Mutual Fund Style Classification from Prospectus

In our class, we have seen an example to use the mutual fund prospectus to classify whether a fund use derivatives. In this project, we will use the same date set of mutual fund prospectus to learn the investment style of a mutual fund.

We have two datasets: 1) a collection of mutual fund prospectus, 2) a CSV form on "Mutual Fund Labels". In the latter file, there is a column on "Investment Strategy". There are 3 main types: "Balanced Fund (Low Risk)", "Fixed Income Long Only (Low Risk)", and "Equity Long Only (Low Risk)". There are also four funds of type Long Short Funds (High Risk) and only one fund of another type.

We will first apply RAG using Langchain and OpenAI API to extract mutual fund investment types from the prospectus, compare with the "Investment Strategy" from the "Mutual Fund Labels" CSV file, in order to see how accurate the "Investment Strategy" labels are. Then we will use the labels provided by RAG as our "ground truth" labels to train our classifier.

Goal and Tasks:

Goal of this project is to use the mutual fund text summaries to predict which investment strategy each fund uses.

1. We will focus our analysis on the three main fund types: Balanced Fund (Low Risk), Fixed Income Long Only (Low Risk), and Equity Long Only (Low Risk). Exclude the funds in the prospectus dataset whose Investment Strategy labels in the "Mutual Fund Labels" are not one of above three. Our dataset contains all the funds whose style is one of above three.
2. Using Langchain and OpenAI API to apply RAG to all fund prospectus in the dataset. Ask the RAG algorithm to classify funds into Balanced Fund, Fixed Income Long Only, and Equity Long Only. Ask the RAG algorithm to return supporting evidence. Compare the RAG results with the Investment Strategy labes in "Mutual Fund Labels". Report your comparison results. Pick several funds where the labels from these two sources are not the same. Comment on which source is closer to the truth.
3. Use the classification results from RAG as the "ground truth" to train your classification algorithm next.
   3.1 Split the data into training, validation, and testing.
   3.2 Following the NLP application in class, use the skip-gram model to build a word embedding dictionary from the mutual fund prospectus in the training set.
   3.3 Design a strategy to build knowledge bases associated to aforementioned three main mutual fund styles.
   3.4 Measure distance of each summary to each knowledge base. Design a classification algorithm to predict the investment strategy of each fund.
   3.5 Use validation data to tune your parameters of your classification algorithms.

3.6 Apply your classification algorithm to predict the investment strategy of each fund in the test data. Report your classification results in the test set.

Report:

1. Submit a research report as if you are a consultant team presenting your results to the company who hire your team to do the analytic works.
2. Document your discussion from tasks above and summary them into a final report. Format of the report should follow

- Executive Summary (summarize your goal and your main findings in a nontechnical language)
- Present your results and discuss your results.
- Present your methodology, compare different methods that you have used, why you think one method is better than others.
- Appendix
- Who have done what in this project
- You code (submitted as a separate file)

It is important that everyone needs to contribute equally in the project. Everyone needs to write part of code. Please include a short paragraph on who did what at the end of the report.

Be succinct and concise on your write-up. Keep the main document within 8 pages and leave all the details in Appendix (no limit on Appendix), but ensure you segment the Appendix into separate sections and refer to the corresponding sections in Appendix from the main document.

Grading criteria:

Quality of the write up, discussion of results, the rigorousness of methodology. All team members will get the same grade on the final project, if all members contribute equally.

This is a group project. Please restrict your group members to be no more than 2.