­­­

|  |
| --- |
| A black and red cover with white text  Description automatically generated    Welcome to our project, DISCOVER!    Have you ever found yourself overwhelmed by the sheer volume of text data that needs to be processed or read? Maybe you need it for work or a school project.    It can be an overwhelming task, consuming valuable time and resources. But imagine a solution that could change the way you handle text data, streamlining the process and ensuring accuracy. |
| A group of people  Description automatically generated  But before we began, we would like to introduce ourselves.     I am Rachel, I will be going into my junior year at James Madison university studying computer Science with a minor in mathematics and entrepreneurship.  A black board with white text  Description automatically generated |
| **Objective**  A black background with text and icons  Description automatically generated​  There are millions of documents, PDFs, and research reports that remain unused all over the web. If processed, this information can be vital in a variety of different fields. ​  One key tool that can help simplify this process is our product DISCOVER. DISCOVER turns text into uniform data while making the computer do the work by summarizing the big data. This allows individuals to draw conclusions that would not have been clear otherwise or would have taken numerous hours to get there. ​  Now, you may be wondering why this is so important.    By processing human-readable data, we can unlock a wealth of valuable insights that serve as a catalyst for driving future advancements. These insights can be harnessed and utilized across various domains, fueling progress, and igniting innovation along the way. |
| A close-up of a black background  Description automatically generated |

|  |
| --- |
| A screenshot of a phone  Description automatically generated**Tools**  In our project, we utilized a diverse array of tools, including GitLab's source code management for version control, enabling seamless collaboration with team members to track and monitor code changes.    Furthermore, we used Jupyter Notebooks along with Python libraries like spaCy, fuzzywuzzy, tkinter and pandas. This powerful combination efficiently data processing and analysis, providing meaningful insights to drive the progress of our project.  Additionally, we integrated machine learning techniques and BART, a pre-trained deep neural network, to achieve our results in creating abstraction-based summaries.  With this combination of tools, we were able to deliver comprehensive and grammatically correct summaries, greatly enhancing the user experience and maximizing the value of our project. |
| **Approach**  Our project approach was structured to allow us to achieve optimal results.  To start, we conducted brainstorming sessions to plan our project. We assessed the feasibility of each step, considering what was achievable and identifying potential limitations.  Next, we divided the workload into groups to ensure progress. ​    The first group focused on data cleaning, a crucial step involving accurate translation of PDFs into text, as well as the removal of unwanted elements.    The second group engaged in exploratory text analysis, extracting vital information from the text. ​    Lastly, we dedicated a group to develop a Graphical User Interface or (GUI) for short to showcasing the results of the exploratory text analysis.​  Then to validate our outcomes, we conducted manual readings of articles, comparing the keywords and phrases generated by our system against our analysis. Additionally, we extensively tested the GUI, inputting various phrases and verifying our outputs.​    Our agile approach allowed for continuous refinement, resulting in a product that has exceptional accuracy. |

|  |
| --- |
| Product demo:  This is a screenshot of our product. Let's take a closer look.  **Blank slide with word filled in:**  You start by entering your keyword or phrase in the text entry field.  For this example we are going to type machine learning.    **Misspelled word slide:**  oh no, it looks like we have a typo here. In real-life situations, users might accidentally misspell their search terms. But no worries, our search system has got them covered! With our spell check feature, we provide suggestions for the correct spelling.    **Titles show up:**  Then when you enter your keyword correctly into the search bar and hit the 'Search' button, you'll instantly see a list of article titles related to your search in the article result section.    **Keywords and summary:**  Now, as you find interesting article titles, you can simply click on them. Our system will display the corresponding summary text and keywords related to the selected article. This way, you can quickly assess if it matches your interests and read more about the content.    Behind the scenes, our product uses advanced technology to make your experience seamless.   The layout of our product is built using tkinter library by using layout managers to move the elements around the screen.  The summary result is pulled from the abstract summary column of our csv, and the keywords and phrases are pulled from their respective columns.  And our spell check feature uses the fuzzywuzzy algorithm.    With just a few clicks, you'll uncover valuable insights and information from the vast amount of text data. |
| **Innovation**  A group of icons on a black background  Description automatically generatedBy combining data science, machine learning, and GUI, we have created a user experience that is interactive and easy to use.  Our product processes high volume text data files while preserving key concepts through abstraction-based summaries that mimic human language. For example, in the article Mia shared earlier our abstract summary highlighted that the AI tools is still in the initial trials of soft testing which we didn’t realize from manually reading the article.  To enhance readability, we designed a GUI that incorporated the results from previous steps. Furthermore, its retraining capabilities make it a powerful tool which can evolve with other technology.  This focus on innovation sets our product apart, enhancing lifetime and pushing the boundaries of text analysis. |

|  |
| --- |
| A black background with text and icons  Description automatically generated |
| A black background with text and icons  Description automatically generated **Relevance**  DISCOVER's relevance is evident through its efficiency, actionable insights, and domain-specific applications.  Let talk about an example where our goal is to study a competing company. DISCOVER automates article analysis, generates keywords and summaries which would save valuable time and effort. This efficiency be beneficial in any context dealing with large volumes of text data.  Moreover, DISCOVER empowers users with actionable insights by extracting key concepts and enhancing decision-making across various domains. In our example DISCOVER can provide valuable insights about the competing company's strategies, strengths, and weaknesses.  DISCOVER's adaptability showcased through its domain-specific applications, providing insights for different industries. Having the abstract summaries unlocks information that we would not have discovered otherwise. For our example we can apply discover to analyze articles about our competing company operations.  With its broad relevance DISCOVER's capabilities drive success and innovation. As the vast amount of online data continues to grow, DISCOVER's importance increases for technical, corporate, and average consumers alike. |

|  |
| --- |
| A black background with text and icons  Description automatically generated |
| A black screen with yellow and white text  Description automatically generated  **Outcomes**  Our product provides an innovative solution to the challenges of processing high volumes of text data. After the abstract summaries, the remaining text is around 16,000 words, representing approximately 90% less text compared to the original document. With its efficient processing system and GUI design, it has been instrumental in improving productivity and maintaining data integrity. It met our expected outcome of accurately processing text data and saving time. We continue to improve it by evaluating our results. Additionally, the significant reduction in the document's length highlights only the essential information, making it easier for users to grasp the key points efficiently. |

|  |
| --- |
| A cartoon character with a question mark  Description automatically generated  Thank you for listening. Now we will take questions? |
|  |