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 apply the similar method to learn the investment style of a mutual fund.

From the NLP application in class, we have two datasets: 1) a collection of mutual fund summaries, 2) a CSV form on "Mutual Fund Labels". In the later 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 Long Short Funds (High Risk) and only one fund of another type.

Goal and Tasks:

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

1. Split the data into training, validation, and testing.
2. Following the NLP application in class, use the skip-gram model to build a word embedding dictionary from the mutual fund summaries in the training set.
3. Design a strategy to build knowledge bases associated to aforementioned three main mutual fund types.
4. Measure distance of each summary to each knowledge base. Design a classification algorithm to predict the investment strategy of each fund.
5. Use validation data to tune your parameters of your classification algorithms.
6. Apply your classification algorithm to predict the investment strategy of each fund in the test data.
7. Instead of building word embedding ourselves, we can also use pre-trained model (for example, sentence Bert) to extract key sentences of each summary. If you use one of pre-trained models, compare the performance of your classification model in the test set with the model using your own word embedding.

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. Sample format of report can be the following

- Executive Summary (summarize your goal and your main finding 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. For single person groups, a small compensation will be awarded.