Capstone Project Proposal Template

Notes:

- This should take no more than one hour to complete – the clearer you are about the business problem you're working to solve with your ML-driven solution, the easier your proposal will be to complete
- This will be uploaded to your repo, which will be a part of your final submission
- Due date for submission is 12/9

Instructions:

1. Download this document as a Word Doc
2. Answer each question using a few sentences, at most
3. Save your completed proposal as a PDF
4. Create a project GitHub repo (if you have yet to do so)
5. Add your instructor as a collaborator (username nickmccarty) to your project repo
6. Add your mentor as a collaborator
7. Push your proposal PDF (created in Step 3) up to your repo
8. Copy the URL corresponding to the location of the PDF in your repo
9. Submit the copied URL using this link

Literary Genre Prediction by Textual Content

Updated 08 November 2022

Business Understanding

- What problem are you trying to solve, or what question are you trying to answer?

While this is not a novel project (much to my honest surprise), genre classification can be a tricky subject in publishing. Authors may not know exactly what they have written and publishers may not be able to easily classify or categorize a given text. Is the book more of a horror novel or something like fantasy or even pure fiction? This becomes quite important during the marketting process, however, as readers who are expecting a specific genre will be disappointed should the book fail to meet their expectations.

This project, then, is an effort to assist the classification process by applying machine learning models trained on existing text-genre sets to predict the genres of novel texts.

- What industry/realm/domain does this apply to?

As this is a literary analysis, the publishing industry has a high degree of investment in this topic, though book sellers might be likewise interested.

- What is the motivation behind your project? (Saying you needed to do a capstone project for flatiron is not an appropriate motivation)

Neil Gaiman once gave a brief speech about his understanding of genre. Even he, as the author of a book, was not always sure what genre his book was supposed to be. There are many aspects which play into the classification of a text into this nebulous category, but there are also social expectations which are inherent to such classification. As a reader and as a scientist, it will be interesting to explore the effect of textual content on the classification of books, and I am curious how effectively such content will predict genre.

Data Understanding

- What data will you collect?

This anlaysis will rely on texts available through Project Gutenberg (for the full corpus) and their genres scraped from sites such as Goodreads, Google, and Wikipedia.

- Is there a plan for how to get the data (API request, direct download, etc.)?

Project Gutenberg recommends downloading specific files through their API, and there are existing tools for scraping genres from the web.

- Are the features that will be used described clearly?

I think so? I will be getting whole text bodies from Project Gutenberg and genres from Goodreads/Google/Wikipedia. These processes are well described and documented.

Data Preparation

- What kind of preprocessing steps do you foresee (encoding, matrix transformations, etc.)?

Any NLP comes with some basic preprocessing and cleaning such as case and tense standardization or the removal of stop words and punctuation. I will also use tokenization and other NLP tools to make the information more computer accessible.

- What are some of the cleaning/pre-processing challenges for this data?

The first challenge will be the selection of texts for this analysis. While it would be better to sample the entire set of texts available, that is infeasible on my hardware and budget, so I will have to limit my training and test sets to more manageable numbers. A second challenge will be the pre-processing of the texts for analysis. Because of the size of dataset I hope to use, the manual verification of all results will be impractical, which could be a source of error, though I shall do what I can to minimize this issue.

Modeling

- What modeling techniques are most appropriate for your problem?

I am not yet certain, though it will be a classification problem.

- What is your target variable? (remember - we require that you answer/solve a supervised problem for the capstone, thus you will need a target)

Genre (with a bonus of profit by regression if I can swing it)

- Is this a regression or classification problem?

Classification.

Evaluation

- What metrics will you use to determine success (MAE, RMSE, etc.)?

This will depend on final model selection.

Tools/Methodologies

- What modeling algorithms are you planning to use (i.e., decision trees, random forests, etc.)?

More analysis is needed before model selection.