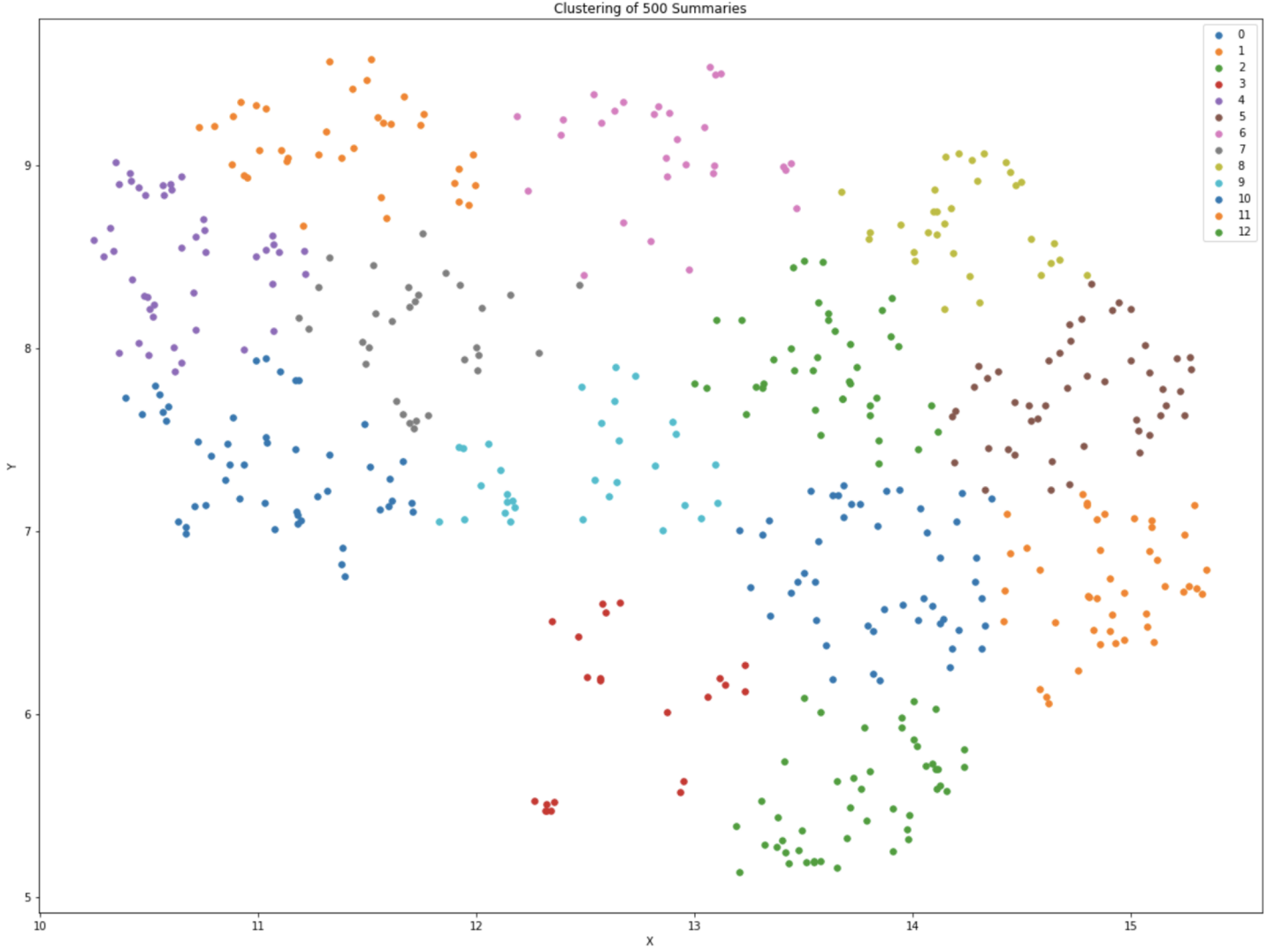


Deliverables - DBScan

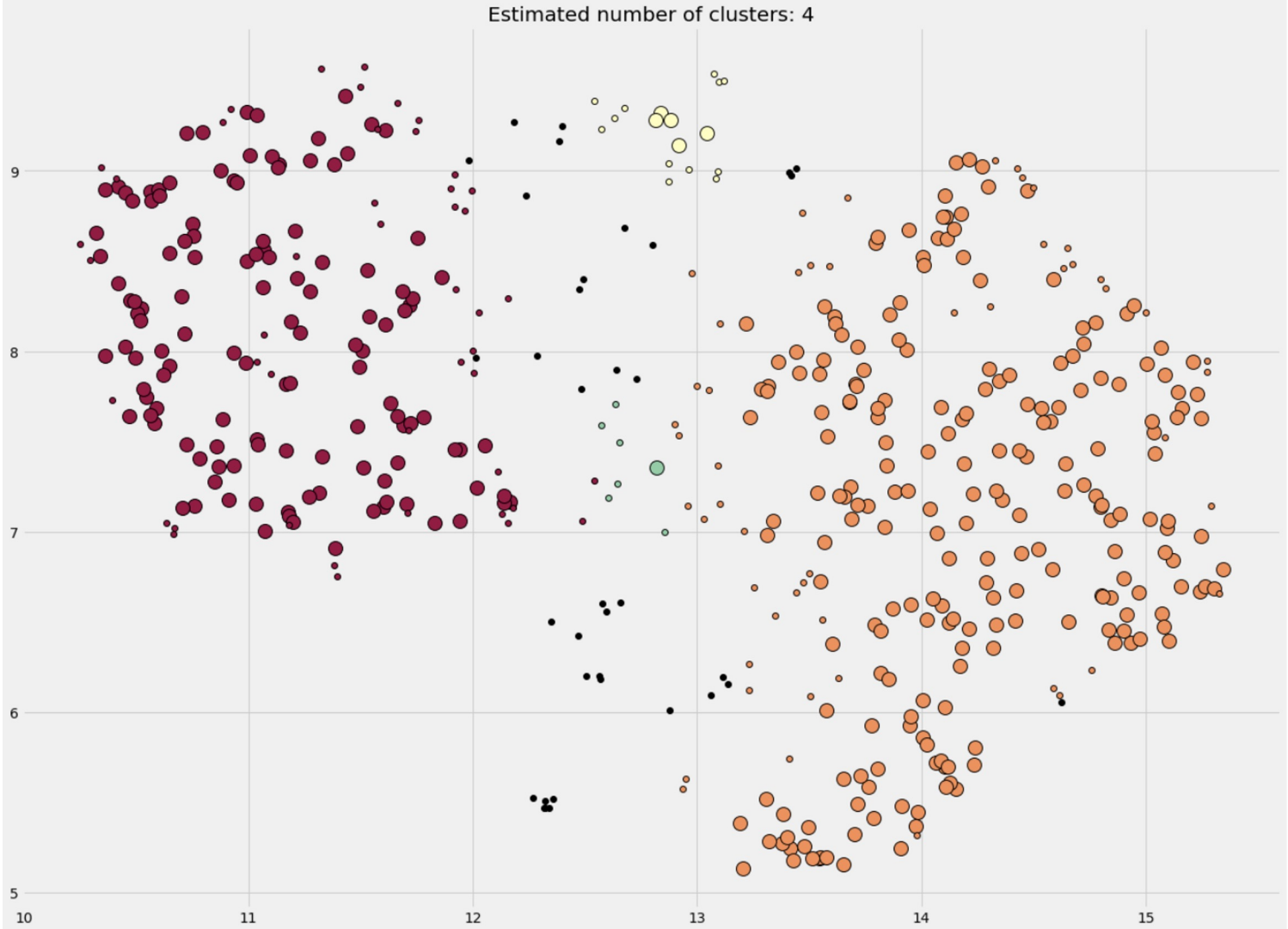
Testing & Evaluation - KMeans



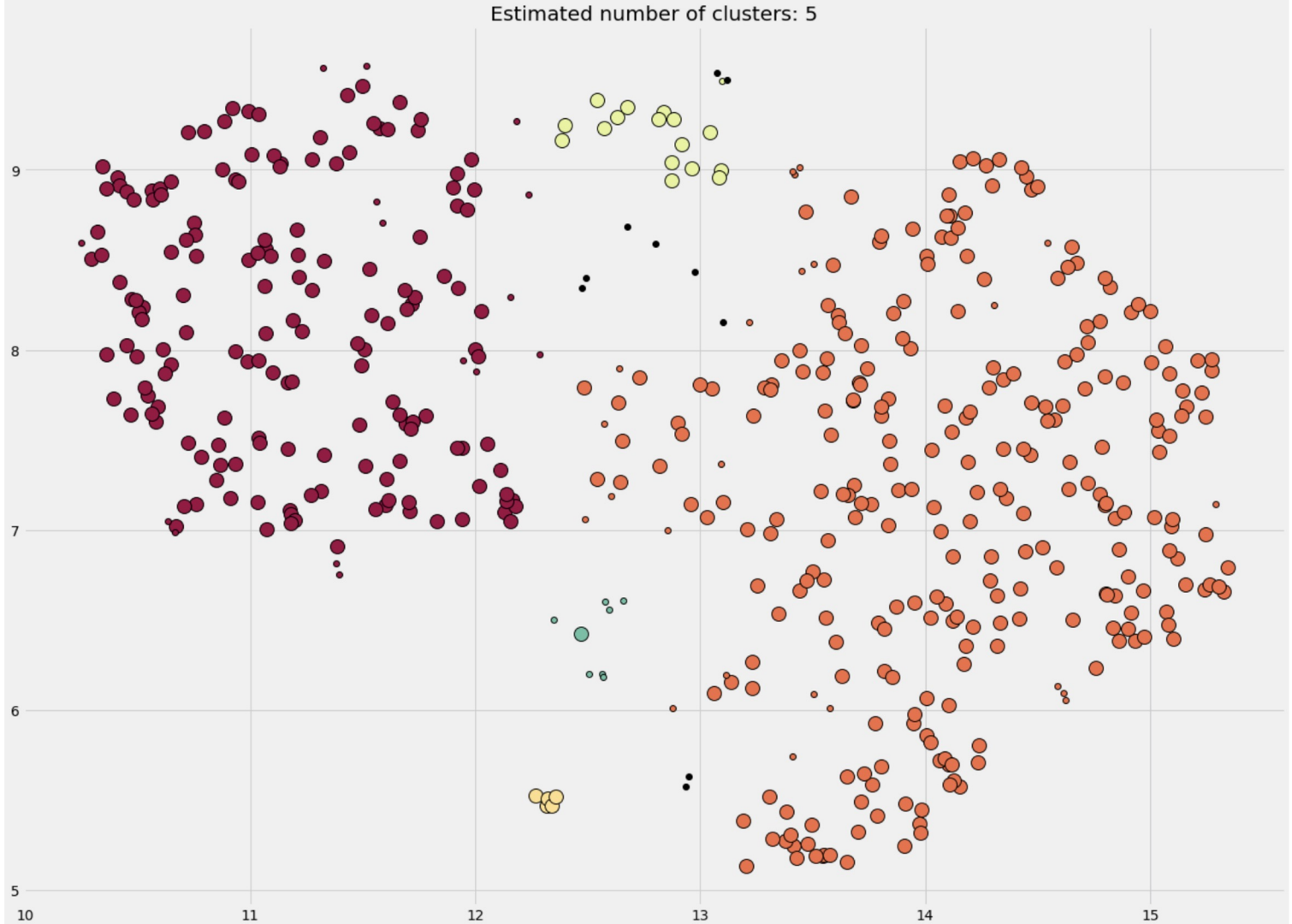
Testing & Evaluation - KMeans



Testing & Evaluation - DBScan



Testing & Evaluation - DBScan



Acknowledgements

- Satvik Chekuri
- Dr. Fox
- Previous CS 5604 teams
  - Chapter Summarization Model
  - Information Retrieval

References

[1] Bhanja, Satyanarayan. “How to Cluster Text Documents Using BERT.” *Theaidigest.in*, 29 Sept. 2020, https://theaidigest.in/how-to-cluster-text-documents-using-bert/. Accessed 3 Mar. 2023.

[2] Caboom.ai. “Speeding up Similarity Search in Recommender Systems Using FAISS.” *Medium*, Artificial Intelligence in Plain English, 13 Aug. 2020, https://ai.plainenglish.io/speeding-up-similarity-search-in-recommender-systems-using-faiss-basics-part-i-ec1b5e92c92d. Accessed 14 Feb. 2023.

[3] Erik Bernhardsson. “Nearest Neighbors and Vector Models – Part 2 – Algorithms and Data Structures.” *Erik Bernhardsson*, Erik Bernhardsson, 19 Apr. 2020, https://erikbern.com/2015/10/01/nearest-neighbors-and-vector-models-part-2-how-to-search-in-high-dimensional-spaces.html. Accessed 14 Feb. 2023.

[4] Frederickson, Ben. “Approximate Nearest Neighbours for Recommender Systems.” *Ben Frederickson*, 11 Oct. 2017, https://www.benfrederickson.com/approximate-nearest-neighbours-for-recommender-systems/. Accessed 14 Feb. 2023.

[5] Ramachandran, Aiswarya. “Building a Movie Recommendation Using FAISS.” *Medium*, Analytics Vidhya, 29 Jan. 2021, https://medium.com/analytics-vidhya/building-a-movie-recommendation-using-faiss-60d65b104dda. Accessed 14 Feb. 2023.

[6] Trabelsi, Eyal. “Comprehensive Guide to Approximate Nearest Neighbors Algorithms.” *Medium*, Towards Data Science, 8 Sept. 2021, https://towardsdatascience.com/comprehensive-guide-to-approximate-nearest-neighbors-algorithms-8b94f057d6b6. Accessed 14 Feb. 2023.

References

[7] Paidiparthy, Manoj Prabhakar. Ramanujan, Ramaraja. Teegalapally, Akshita. Muralikrishnan, Madhuvanti. Balar, Romil Khimraj. Juvekar, Shaunak. Murali, Vivek. 11 Jan. 2023, *CS5604 (Information Retrieval) F2022 End User (EU) Team Project*, http://hdl.handle.net/10919/114080. Accessed 6 Apr. 2023.

[8] Ganesan, Kaushik. Nanjundan, Deepak. Srivastava, Deval. Neog, Abhilash. Jayaprakash, Dharneeshkar. Shah, Aditya. 11 Jan. 2023, *CS5604 (Information Retrieval) Language Models, Classification, Summarization & Segmentation*, http://hdl.handle.net/10919/114077. Accessed 6 Apr. 2023.

[9] Devlin, Jacob, et al. 'BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding'. *CoRR*, vol. abs/1810.04805, 2018, http://arxiv.org/abs/1810.04805. arXiv. Accessed 6 Apr. 2023.

[10] McInnes, Leland, et al. 'UMAP: Uniform Manifold Approximation and Projection'. *Journal of Open Source Software*, vol. 3, no. 29, The Open Journal, 2018, p. 861, https://doi.org10.21105/joss.00861. Accessed 6 Apr. 2023.

[11] Meta, Inc. "Welcome to FAISS Documentation." *Welcome to Faiss Documentation - Faiss Documentation*, Meta, Inc, 2015, https://faiss.ai/. Accessed 6 Apr. 2023.

[12] Spotify USA, Inc.*ANNOY Library*. 2015, https://github.com/spotify/annoy. Accessed 6 Apr. 2023.

[13] Pearson, Karl. 'LIII. On Lines and Planes of Closest Fit to Systems of Points in Space'. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, Taylor & Francis, 1901, pp. 559–572, https://doi.org10.1080/14786440109462720. Accessed 8 Apr. 2023.

References

[14] Elastic, *"*Elasticsearch*"*. 2015, https://github.com/elasticsearch/elasticsearch. Accessed 8 Apr. 2023.

[15] Dorfer, Thomas A. "Density-Based Clustering: DBSCAN vs. HDBSCAN." *Medium*, Towards Data Science, 6 Dec. 2022, https://towardsdatascience.com/density-based-clustering-dbscan-vs-hdbscan-39e02af990c7. Accessed 8 Apr. 2023.

[16] Ginni. "What Is the Difference between K-Means and DBSCAN." *Tutorials Point*, 14 Feb 2022, https://www.tutorialspoint.com/what-is-the-difference-between-k-means-and-dbscan. Accessed 8 Apr. 2023.

[17] McInnes, Leland. Healy, John. Astels, Steve. "Comparing Python Clustering Algorithms." *Comparing Python Clustering Algorithms - Hdbscan 0.8.1 Documentation*, 2016, https://hdbscan.readthedocs.io/en/latest/comparing_clustering_algorithms.html. Accessed 8 Apr. 2023.

[18] Trivedi, Aashka. Trivedi, Anya. Varshney, Sourabh. Joshipura, Vidhey. Mehta, Rupa. Dhanani, Jenish. IEEE, 2020, *Extracted Summary Based Recommendation System for Indian Legal Documents*, ieeexplore.ieee.org/abstract/document/9225290. Accessed 10 Apr. 2023.

[19] Radev, Dragomir R. Fan, Weiguo. Zhang, Zhu. *WebInEssence: A Personalized Web-Based Multi-Document Summarization and Recommendation System*, Sept. 2001, citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=40fccfffe550e84797d146f856e5d437f52ed422. Accessed 10 Apr. 2023.

References

[20] Trabelsi, Eyal. “Comprehensive Guide to Approximate Nearest Neighbors Algorithms.” *Medium*, Towards Data Science, 8 Sept. 2021, https://towardsdatascience.com/comprehensive-guide-to-approximate-nearest-neighbors-algorithms-8b94f057d6b6. Accessed 10 Apr. 2023.

[21] Real Python. “K-Means Clustering in Python: A Practical Guide.” *Real Python*, Real Python, 30 Jan. 2023, https://realpython.com/k-means-clustering-python/#choosing-the-appropriate-number-of-clusters. Accessed 12 Apr. 2023.

[22] Arvai, Kevin. *Kneed Package,* kneed. 0.7.0, 8 2020, https://doi.org10.5281/zenodo.6496267. Accessed 12 Apr. 2023.

[23] Frenzel, Charles. “Tuning with HDBSCAN.” *Medium*, Towards Data Science, 11 Sept. 2021, towardsdatascience.com/tuning-with-hdbscan-149865ac2970.