CS 410 Project Progress Report – Media Match

Team: Kanjana Chandren (kanjana3)

1) Which tasks have been completed?

I have completed the following tasks so far:

a) Cleaned and prepared a dataset that has top selling books over the years and movies of all genres over the years. I then wrote a python script to extract the keyworks and genres in separate columns for each book and movie.

b) Created an Amazon Kendra Index and uploaded the dataset to an S3 bucket. This will be used as the vector database that will provide extra content (RAG framework) to the LLM.

2) Which tasks are pending?

a) Write code to connect the Kendra Index to a foundational model.

b) Build the front end of the chatbot and deploy it to Amazon CloudFront so that it will be publicly useable.

3) Are you facing any challenges?

a) Number of books and movies

There were too many books and movies that have been released, not including books/movies in different languages. I decided to scale back and only include the top 50 for each category in the past 20 years. Additionally for the sake of simplicity I’ve only included books and movies in the English language.

Nothing else so far….