2025-05-02

* Finished web scraping portion of project. I first navigate to login page to extract CSRF token because Ravelry.com requires user to login to view all page content. Then I navigate to the projects section on the pattern page where it lists all posted projects in cards. I am web scraping this data:
  + Username
  + Yarn name and colorway used for project
  + Project notes which act as “reviews” for this purpose of this project.
  + Status: Is the project completed or still unfinished?
  + Date: Date which project was started
  + Url: Link to project notes page for full details
* Contemplating whether to web scrape more data in future, for example: Date which project was completed (if completed)

2025-05-03

After accumulating data, I export as an excel file for later use such as uploading to a database in SQL Server. I also trained a NLP model to try and interpret human language using a [guide](https://medium.com/@robdelacruz/sentiment-analysis-using-natural-language-processing-nlp-3c12b77a73ec) written by Robert De La Cruz.

The data that used to train the model is an Amazon reviews dataset, “which consists of reviews from Amazon customers downloaded from [Xiang Zhang’s Google Drive](https://drive.google.com/drive/folders/0Bz8a_Dbh9Qhbfll6bVpmNUtUcFdjYmF2SEpmZUZUcVNiMUw1TWN6RDV3a0JHT3kxLVhVR2M?resourcekey=0-TLwzfR2O-D2aPitmn5o9VQ). The dataset spans 18 years, including ~35 million reviews up to March 2013. Reviews include product and user information, ratings, and a plaintext review. “

The guide trains the model to only predict ‘positive’ and ‘negative’ sentiments, but I decided to train my model with a third ‘neutral’ sentiment. The Amazon reviews dataset includes review ratings from 1-5 stars and Robert converts reviews with 1- star ratings as ‘negative’ and 2- star ratings as positive, although I am not sure why.

For my model, I converted 1- and 2-star ratings to negative, 3- star as neutral, and 4- and 5- star as positive. However, I am wondering if a 3- star rating is considered positive rather than neutral and I would imagine that it depends on an individual’s personal rating system, which adds some nuance. Regardless, I moved forward with this rating system and produced the following report:

A screenshot of a computer screen

AI-generated content may be incorrect.

2025-05-04

The report for the model shows values lower than I anticipated, but I wanted to try and use the model anyway on the web-scraped data from Ravelry.com to predict sentiments. I added a new column, predictions, into the data frame and looked over the results. In general, I noticed that the model was quite inaccurate for this data. The model itself was not especially precise or accurate to begin with and it seems like I need to reevaluate how I plan to train the model. I also see a potential issue in how using an Amazon reviews dataset could cause problems for predicting sentiment in project notes of a knitting project since they are not extremely similar. It’s possible to create my own dataset to train the model but this would take an extremely long time.