**Project Topic:** I will be exploring and implementing the PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive Summarization) model to summarize a large corpus of text. As the name states, PEGASUS will provide abstract summarization, meaning it won’t use sentences explicitly found in the text to form a summary. Instead, it will create a brand-new summary with brand-new sentences based on its “understanding” of the corpus.

**Team Members:** Reg Gonzalez (rdg170330)

**Technique/Algorithm(s) to implement:** The technique/algorithm I plan to implement is the PEGASUS model. According to the abstract for the official “PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization” paper written by Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu, “In PEGASUS, important sentences are removed/masked from an input document and are generated together as one output sequence from the remaining sentences, similar to an extractive summary.” Besides the summary itself, there are other evaluation metrics such as the ROGUE score, which can be utilized to analyze the effectiveness of the model’s summarization.

**Dataset:** <https://www.gutenberg.org/cache/epub/14461/pg14461.txt>

This dataset is the entire book of *Letters on Demonology and Witchcraft* by Walter Scott. Using PEGASUS, I will summarize the contents of this book. Unlike other, more traditional datasets, there are no features or instances. This is simply a text file containing the entire contents of the book.

*Note:* As this project progresses, I may or may not change this dataset to another large corpus and/or include more corpora to summarize.

**Coding language/Technique:** I’ll code this project in a Databricks notebook using PySpark. I also plan to use DataFrames to store the contents of the book as well as the abstract summary provided by PEGASUS. The PEGASUS technique/algorithm was already described above.