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

February 2023:
- Research recommendation systems and Python libraries
- Collect chapter summaries & create embeddings

March 2023:
- Run clustering algorithms & perform data analysis
- Prepare for integration with front-end

April 2023:
- Integrate with front-end
- 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
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
General Approach

- Preprocessing
- BERT
- UMAP
- Search query
- Clustering Algorithm
- Sorting Algorithm
- Front-end

KMeans, DBScan or HDBScan using user specifications
Sort output of each cluster to retrieve two nearest results to first $n$ chapters
Append two nearest documents to search results.
Recommendation Experiments

Choose Clustering Algorithm

- Clustering Algorithm
- Number Of Clusters
- Number of Documents
- Minimum Cluster Size
- PCA

*UMAP is already applied

Message:

ID | Title & Chapter | Document Viewer |
---|-----------------|----------------|

No rows
## Recommendation Experiments

### Choose Clustering Algorithm

<table>
<thead>
<tr>
<th>Clustering Algorithm</th>
<th>Number Of Clusters</th>
<th>Number of Documents</th>
<th>Minimum Cluster Size</th>
<th>PCA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>UMAP is already applied</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Search ETDs

- Search Term: machine learning

### Message:

### Deliverables

<table>
<thead>
<tr>
<th>ID</th>
<th>Title &amp; Chapter</th>
<th>Document Viewer</th>
</tr>
</thead>
<tbody>
<tr>
<td>204115</td>
<td>On Building Generalizable Learning Agents - Chapter 5</td>
<td>View Document</td>
</tr>
<tr>
<td>210880</td>
<td>The Child as an Active Learner - Chapter 4</td>
<td>View Document</td>
</tr>
<tr>
<td>204919</td>
<td>Applications of Near-Term Quantum Computers - Chapter 8</td>
<td>View Document</td>
</tr>
<tr>
<td>197922</td>
<td>Behavior of Machine Learning Algorithms in Adversarial Environments - Chapter 2</td>
<td>View Document</td>
</tr>
<tr>
<td>216024</td>
<td>Exploratory model analysis for machine learning - Chapter 2</td>
<td>View Document</td>
</tr>
<tr>
<td>202213</td>
<td>Taming Evasions in Machine Learning Based Detection Pipelines - Chapter 1</td>
<td>View Document</td>
</tr>
<tr>
<td>201995</td>
<td>Computational Trade-offs in Statistical Learning - Chapter 1</td>
<td>View Document</td>
</tr>
<tr>
<td>198346</td>
<td>System Design for Large Scale Machine Learning - Chapter 2</td>
<td>View Document</td>
</tr>
<tr>
<td>198345</td>
<td>System Design for Large Scale Machine Learning - Chapter 1</td>
<td>View Document</td>
</tr>
<tr>
<td>204774</td>
<td>Statistical Learning Towards Gamification in Human-Centric Cyber-Physical Systems - Chapter 6</td>
<td>View Document</td>
</tr>
<tr>
<td>211448</td>
<td>Distributed Machine Learning with Communication Constraints - Chapter 1</td>
<td>View Document</td>
</tr>
</tbody>
</table>

Copyright © Virginia Tech 2023
Recommendation Experiments

Choose Clustering Algorithm

- Clustering Algorithm: KMeans
- Number of Clusters: 3
- Number of Documents: 10
- Minimum Cluster Size

*UMAP is already applied

Message:

<table>
<thead>
<tr>
<th>ID</th>
<th>Title &amp; Chapter</th>
<th>Document Viewer</th>
</tr>
</thead>
<tbody>
<tr>
<td>204115</td>
<td>On Building Generalizable Learning Agents - Chapter 5</td>
<td>View Document</td>
</tr>
<tr>
<td>210880</td>
<td>The Child as an Active Learner - Chapter 4</td>
<td>View Document</td>
</tr>
<tr>
<td>204919</td>
<td>Applications of Near-Term Quantum Computers - Chapter 8</td>
<td>View Document</td>
</tr>
<tr>
<td>197922</td>
<td>Behavior of Machine Learning Algorithms in Adversarial Environments - Chapter 2</td>
<td>View Document</td>
</tr>
<tr>
<td>60</td>
<td>Prediction Methods for Astronomical Data Observed with Measurement Error - 2</td>
<td>View Document</td>
</tr>
<tr>
<td>1690</td>
<td>Infritary Limits of Finite Structures - 2</td>
<td>View Document</td>
</tr>
<tr>
<td>216624</td>
<td>Exploratory model analysis for machine learning - Chapter 2</td>
<td>View Document</td>
</tr>
<tr>
<td>202213</td>
<td>Taming Evasions in Machine Learning Based Detection Pipelines - Chapter 1</td>
<td>View Document</td>
</tr>
<tr>
<td>577</td>
<td>Teacher Educators: Addressing the Needs of All Learners - 1</td>
<td>View Document</td>
</tr>
<tr>
<td>1679</td>
<td>Hibernation in Turkish hamsters: effects on incisor dentin morphology and implications for studying hibernation in evolutionary and historical contexts - 1</td>
<td>View Document</td>
</tr>
<tr>
<td>201995</td>
<td>Computational Trade-offs in Statistical Learning - Chapter 1</td>
<td>View Document</td>
</tr>
</tbody>
</table>

Copyright © Virginia Tech 2023

Deliverables - KMeans
### Recommendation Experiments

Choose Clustering Algorithm

- **Clustering Algorithm**: DBScan

- **Number Of Clusters**: 10

- **Minimum Cluster Size**: 10

- **PCA**: False

*UMAP is already applied

---

**Message:**

- **ID**: 204115
- **Title & Chapter**: On Building Generalizable Learning Agents - Chapter 5

- **ID**: 210880
- **Title & Chapter**: The Child as an Active Learner - Chapter 4

- **ID**: 204919
- **Title & Chapter**: Applications of Near-Term Quantum Computers - Chapter 8

- **ID**: 197922
- **Title & Chapter**: Behavior of Machine Learning Algorithms in Adversarial Environments - Chapter 2

- **ID**: 60
- **Title & Chapter**: Prediction Methods for Astronomical Data Observed with Measurement Error - 2

- **ID**: 1690
- **Title & Chapter**: Infinitary Limits of Finite Structures - 2

- **ID**: 216624
- **Title & Chapter**: Exploratory model analysis for machine learning - Chapter 2

- **ID**: 202313
- **Title & Chapter**: Taming Evasions in Machine Learning Based Detection Pipelines - Chapter 1

- **ID**: 577
- **Title & Chapter**: Teacher Educators: Addressing the Needs of All Learners - 1

- **ID**: 1679
- **Title & Chapter**: Hibernation in Turkish Hamsters: effects on incisor dentin morphology and implications for studying hibernation in evolutionary and historical contexts - 1

- **ID**: 201995
- **Title & Chapter**: Computational Trade-offs in Statistical Learning - Chapter 1

---

**Deliverables - DBScan**

---

**Copyright © Virginia Tech 2023**
Clustering of 500 Summaries

Testing & Evaluation - KMeans
Estimated number of clusters: 4
Acknowledgements

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


References


References


[15] Dorfer, Thomas A. “Density-Based Clustering: DBSCAN vs. HDBSCAN.” Medium, Towards Data Science, 6 Dec. 2022,


References


