Final Project Presentation
Team 3- Object Detection and Topic Modeling

Pradyumna Upendra Dasu, Amr Ahmed Aboelnaga , Anushka Sivakumar, Jayanth Narla,
Ragul Seetharaman, Sahana Bhaskar, Shankar Srinidhi Srinivas

Under the guidance of **Dr. Edward Fox**

SME: Chenyu Mao

30 November 2023
CS5604, Virginia Tech, Blacksburg, VA 24061

Sections Covered

- Team
- Deliverables and responsibility
- Project management
- Object detection
  - Tasks completed
  - Tasks in progress
  - Future work
- Chapter segmentation
  - Methodology
  - Tasks completed
  - Tasks in progress
  - Future work
- Topic modeling
  - Tasks completed
  - Future work

TEAM

Subject Matter Expert - Chenyu Mao (mchenyu@vt.edu)

Email (for team)**:** team3_CS5604-g@vt.edu Slack

Name	Email	Responsibility
Anushka Sivakumar	anushkas01@vt.edu	Object detection: XML generation with IDs, API integration, multiprocessing
Amr Ahmed Aboelnaga	amraboelnaga@vt.edu	Image generation, object detection, image classification, chapter segmentation using table of contents.
Jayanth Narla	jnarla@vt.edu	Topic modeling API integration
Pradyumna Upendra Dasu	pradyumnaupendra@vt.edu	Topic modeling experimentation and data cleaning
Ragul Seetharaman	ragul@vt.edu	Object detection and documentation
Sahana Bhaskar	sahanab@vt.edu	XML generation, API integration, experimentation with multiprocessing
Shankar Srinidhi Srinivas	shankarsrinidhi@vt.edu	Topic modeling and documentation

DELIVERABLE

A system supporting searching etc. using topics, where users can search on derived digital objects, and where experimenters can further research about objects and topics.

RESPONSIBILITIES

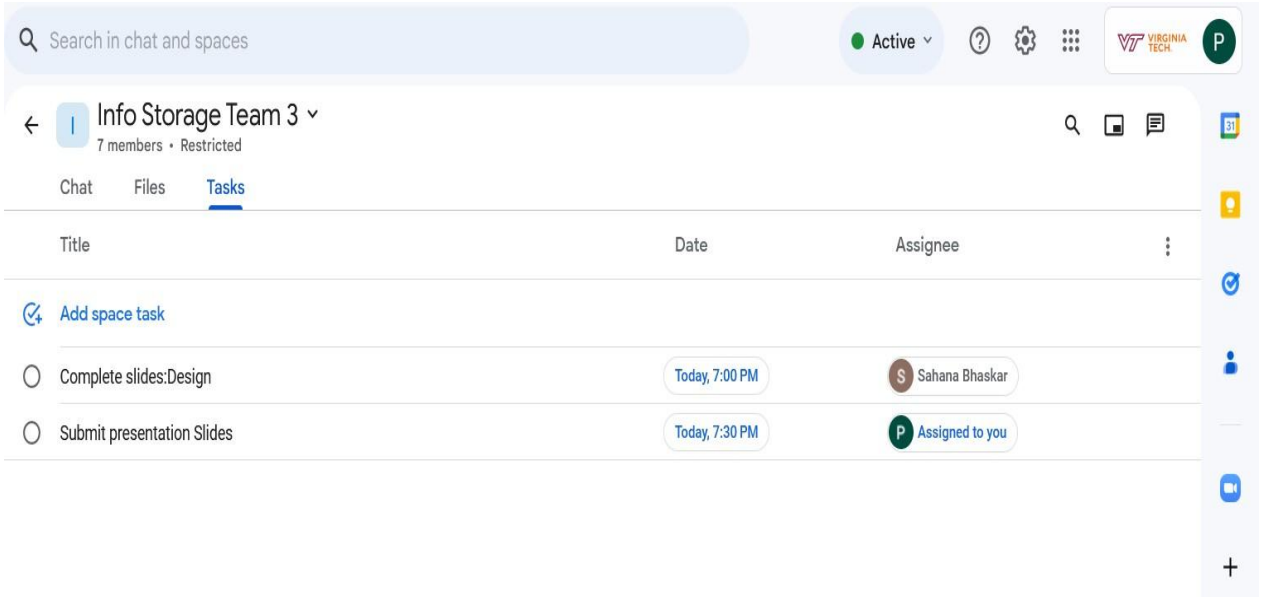
1. *Responsibility 1:*
    - Analyzing documents, considering object detection and topic models.
    - Improving the current methods and deploying them widely.
2. *Responsibility 2*:

Building on the existing system working with object detection and topic analysis.

- Help **Team 1** so object detection results will be represented in the KG.
- Help **Team 2** so detected objects and topics can be integrated into indexing.
- Help **Team 4** so classification and summarization can be over chapters.
- Help **Team 6** for integration with the UI.

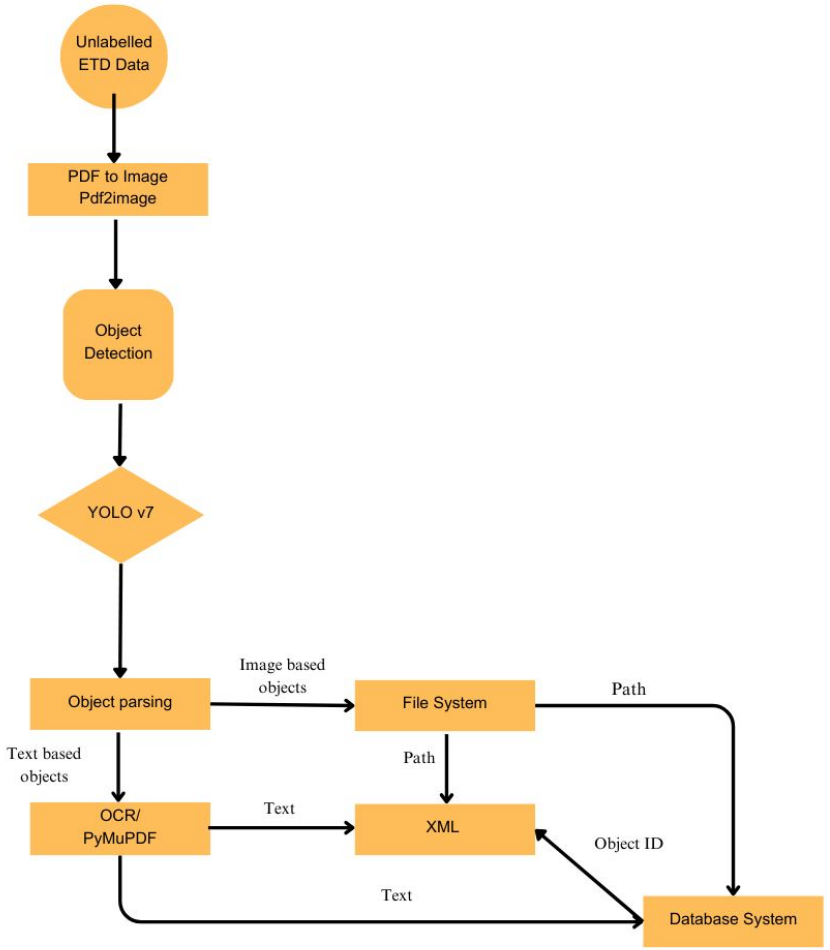
PROJECT MANAGEMENT

1. *Weekly meeting on Wednesdays via Zoom.*
2. *Using Google Tasks for collaboration and task management.*
3. *Slack for sharing resources and links.*



Object detection

DESIGN


Unlabelled ETD Data
PDF to Image Pdf2image
Object Detection
YOLO v7
Object parsing
Image based objects
File System
Path
Text based objects
Path
OCR/ PyMuPDF
Text
XML
Object ID
Text
Database System

TASKS COMPLETED

- Object detection setup in Endeavour.
- Modified the object detection pipeline to save the page images as well as objects detected.
- Object detection run on 200 ETDs that has already been segmented in past work.
- Updated the XML generation code to include unique object IDs that Team 1 needs for recording structural connections and reran the object detection code on the 200 ETDs.
- Defined and integrated the required APIs to populate the database and save the files corresponding to the ETD.

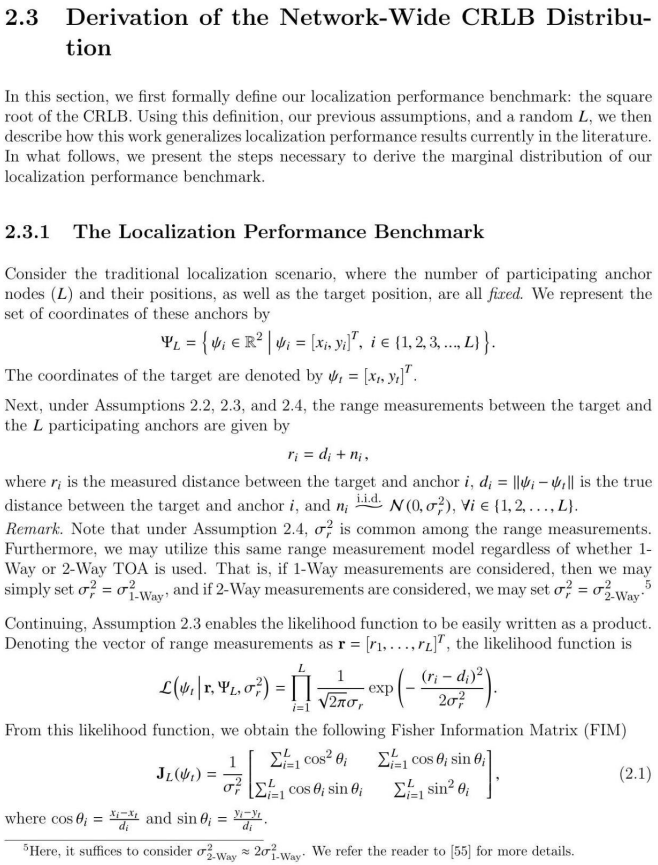
TASKS COMPLETED

Example results of an ETD through the object detection pipeline

TASKS COMPLETED

➢ Objects Detected from Page Image

Image of the page



2.3 Derivation of the Network-Wide CRLB Distribution

In this section, we first formally define our localization performance benchmark: the square root of the CRLB. Using this definition, our previous assumptions, and a random L, we then describe how this work generalizes localization performance results currently in the literature. In what follows, we present the steps necessary to derive the marginal distribution of our localization performance benchmark.

2.3.1 The Localization Performance Benchmark

Consider the traditional localization scenario, where the number of participating anchor nodes (L) and their positions, as well as the target position, are all *fixed*. We represent the set of coordinates of these anchors by

$$\Psi_L = \left\{ \psi_i \in \mathbb{R}^2 \,\middle|\, \psi_i = [x_i, y_i]^T,\ i \in \{1, 2, 3, ..., L\} \right\}.$$

The coordinates of the target are denoted by $\psi_t = [x_t, y_t]^T$.

Next, under Assumptions 2.2, 2.3, and 2.4, the range measurements between the target and the L participating anchors are given by

$$r_i = d_i + n_i,$$

where r_i is the measured distance between the target and anchor i, $d_i = \|\psi_i - \psi_t\|$ is the true distance between the target and anchor i, and $n_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_r^2)$, $\forall i \in \{1, 2, \ldots, L\}$.

Remark. Note that under Assumption 2.4, σ_r^2 is common among the range measurements. Furthermore, we may utilize this same range measurement model regardless of whether 1-Way or 2-Way TOA is used. That is, if 1-Way measurements are considered, then we may simply set $\sigma_r^2 = \sigma_{\text{1-Way}}^2$, and if 2-Way measurements are considered, we may set $\sigma_r^2 = \sigma_{\text{2-Way}}^2$.[5]

Continuing, Assumption 2.3 enables the likelihood function to be easily written as a product. Denoting the vector of range measurements as $\mathbf{r} = [r_1, \ldots, r_L]^T$, the likelihood function is

$$\mathcal{L}\left(\psi_t \,\middle|\, \mathbf{r}, \Psi_L, \sigma_r^2\right) = \prod_{i=1}^{L} \frac{1}{\sqrt{2\pi}\sigma_r} \exp\left(-\frac{(r_i - d_i)^2}{2\sigma_r^2}\right).$$

From this likelihood function, we obtain the following Fisher Information Matrix (FIM)

$$\mathbf{J}_L(\psi_t) = \frac{1}{\sigma_r^2} \begin{bmatrix} \sum_{i=1}^{L} \cos^2\theta_i & \sum_{i=1}^{L} \cos\theta_i \sin\theta_i \\ \sum_{i=1}^{L} \cos\theta_i \sin\theta_i & \sum_{i=1}^{L} \sin^2\theta_i \end{bmatrix}, \tag{2.1}$$

where $\cos\theta_i = \frac{x_i - x_t}{d_i}$ and $\sin\theta_i = \frac{y_i - y_t}{d_i}$.

[5] Here, it suffices to consider $\sigma_{\text{2-Way}}^2 \approx 2\sigma_{\text{1-Way}}^2$. We refer the reader to [55] for more details.

Text detected from the page

```
494935183
2.3 Derivation of the Network-Wide CRLB Distribu-
tion

494935184
In this section, we first formally define our localization performance benchmark: the square
root of the CRLB. Using this definition, our previous assumptions, and a random we then
L,
describe how this work generalizes localization performance results currently in the literature.
In what follows, we present the steps necessary to derive the marginal distribution of our
localization performance benchmark.

494935185
2.3.1 The Localization Performance Benchmark

494935186
Consider the traditional localization scenario, where the number of participating anchor
nodes and their positions, as well as the target position, are all fixed. We represent the
(L)
set of coordinates of these anchors by
□ □
□□□

494935188
∈ ∈ {1, L}
□ □
=
The coordinates of the target are denoted by
ψt [xt, yt]T.
□□□
Next, under Assumptions 2.2, 2.3, and 2.4, the range measurements between the target and
the participating anchors are given by
L
```

TASKS COMPLETED

Images detected

$$\Psi_L = \left\{ \psi_i \in \mathbb{R}^2 \,\middle|\, \psi_i = [x_i, y_i]^T,\ i \in \{1, 2, 3, \ldots, L\} \right\}.$$

$$r_i = d_i + n_i\,,$$

$$\mathcal{L}\left(\psi_t \,\middle|\, \mathbf{r}, \Psi_L, \sigma_r^2\right) = \prod_{i=1}^{L} \frac{1}{\sqrt{2\pi}\sigma_r} \exp\left(-\frac{(r_i - d_i)^2}{2\sigma_r^2}\right).$$

$$\mathbf{J}_L(\psi_t) = \frac{1}{\sigma_r^2} \begin{bmatrix} \sum_{i=1}^{L} \cos^2\theta_i & \sum_{i=1}^{L} \cos\theta_i \sin\theta_i \\ \sum_{i=1}^{L} \cos\theta_i \sin\theta_i & \sum_{i=1}^{L} \sin^2\theta_i \end{bmatrix},$$

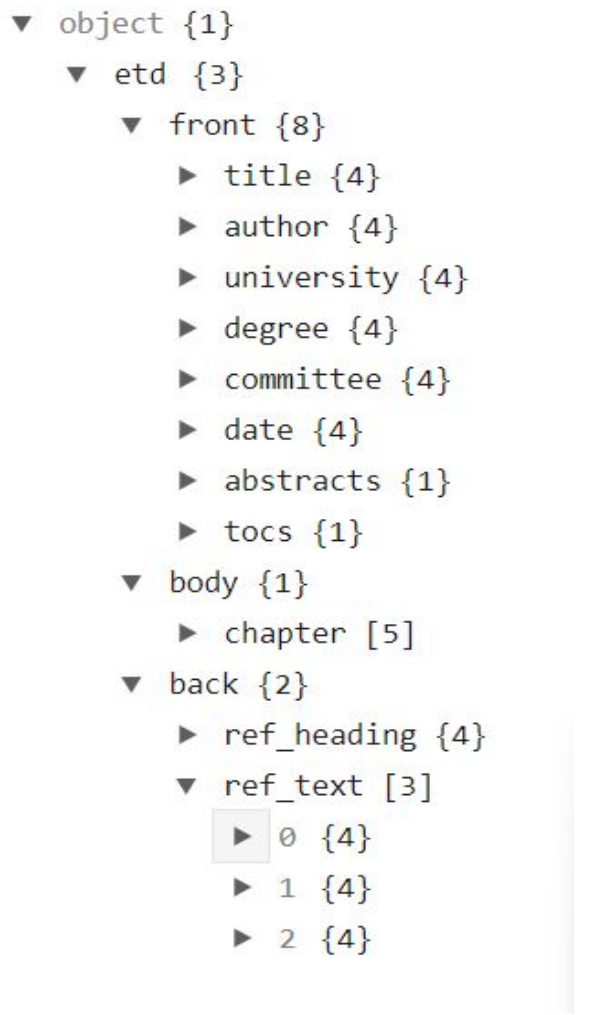
TASKS COMPLETED

- XML Generated

The XML schema contains the ETD ID as the root element and 3 subsections:-

1) Front
2) Body
3) Back

XML Schema:



TASKS COMPLETED

XML Generated

Front Section

```
- <etd>
  - <front>
      <title pg_no="0" conf="0.88" bbox="[294.33, 375.08, 1498.85, 497.25]">UNA REVOLUCION, NI MAS NI MENOS: THE ROLE OF THE ENLIGHTENMENT IN THE SUPREME JUNTAS IN QUITO, 1765-1822</title>
      <author pg_no="0" conf="0.84" bbox="[687.67, 922.45, 1102.1, 986.7]">Beau James Brammer, B.A.</author>
      <university pg_no="0" conf="0.84" bbox="[700.9, 1151.81, 1098.08, 1214.31]">The Ohio State University</university>
      <degree pg_no="0" conf="0.76" bbox="[291.6, 652.02, 1507.71, 762.29]">Presented in Partial Fulfillment of the Requirements for The Degree Master of Arts in the Graduate School of The Ohio State University</degree>
      <committee pg_no="0" conf="0.87" bbox="[645.96, 1376.1, 1148.21, 1686.49]">Master's Examination Committee: Kenneth Andrien, Adviser Stephanie Smith Alan Gallay</committee>
      <date pg_no="0" conf="0.79" bbox="[847.73, 1236.25, 949.35, 1285.87]">2010</date>
    - <abstracts>
      - <abstract>
          <abs_heading pg_no="2" conf="0.75" bbox="[801.07, 335.21, 973.89, 405.18]">Abstract</abs_heading>
          <abs_text>This thesis examines the role the European Enlightenment played in the political sphere during the late colonial era in the Audiencia of Quito. Until the eighteenth century, Creole elites controlled the local economic and governmental sectors. With the ascension of the Bourbon dynasty in 1700, however, these elites of Iberian descent began to lose their power as new European ideas, emerging from the Enlightenment, led to a process of consolidating and centralizing power into the hands of Peninsular Spanish officials. Known as the Bourbon Reforms, these measures led to Creole disillusionment, as they began losing power at the local level. Beginning in the 1770s and 1780s, however, Enlightenment ideas of "nationalism" and "rationality" arrived in the Andean capital, making their way to the disgruntled Creoles. As the situation deteriorated, elites began to incorporate these new concepts into their rhetoric, presenting a possible response to the Reforms. When Napoleon invaded Spain in 1808, the Creoles expelled the Spanish government in Quito, creating an autonomous movement, the Junta of 1809, using these Enlightenment principles as their justification. I argue, however, that while these 'modern' principles gave the Creoles an outlet for their grievances, it is their inability to find a common ground on how their government should interpret these new ideas which ultimately lead to the Junta's failure.This conclusion challenges previous historiography which contends that the political and economic turmoil in Quito were the only prominent factors leading to the Junta Era of 1809 to 1812 and when discussed, scholars view the Enlightenment as a catalyst for beneficial change in the region. This thesis contends that the Enlightenment principles adopted by local elites, while giving them the opportunity to revolt, also divided the Creole elite, ultimately ending the possibility of any successful autonomous movement. In the end, I contend that it is necessary for scholars to look at both the positive and negative ramifications of Enlightenment principles when studying the Latin American movements for independence.This conclusion challenges previous historiography which contends that the political and economic turmoil in Quito were the only prominent factors leading to the Junta Era of 1809 to 1812 and when discussed, scholars view the Enlightenment as a catalyst for beneficial change in the region. This thesis contends that the Enlightenment principles adopted by local elites, while giving them the opportunity to revolt, also divided the Creole elite, ultimately ending the possibility of any successful autonomous movement. In the end, I contend that it is necessary for scholars to look at both the positive and negative ramifications of Enlightenment principles when studying the Latin American movements for independence.</abs_text>
      </abstract>
    </abstracts>
    - <tocs>
      - <tocs>
          <toc_heading pg_no="7" conf="0.83" bbox="[750.83, 332.7, 1043.14, 409.03]">Table of Contents</toc_heading>
          <toc_text>Page Abstract .....ii Dedication.....iv Acknowledgements .....v Vita .....vi List of Graphs.....viii List of Abbreviations.....ix Chapters: 1 Introduction..... .....1 2 Historiography.....6 3 Charles III and the Quito Insurrection of 1765.....21 4 Contrasting Presidencies: Jose García de Leon y Pizarro and the Baron de Carondelet..... .....28 5 Napoleonic Invasion and the Junta of 1809.....38 6 Massacre of August 2nd and the Junta of 1810.....56 7 An Era of Uneasy Peace: Quito from 1812 to 1822.....69 8 Conclusion.....79 References.....</toc_text>
```

TASKS COMPLETED

Body Section


```
- <body>
    + <chapter>
    + <chapter>
    + <chapter>
    + <chapter>
    + <chapter>
    + <chapter>
    + <chapter>
    - <chapter>
        <title pg_no="78" conf="0.58" bbox="[527.32, 399.93, 1366.62, 492.89]">Chapter 7: An Era of Uneasy Peace: Quito 1812 to 1822</title>
        <sections/>
      - <section>
            <name pg_no="0" conf="0" bbox="[0, 0, 0, 0]"/>
          - <paragraphs>
                <para pg_no="78" conf="0.94" bbox="[261.18, 558.39, 1532.84, 1726.42]">The end of the Junta of 1810 marked the end of the Junta Era in Quito, leading to a ten-year period
                marked by uneasy peace and questions over the direction of royalist rule in the Audiencia. The royalist government of Toribio Montes centered its attention on how to handle
                the insurgents left in the city and their response to the reimposition of royal rule to forestall another autonomous movement. Empire-wide conflicts and events, such as the
                Hidalgo Revolt in Mexico from 1810 until 1813 and the return of Ferdinand VII to assume the throne as King of Spain in 1814, directly affected royal policy in the Audiencia
                and governmental reaction to the Juntas.132 Local royalists focused on the political situation in Spain as they waited to see how Ferdinand VII"s return affected the Andean
                capital. An uneasy peace settled on the city, but the royalists remained alert to any suspicious actions possibly connected to an autonomous movement. In response to the
                heightened tensions, a new president, Don Toribio Montes, instituted a new conciliatory policy focused on finding a balance between punishing the remaining insurgents and
                incorporating them into the new royalist society. By doing so, Montes was able to reassert control in this turbulent, yet peaceful, era in Quito until the Battle of</para>
                <para pg_no="79" conf="0.97" bbox="[258.08, 180.89, 1535.83, 1776.69]">Pichincha in 1822, when insurgent troops from Venezuela and Colombia forced their way into the city,
                officially separating the city from Spanish rule.133 When examining the effects Juntas of 1809 and 1810 in post-Junta Era Quito, two key characteristics define these
                autonomous movements: their regionalist nature and the ideological divisions between the supporters of the Junta and the Crown. A letter written by a Creole royalist,
                Ramón Nuñez del Arco in 1813, presents detailed information about those involved in the Juntas. He sent to Spain a list of those who supported the Juntas from 1809-1812.
                81% (361) of supporters were Creoles in Quito. Del Arco also shows that 61% (154) of royalists were Creoles.134 This division among Creoles added to the insurgents"
                inability to spread the autonomous movement and provides evidence of their lack of support outside the North-Central Sierra. In this intriguing situation, as Montes was
                forced to appease the royalist government in Cádiz and at the same time relieve uneasy tensions surrounding the large amount of locals who supported the movement. The
                lack of insurgent support outside the city enabled the President to focus on what went on Quito instead of the peripheral regions within the Audiencia. Del Arco"s letter
                confirms that the Juntas of 1809 and 1810 were regionalist movements, as support did not spread outside the North-Central Sierra. As Graph 1 shows, unsurprisingly, only
                2% of the local population supporting the Juntas came from outside the Andean region (peninsulars), and all of them resided in Quito during the Junta Era. Of the remaining
                population, less than 1% (7) came from regions outside the North- Central Sierra, further showing the inability of the local elites to incorporate Creoles</para>
                <para pg_no="80" conf="0.93" bbox="[255.17, 181.66, 1525.82, 1716.18]">outside the Andean region. Old textile producing strongholds, however, such as Ambato and
                Riobamba, sided with Quito during the Junta era. This helps demonstrate why local Creoles were unable to gain support for the Juntas in Cuenca, Guayaquil, and other
                regions in the Audiencia. Another key concept that Del Arco"s letter shows is that not all Creoles in Quito supported the autonomous movements, instead a large portion of
                the elite population allied with the Spanish Crown. Graph 2 shows that of the 154 people identified as being Royalist supporters, 94, or 61%, were Creoles from Quito. Only
                27% of Royalist support came from Spaniards (peninsulars), with the other 11% coming from surrounding regions. Such a trend shows the regionalist nature of the
                movement, as well as the importance of the Creole split in support. A once unified elite population fragmented, making it difficult to forge a consensus over the future of the
                Andean capital. Such a division, also, was not common among other autonomous movements across Latin America during the independence era, creating a unique situation
                for the Andean capital. Such a disparity in support from within the city only added to the inability of local insurgents to create a stable governmental structure. The
                enlightened elite population within Quito, the ones responsible for both Supreme Juntas, followed a different path than those in other regions across the Audiencia. From a
                political standpoint, the Quiteños were not the only ones losing local power and control, therefore presenting the question, what made the situation in Quito</para>
                <para pg_no="83" conf="0.96" bbox="[258.23, 181.3, 1530.06, 1855.92]">different from other regions in the Audiencia ultimately leading to revolt? The answer lies in its
                ideological make up based on its political affiliation within the Spanish Empire. Quito, being the capital of the Audiencia and primary supplier of textiles to the silver mines to
```

TASKS COMPLETED

Body Section

```
        <footnotes/>
        <algorithms/>
        <equations/>
      - <figures>
          - <figure>
                <path pg_no="81" conf="0.92" bbox="[269.43, 176.61, 1496.19, 1111.11]">static/107411/detected_images/107411_81_8QOA.jpg</path>
                <caption pg_no="81" conf="0.87" bbox="[268.65, 1139.65, 1515.71, 1467.1]">Criollos Graph 1: Breakdown of Insurgents, those supporting the Juntas, based on Origin. refer
                    to Creoles within the city of Quito. Each part of the graph resembles a local within the Andean region (except Spaniards). Source: Carta de Ramón Nuñez del Arco sobre
                    la Junta Suprema en Quito, Quito, 20 May, 1813, Quito 257, AGI.</caption>
            </figure>
          - <figure>
                <path pg_no="82" conf="0.91" bbox="[264.52, 214.69, 1524.16, 1136.64]">static/107411/detected_images/107411_82_CS3H.jpg</path>
                <caption pg_no="82" conf="0.87" bbox="[267.04, 1181.98, 1510.18, 1508.23]">Graph 2: A breakdown of Royalists, those against the Juntas, based on Origin. Each part of the
                    graph resembles a local within the Andean region (except Spaniards). Source: Carta de Ramón Nuñez del Arco sobre la Junta Suprema en Quito, Quito, 20 May 1813,
                    Quito 257, AGI.</caption>
            </figure>
        </figures>
        <tables/>
    </section>
</chapter>
```

TASKS COMPLETED

Back Section


```
- <back>
    <ref_heading pg_no="92" conf="0.36" bbox="[793.52, 373.18, 994.59, 441.44]">References</ref_heading>
    <ref_text pg_no="92" conf="0.75" bbox="[259.87, 464.55, 1540.44, 1909.08]">Primary Sources – Archivo General de Indias (AGI) Seville, Spain Cuentas de caja de Quito, 1810-1817, Legajo 256-260, 262,275. Cinco Cartas sobre la revolución de Quito, 1809, Estado, 72. – Archivo Nacional de Madrid (ANM) Madrid, Spain Published Primary Sources Quito: 1809-1812, Según los Documentos del Archivo Ribadeneira, Alfredo Ponce. Nacional de Madrid. Madrid: Imprenta, 1960. A Historical and Descriptive Narrative of Twenty Years Residence Stevenson, William B. In South America: Volume III. London: Hurst, Robinson, and Company, 1825. Secondary Literature Sovereignty and Revolution in the Iberian Atlantic, Adelman, Jeremy, Princeton, New Jersey: Princeton University Press, 2006. Imagined Communities: Reflections on the Origins and Spread of Anderson, Benedict, Nationalism, London: Verso, 1991. Andrien, Kenneth J. "Economic Crisis, Taxes, and the Quito Insurrection of 1765," Past and Present, 129 (1990), 104-131. "Soberanía y Revolución _____. en el Reino de Quito, 1809-1810." Presented for the El Umbral de las revoluciones hispánicas: el bienio 1808-1810. April 10-11, 2008. The Kingdom of Quito, 1690-1830: The State and Regional Development. _____. Cambridge: Cambridge University Press, 1995. Spain and the Loss of Anna, Timothy. America, Lincoln: University of Nebraska Press, 1983.</ref_text>
    <ref_text pg_no="93" conf="0.92" bbox="[257.59, 330.74, 1536.18, 1901.38]">The Wars of Independence in Spanish Archer, Christon I, America, Wilmington, Delaware: Scholarly Resources, 2001. Familia, Honor, y Poder: La Nobleza de la Cuidad de Quito en la Buschges, Christian, Época Colonial Tardía (1765-1822), Quito: FONSAL, 2007. 1808: La eclosión juntera en el mundo Chust, Manuel. hispano, México City: El Colegio de México, 2007. Ideology: An Eagleton, Terry, Introduction, London: Verso, 1991. Cultural Theory: The Key Edgar, Andrew and Sedgwick, Peter, Concepts, London: Routledge, 2002. Echeverri, Marcela, "Popular Royalists, War, and Politics in Southwestern New Granada, 1808-1820," Presented for the International Seminar on the History of the Atlantic World, 1500-1825, Cambridge Mass, 2008. Elliott, John Huxtbole. "A Europe of Composite Monarchies," Past and Present 42 (1969). The Baron de Carondelet as Agent of the Bourbon Reforms: A Fiehrer , Thomas Marc, Study of Spanish Colonial Administration in the Years of the French Revolution, Dissertation, Tulane University, New Orleans 1977. La Revolución de Quito del 10 de Agosto de 1809, Gabriel Navarro, José. Quito, 1962. Gilmore, Robert L. "The Imperial Crisis, Rebellion, and the Viceroy: Nueva Granada in The Hispanic American Historical 1809," Review, Vol. 40, No. 1 (Feb. 1960). Historia eclesiástica del Ecuador desde los tiempos de la González Suárez, Fredrico. conquista hasta nuestros días, Quito: Imprenta del Clero, 1901. Historia general del república del Ecuador. _____. Quito: Imprenta, 1890. Modernidad E Independencias: Ensayos sobre las revoluciones Guerra, François-Xavier, hispánicas, Madrid: Editorial Mapfre, 1992. Hamerly, Michael T. "Selva Alegre, President of the Quiteña Junta of 1809: Traitor or Patriot," Hispanic American Historical Review 48 (1968).</ref_text>
    <ref_text pg_no="94" conf="0.94" bbox="[257.58, 290.33, 1522.33, 1902.14]">Upholding Justice: Society, State, and the Penal System Herzog, Tamar. (1650-1750), Ann Arbor: University of Michigan, 2004. "Las Primeras Juntas Quiteñas," La Independencia en Los Landázuri Camacho, Carlos. Países Andinos: Nuevas Perspectivas. Quito: Universidad Andina Simón Bolívar, 2004. Nueva Historia del Ecuador: _____."La Independencia del Ecuador: 1808-1822," Volumen 6, Independencia Y Periodo Colombiano. edited by Enrique Ayala Mora, Quito: Corporación Editora Nacional, 1983. Breve Historia contemporánea del Lara, Jorge Salvador, Ecuador, México City: Fondo de Cultura Económica, 1994. The Spanish American Revolutions 1808-1826. Lynch, John. New York: Norton & Company, 1973. Actas en La Revolución quiteña, María Borrero, Manuel. 1809-1812, Quito: Editorial Espejo, 1962. Historia de la Revolución de la República de Colombia, Ed. by Manuel, Restrepo, José. Jorge Salvador, La Revolución de Quito 1809-1812, según los primeros relatos e historias por autores extranjeros, Quito: Corporación Editora Nacional (1982). Marchena Fernández, Juan. "Los Procesos de Independencia en los Países Andinos: Ecuador y Bolivia," Debates sobre las Independencias Iberoamericanas. edited by Manuel Chust and José Antonio Serrano. Madrid: Iberoamericana, 2007. "The „Rebellion of the Barrios": Urban Insurrection in Bourbon McFarlane, Anthony. Hispanic American Historical Review, Quito," 69, (1989), 283-330. Independence and Revolution in McFarlane, Anthony & Eduardo Posada-Carbó, ed. Spanish America: Perspectives and Problems, London: University of London, 1999. The People of Quito, 1690-1810: Change and Unrest in the Minchom, Martin. Underclass. Boulder: Westview Press, 1994. Sublevaciones Indígenas en la Audiencia de Quito: Desde Moreno Yánez, Segundo. comienzos del siglo XVIII hasta finales de la Colonia. Quito: Universidad Católica, 1976.</ref_text>
```

TASKS COMPLETED

XML with object ID


```
- <front>
    <title obj_id="6452748" pg_no="0" conf="0.89" bbox="[337.92, 174.19, 1450.04, 305.56]">IMMIGRATION, REGIONAL RESILIENCE, AND LOCAL ECONOMIC DEVELOPMENT POLICY</title>
    <author obj_id="6452749" pg_no="0" conf="0.78" bbox="[815.84, 1033.06, 992.56, 1093.26]">Xi Huang</author>
    <university obj_id="6452751" pg_no="0" conf="0.36" bbox="[659.27, 1565.04, 1132.3, 1667.64]">Georgia State University Georgia Institute of Technology</university>
    <degree obj_id="6452750" pg_no="0" conf="0.81" bbox="[626.15, 1215.67, 1169.21, 1367.24]">In Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in Public
      Policy</degree>
    <committee obj_id="6452755" pg_no="1" conf="0.88" bbox="[256.73, 1088.42, 1531.76, 1766.62]">Approved by: Dr. Cathy Yang Liu Dr. Juan Rogers Andrew Young School of Policy Studies
      School of Public Policy Georgia State University Georgia Institute of Technology Dr. Ann-Margaret Esnard Dr. Ross Rubenstein Andrew Young School of Policy Studies Andrew Young
      School of Policy Studies Georgia State University Georgia State University Dr. Carlianne Patrick Andrew Young School of Policy Studies Georgia State University</committee>
    <date obj_id="6452752" pg_no="0" conf="0.78" bbox="[775.36, 1688.26, 1019.96, 1741.81]">December 2017</date>
  - <abstracts>
    - <abstract>
        <abs_heading obj_id="6452758" pg_no="2" conf="0.44" bbox="[257.94, 342.48, 1529.8, 1934.24]">Abstract</abs_heading>
        <abs_text>I would like to thank my committee members for their critical reviews, wise advisements and unwavering support throughout the process of writing this dissertation.
          Without them, this project would not have been possible. It is my privilege to have Cathy Yang Liu as a mentor, collaborator, and friend. She provided valuable guidance and
          warm encouragement at each step of this project. Her knowledge and advice has been a nurturing force in my academic development and will continue to shape my growth as a
          scholar in future positions. Ann-Margaret Esnard supported this project since its inception, and her expertise on resilience broadened my perspectives. She also went beyond
          the mandate of a committee member and provided insightful advice on how to navigate the academic world. Carlianne Patrick, an urban economist, offered a critical eye to the
          methodologies, which helped ensure the empirical rigor of this work. Juan Rogers pushed me to improve this project theoretically with his excellent and thoughtful feedback
          and comments. Ross Rubenstein provided timely feedback on the research design of this research and has been a source of ideas and advice. Together they have made
          dissertation writing a rewarding experience and set the examples that I hope to achieve in the future. I am also grateful for the Andrew Young School of Policy Studies, the GSU
          Foundation, and the Coca-Cola Scholars Foundation for the financial support of this dissertation. Outside of my dissertation committee, Greg Lewis has been a supportive
          mentor through my time at Georgia State University. His passion for research and pursuit of excellence showcases the essence of a true scholar. I am grateful to Joseph Hacker
          andI would like to thank my committee members for their critical reviews, wise advisements and unwavering support throughout the process of writing this dissertation.
          Without them, this project would not have been possible. It is my privilege to have Cathy Yang Liu as a mentor, collaborator, and friend. She provided valuable guidance and
          warm encouragement at each step of this project. Her knowledge and advice has been a nurturing force in my academic development and will continue to shape my growth as a
          scholar in future positions. Ann-Margaret Esnard supported this project since its inception, and her expertise on resilience broadened my perspectives. She also went beyond
          the mandate of a committee member and provided insightful advice on how to navigate the academic world. Carlianne Patrick, an urban economist, offered a critical eye to the
          methodologies, which helped ensure the empirical rigor of this work. Juan Rogers pushed me to improve this project theoretically with his excellent and thoughtful feedback
          and comments. Ross Rubenstein provided timely feedback on the research design of this research and has been a source of ideas and advice. Together they have made
          dissertation writing a rewarding experience and set the examples that I hope to achieve in the future. I am also grateful for the Andrew Young School of Policy Studies, the GSU
          Foundation, and the Coca-Cola Scholars Foundation for the financial support of this dissertation. Outside of my dissertation committee, Greg Lewis has been a supportive
          mentor through my time at Georgia State University. His passion for research and pursuit of excellence showcases the essence of a true scholar. I am grateful to Joseph Hacker
          and</abs_text>
      </abstract>
    </abstracts>
  - <tocs>
    - <tocs>
        <toc_heading obj_id="6452763" pg_no="4" conf="0.84" bbox="[692.89, 178.13, 1098.83, 256.01]">TABLE OF CONTENTS</toc_heading>
        <toc_text>ACKNOWLEDGEMENTS ..... iv LIST OF
          TABLES ..... vii LIST OF FIGURES ..... viii

  </paragraphs>
  <footnotes/>
  <algorithms/>
  <equations/>
  - <figures>
    - <figure>
        <path obj_id="6452897" pg_no="46" conf="0.89" bbox="[219.77, 240.51, 1489.41, 1148.6]">static/41469/detected_images/41469_46_M560.jpg</path>
        <caption obj_id="6452898" pg_no="46" conf="0.66" bbox="[267.47, 1182.79, 1509.63, 1274.47]">Figure 2.3 Two-way scatter plot of RCI score 2010 and change in foreign-
          born share 2000-2010</caption>
```

TASKS COMPLETED

API integration

TASKS COMPLETED

- Team 3 APIs

Request Method	Description	Request parameters	Database/File System and Response (if any)
GET	Get ETD ID of the PDF file	-	Database Table: etds
GET	GET ETD with ETD ID	-	File System

POST	Save page image (page of PDF)	data = {
		etds_id = ETD ID,
		page_no = page number of the page,
		page_image = .jpg image,
		}
			Table: etd_pages and File System.
			Responds with pageID

GET	Retrieve page image with pageID	data = {
		pageID = page ID received after saving the page.
		}
			File System

TASKS COMPLETED

Request Method	Description	Request parameters	Database/File System and Response (if any)

POST	Populate the ETD_metadata table with the ETD metadata extracted	data = {
		etds_id = ETD ID,
		title = Title of the ETD,
		author = Author of the ETD,
		institution = Institution Name extracted ,
		department = ,
		committee = List of committee members,
		year = Extracted from the date ETD was published,
		degree = Degree ,
		degree_level = Degree level,
		rights = List of Rights (eg:"Open Access", "Attribution required"),
		keywords = List of keywords extracted from text,,
		abstract = abstract text,
		abstract_additions = ["Additional note 1", "Additional note 2"]
		}
			Database Table: ETD metadata

TASKS COMPLETED

Request method	**Description**	**Request Parameters**	**Database/File System and Response (if any)**

POST	Store the detected objects	data = {
		etd_id = ETD ID of the ETD,
		object_type = Text category (paragraph, abstract, etc.) or Image category (figure, equation, etc),
		object_format = Text or Image
		object_text = Text extracted from text based objects (null if storing image)
		object_image = .jpg of image-based object (null if storing text)
		}
			Database Table: objects and File System.
			
			Responds with objectID

POST	Store the generated XML file	data = {
		File = .xml file generated for the ETD
		etds_id = ETD ID
		landing = provide landing/url
		}
			Filesystem

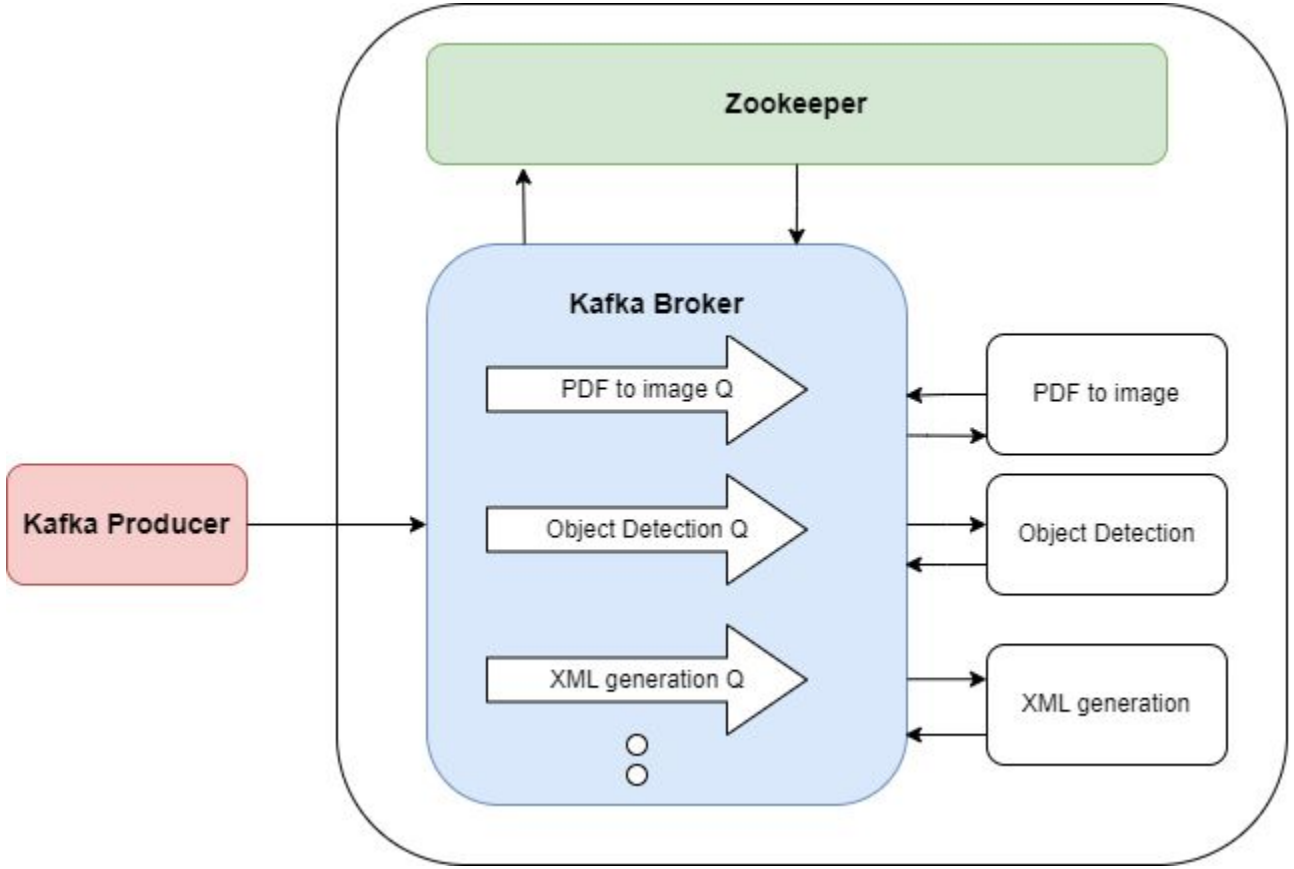
TASKS COMPLETED

- Image generation for 450,000 ETDs, turning each page into an image.

FUTURE OBJECTIVES

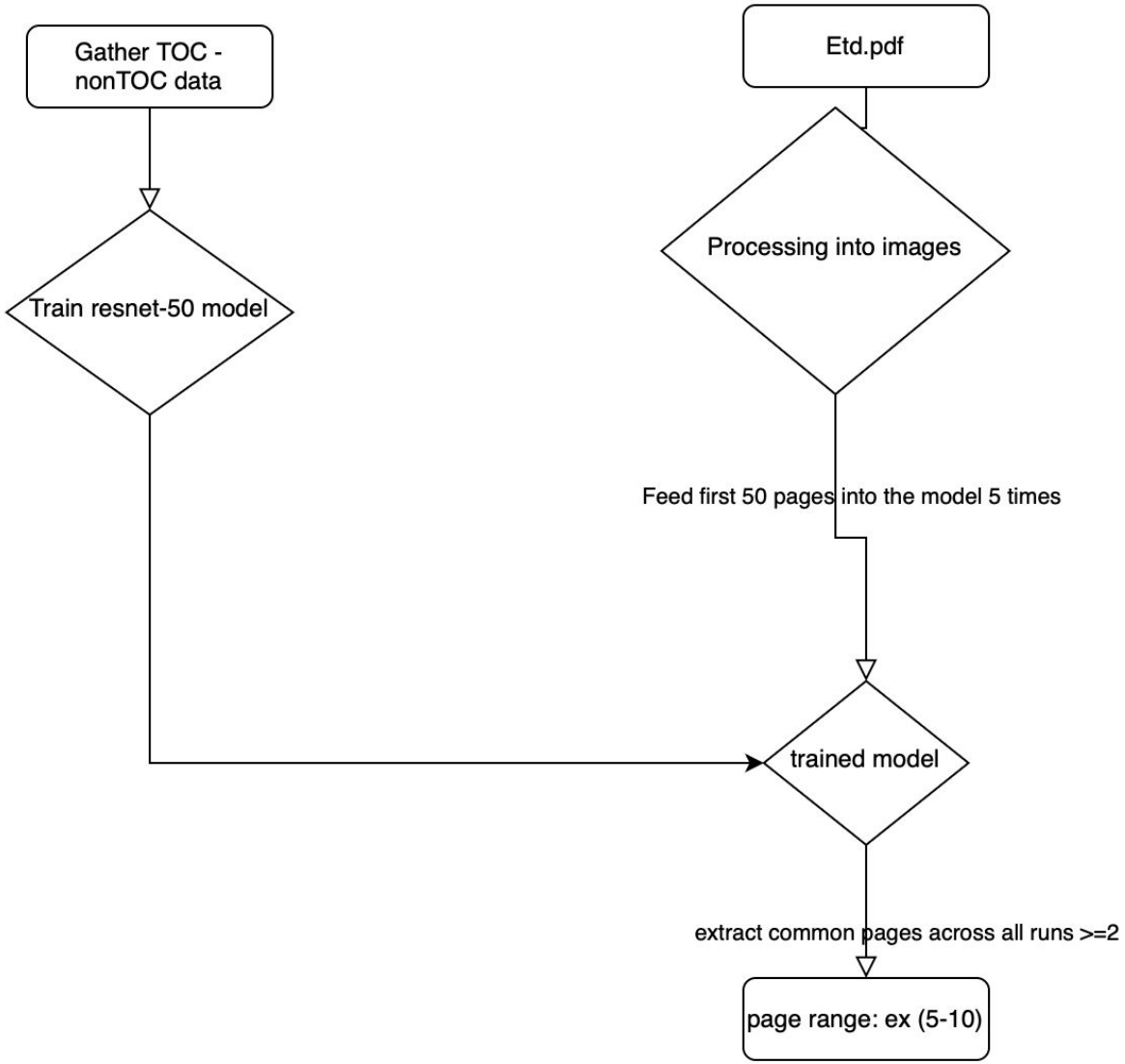
- Improve the object detection model and experiment with models such as YOLO v8.
- Running object detection algorithm on the images generated by the image generation algorithm for the 500,000 ETDs.
- Improve the object detection pipeline to use Kafka so it can speedily handle batches.
- Populate both the database and the file system with the extracted data and files generated by passing the ETD through the object detection pipeline using the integrated APIs.

Architecture for bulk processing ETDs



Chapter Segmentation

METHODOLOGY: EXTRACTING TOC PAGES



METHODOLOGY: EXTRACTING TOC PAGES

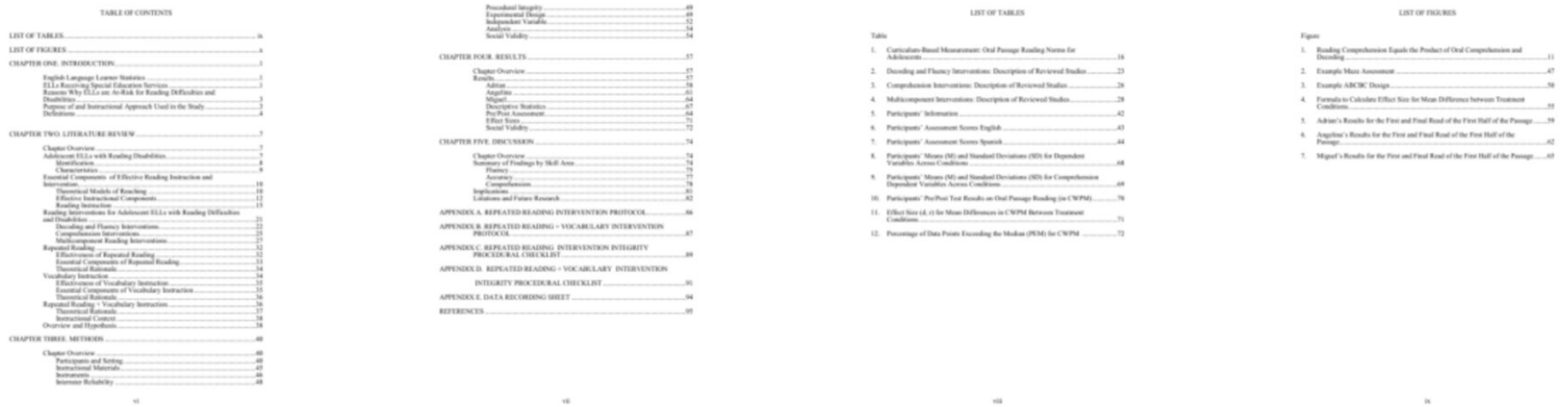
```
Common TOC pages for ./mydata/all_selected_folders/Educational_psychology/3449297/3449297.pdf: [7, 8]
```

CONTENTS

METHODOLOGY: EXTRACTING TOC PAGES

```
Common TOC pages for ./mydata/all_selected_folders/Civil_engineering/3451486/3451486.pdf: [8, 9, 10, 11, 12]
```

```
Common TOC pages for ./mydata/all_selected_folders/Special_education/3628396/3628396.pdf: [9, 10, 11, 12]
```



METHODOLOGY: LLM APPOACH

- Parse each of the table of contents pages by blocks (the basic parsing unit of most parsing libraries).
- Extract lines from blocks.
- Construct similarly structured text with similar indentation to the PDF text version.
- Pass in a prompt “turn this table of contents page into a hierarchical JSON object that has each chapter title and starting page:
  [text]
  ”
- Feed the prompt to LLaMA 2 13B model with long context such as the together-api LLaMA 2 model version.

METHODOLOGY: Using ChatGPT

```
messages=[
    {"role": "user", "content": f"""
     Given the following Table of Contents segment, Convert it into a hierarchical JSON
     representation, where each entry consists of the item name, start page, end page, and subsections."""
     },
    {"role": "assistant", "content": "great i can do that, can you give me an example of input to output"},
    {"role": "user", "content": "sure, here is an example "+"""
     Input:
        Chapter 1: Title . . . . . 1-4
        1.1 Subsection . . . . . 2-3
        1.1.1 Sub-subsection . . . . . 2
        1.2 Subsection . . . . . 4
        Output:
        {{
            "Chapter 1: Title": {{
                "start_page": 1,
                "end_page": 4,
                "subsections": {{
                    "1.1 Subsection": {{
                        "start_page": 2,
                        "end_page": 3,
                        "subsections": {{
                            "1.1.1 Sub-subsection": {{
                                "start_page": 2,
                                "end_page": 2,
                                "subsections": {{}}
                            }},
                        }}
                    }},
                    "1.2 Subsection": {{
                        "start_page": 4,
                        "end_page": 4,
                        "subsections": {{}}
                    }}
                }}
            }}
        }}
     """
     },
    {"role": "assistant", "content": "can you provide the table of contents text that you want to carry this operation on."},
    {"role": "user", "content": """
        {toc_text}
     """
     }
```

METHODOLOGY: Using LLaMA 2-7b-32k

```
6.3.3 Policy Generator . . . . . 148
6.4 Implementation of DOM-ACP . . . . . 149
6.4.1 V8 Binding . . . . . 150

ix

6.4.2 Supporting Access Control Policy . . . . . 151
6.4.3 Enforcing Access Control Policy . . . . . 152
6.4.4 Generating Access Control Code . . . . . 153
6.4.5 Summary . . . . . 155
6.5 Case Studies . . . . . 155
6.5.1 Protecting Real Websites . . . . . 155
6.5.2 Protecting e-Commerce and e-Pay . . . . . 156
6.6 Performance Evaluation . . . . . 158
6.7 Related Work . . . . . 159
6.8 Summary . . . . . 161
"""


model = AutoModelForCausalLM.from_pretrained("togethercomputer/Llama-2-7B-32K-Instruct",
    trust_remote_code=True, torch_dtype=torch.float16).to("cuda")
prompt=f"""
[INST]
Given the toc content below:
{text}
can you tell me where each chapter begins and where it ends in this json format "chapter's title: {{title:begin:1, e
[/INST]
""", return_tensors="pt").to("cuda")
output = model.generate(input_ids, max_length=6000, repetition_penalty=1.1,top_p=0.7, top_k=50,temperature=0.1)
output_text = tokenizer.decode(output[0], skip_special_tokens=True)

print(output_text)
```

METHODOLOGY: HEURISTIC APPOACH #1

- Parse each of the table of contents pages by blocks (the basic parsing unit of most parsing libraries).
- Extract lines from blocks.
- Extract spans from lines.
- Align all the spans horizontally, keeping track of indentation level to construct functional lines.
- Extract page numbers at the end of each line.
- Construct JSON object based on the indentation level recursively.

METHODOLOGY: HEURISTIC APPOACH #1

TABLE OF CONTENTS

METHODOLOGY: HEURISTIC APPOACH #2

- Parse each of the table of contents pages by blocks (the basic parsing unit of most parsing libraries).
- Extract lines from blocks.
- Extract spans from lines.
- Group by clustering column-wise numbers detected from the spans.
- Take the right-most column to be the page numbers detected for chapters/subchapters.

METHODOLOGY: HEURISTIC APPOACH #2

TABLE OF CONTENTS

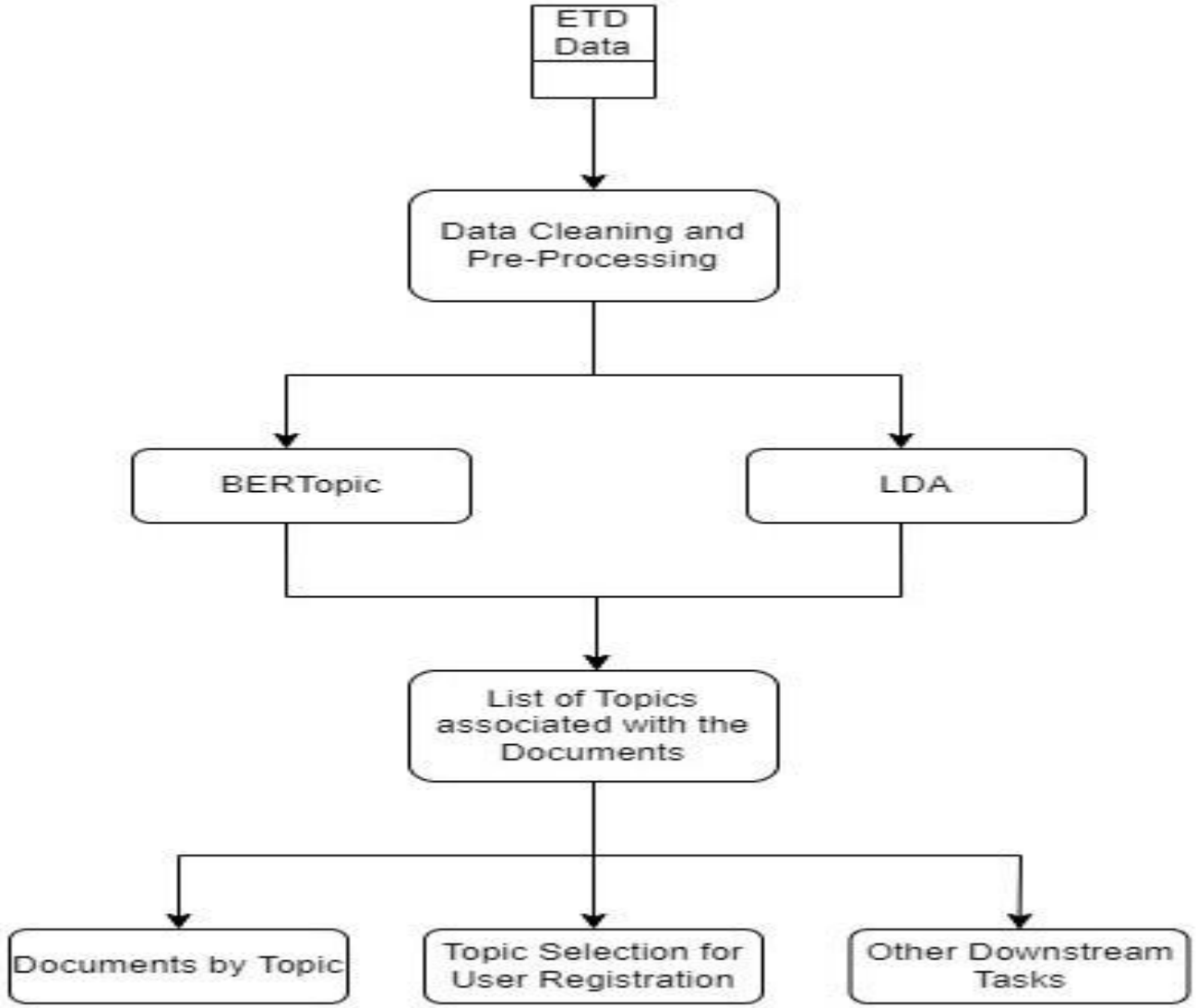
FUTURE WORK

- Merge heuristic approach 1, 2 in order to get a hierarchical JSON object.
- Run the merged algorithm on both OCR and digital data.
- Generate training data for LLaMA 2-7b fine tuning.
- Fine tune LLaMA 2-7b to generate JSON object dependently.

Topic Modeling

TASKS COMPLETED

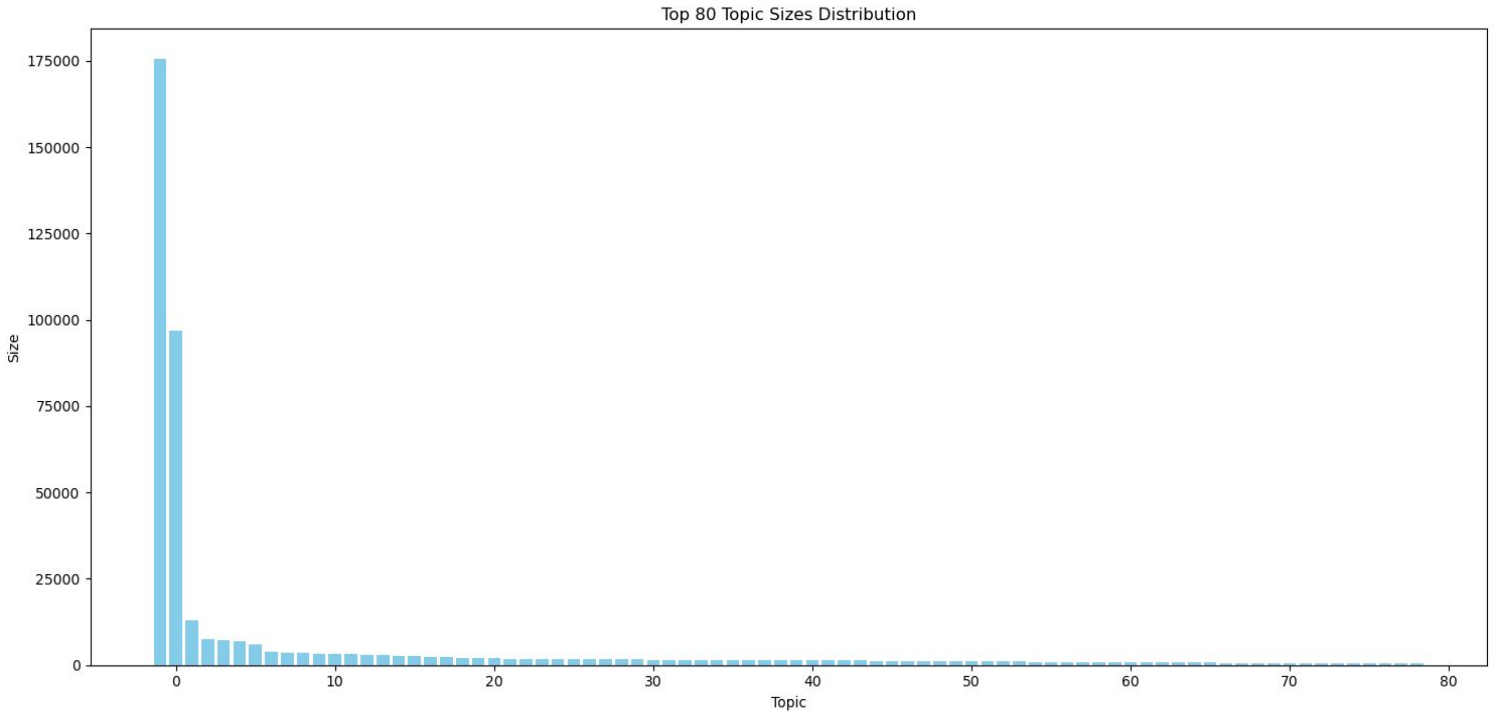
- Topic modeling setup in Endeavour.
- Topic modeling run on cleaned dataset of ~334k ETDs.
- Updated the LDA code to generate JSON, CSV and TSV formats of topic modeling results to be used by Team 2 and Team 6.
- Defined the required APIs to populate the database with topic modeling data corresponding to the ETD.


ETD Data
Data Cleaning and Pre-Processing
BERTopic
LDA
List of Topics associated with the Documents
Documents by Topic
Topic Selection for User Registration
Other Downstream Tasks

BERTopic

Topic modeling:

- BERTopic (Bidirectional Encoder Representations from Transformers) was the topic modeling technique used
- 100 topics were found among 423589 ETDs.
- Each topic is a set of 10 words.
- 53 topics have more than 1000 ETDs.
- Topic 0 has ETDs with incomplete data: 96898 ETDs. .




```
Counter({-1: 175533,
         0: 96898,
         1: 12934,
         2: 7537,
         3: 7222,
         4: 6933,
         5: 6140,
         6: 3889,
         7: 3689,
         8: 3475,
         9: 3267,
         10: 3246,
         11: 3148,
         12: 2967,
         13: 2856,
         14: 2640,
         15: 2627,
         16: 2304,
         17: 2264,
         18: 2056,
         19: 1964,
         20: 1950,
         21: 1851,
         22: 1849,
         23: 1844,
         24: 1763,
         25: 1755,
         26: 1751,
         27: 1716,
         28: 1670,
         29: 1605,
         30: 1587,
         31: 1536,
         32: 1533,
         33: 1519,
         34: 1516,
         35: 1481,
         36: 1470,
         37: 1453,
         38: 1430,
         39: 1388,
         40: 1383,
         41: 1380,
```

```
{-1: [('study', 0.005805585495941566),
  ('data', 0.004609646736949144),
  ('using', 0.004579859552806581),
  ('also', 0.004516867985271672),
  ('used', 0.004473042240779303),
  ('results', 0.004358180694064183),
  ('research', 0.0043580848697494096),
  ('model', 0.00432249435388675),
  ('two', 0.004230305142624809),
  ('analysis', 0.003960885053166374)],
 0: [('none', 1.8300648724111515),
  ('available', 0.2008447769282615),
  ('availsble', 0.00013680623715374676),
  ('dnot', 0.00013680623715374676),
  ('availalble', 0.00013680623715374676),
  ('noneq', 0.00013680623715374676),
  ('availabl', 0.00013680623715374676),
  ('availabel', 0.00013680623715374676),
  ('decembers', 0.00012965347640520782),
  ('heggie', 0.00010359019390954866)],
 1: [('students', 0.029739771263338505),
  ('school', 0.018028802160341854),
  ('teachers', 0.017396237549079784),
  ('education', 0.016129393783934146),
  ('student', 0.015584055701763968),
  ('study', 0.015278653269911659),
  ('college', 0.012438620820559921),
  ('reading', 0.010592136354200005),
  ('learning', 0.010428222515586987),
  ('research', 0.010070682022437777)],
 2: [('children', 0.014671749108198589),
  ('health', 0.0116635935205887),
  ('sexual', 0.011266476940038753),
  ('study', 0.011247078941805582),
  ('participants', 0.010501776688898857),
  ('parents', 0.010124481054247009),
  ('social', 0.01009143435039494),
  ('hiv', 0.010026333447645771),
  ('relationship', 0.00968379611380501),
  ('asd', 0.00900134608487201)],
 3: [('art', 0.013324896268463834),
  ('dance', 0.009080105257467587),
  ('music', 0.008453900503936603),
```

LDA

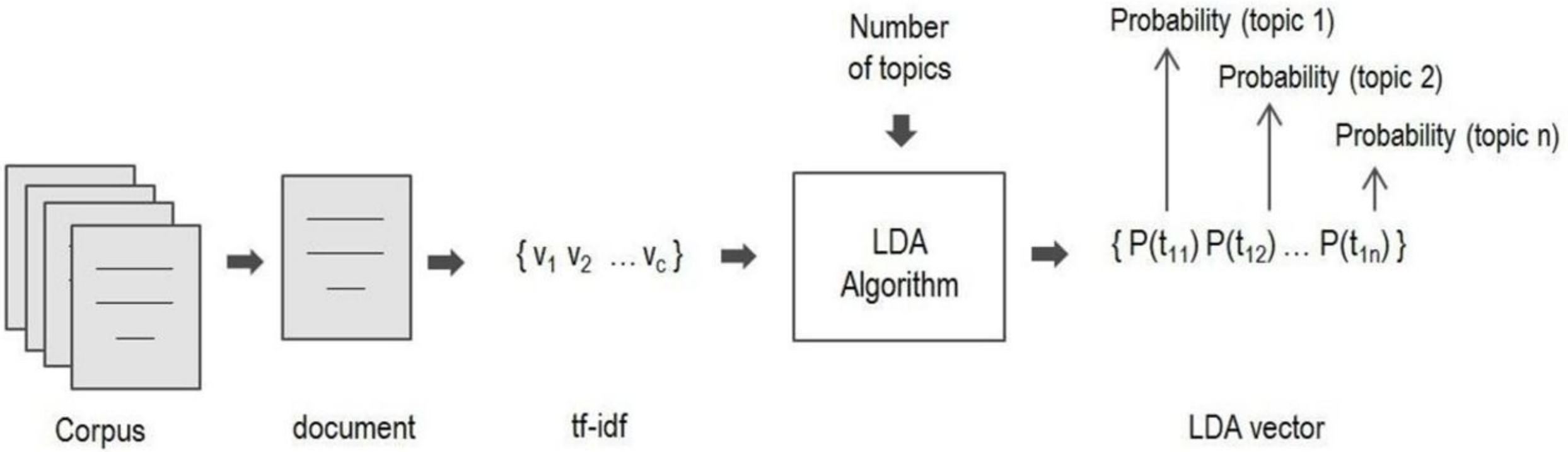
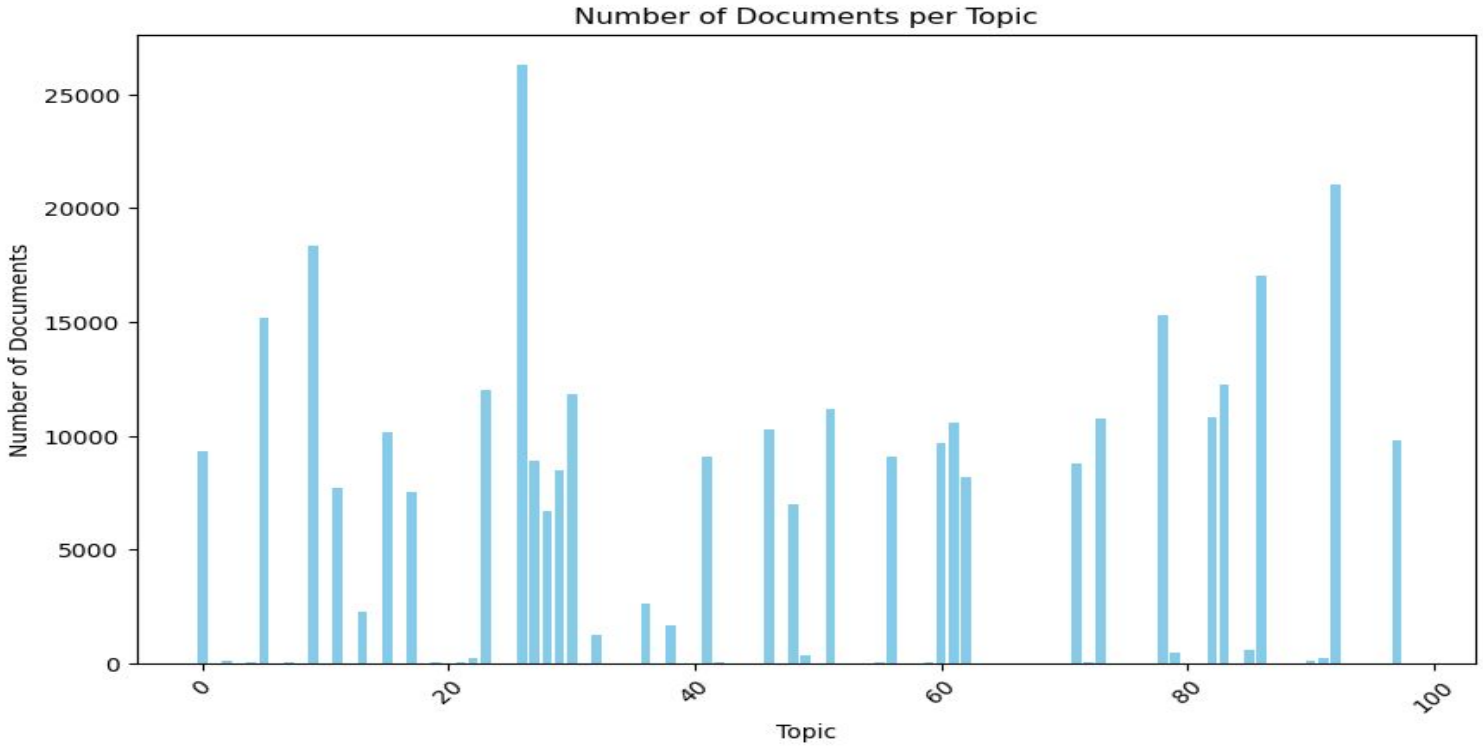


Fig 1. LDA model

LDA

Topic modeling:

- LDA (**latent Dirichlet allocation**) was the algorithm used.
- 73 topics were found among ~334k ETDs.
- Each topic is a set of 10 keywords.
- List of topics passed to team 2 and team 6.




```
['temperature, heat, thermal, temperatures, pressure, high, flow, model, liquid, transfer'
 'algorithm, network, algorithms, problem, system, data, networks, performance, optimization, systems'
 'self, study, participants, relationship, attitudes, perceived, research, social, variables, results'
 'financial, market, model, risk, chapter, find, value, economic, information, impact'
 'architecture, view, thesis, place, space, world, work, nature, full, human'
 'students, school, teachers, education, student, study, learning, teacher, college, schools'
 'cells, cell, expression, protein, mice, signaling, activity, proteins, gene, activation'
 'electron, ion, charge, energy, molecular, dynamics, molecules, state, structure, transfer'
 'vehicle, cost, energy, system, costs, safety, model, design, systems, fuel'
 'works, century, writing, de, music, work, historical, thesis, modern, history'
 'political, government, public, states, economic, international, state, united, policy, national'
 'imaging, optical, detection, sensor, 3d, laser, resolution, system, using, image'
 'community, social, study, research, organizations, interviews, members, organization, communities, managemen
 'model, models, estimation, data, test, method, regression, parameter, methods, distribution'
 'flow, velocity, pressure, model, boundary, motion, fluid, layer, numerical, wave'
 'materials, surface, properties, films, thin, material, magnetic, polymer, film, metal'
 'cultural, identity, war, culture, social, american, political, history, society, century'
 'soil, climate, water, north, area, variability, land, data, change, study'
 'reaction, synthesis, compounds, acid, reactions, metal, carbon, products, chemical, organic'
 'problem, theory, equations, problems, space, finite, solution, method, equation, solutions'
 'power, system, design, frequency, devices, device, performance, control, high, channel'
 'structures, study, 05, 10, 100, 12, 15, 20, 25, 30'
 'learning, data, model, neural, models, based, features, system, information, methods'
 'chapter, major, ii, purpose, describes, described, review, four, investigation, three'
 'health, care, patients, risk, patient, medical, mental, study, disease, clinical'
...
```

FUTURE WORK

- Populate tables using API :
  a. “Add to Topic_models” to add one row for a topic modeler.
  b. “Add to Collection_topics” to store the set of topics from that modeler.
  c. “Add to ETD_topics” to store its topic modeling results for the ETDs.
  d. “Add to Object_topics” to store the topic modeling results using the topic model build on the ETD metadata, applied to chapter summaries, which are obtained using the API “Get from Object_summaries”.

FUTURE WORK

- Analyze and implement the LDA model on chapter summaries and try to improve results by running more experiments.
- Analyze BERTopic results and run experiments to improve the model to get better outputs.
- Explore and compare more topic models like ProdLda and NeurLDA with BERTopic and LDA.
- Integrating Kafka for Bulk processing ETDs to assign topics.

Thank You!


UNIVERSITY