Background

ESG is the practice of characterizing responsible and sustainable investments with three components: environmental, social, and governance. Investors who incorporate ESG into their investment strategy evaluate the environmental and social impact of a firm and the behavior of its corporate management. Important environmental concerns include greenhouse gas emissions, carbon footprint, and pollution. Social concerns include health and safety, workers’ rights, and, for software companies, data security. Governance concerns include CEO pay and corruption.

Responsible investing, ESG in particular, is currently defined very loosely. There is no standard way of evaluating a firm’s environmental, social, and governance attributes. Norway’s sovereign wealth fund is a leader in this area.

There are several companies that provide ESG scores. For instance, Refinitiv, formerly the Financial and Risk business of Thomson Reuters, analyzes stock exchange filings, news reports, and annual corporate reports in order to calculate ESG scores. They rely on a combination of algorithms and human content research analysts to gather and process data.
Overview

In collaboration with the Yale School of Management, I will explore the utility of statistical text analysis in ESG prediction. Because ESG is loosely defined, I will first review existing literature on the topic to understand how the investment community in general conceives of it. I will then determine a set of parameters that are useful for characterizing ESG. Next, I will gather a text dataset to use to predict these parameters. I will likely use similar types of data as Refinitiv, incorporating annual company reports and news reports. Finally, I will build a statistically-robust model to predict the ESG parameters from the text data.

The specific techniques I will use depend heavily on the ESG parameters that I decide to predict. Gathering and cleaning the data will likely involve web scraping and text processing algorithms. I expect my predictive model will incorporate some sort of frequency analysis on words or phrases.

Timeline

The literature review will take no more than a week. I will spend a three to four weeks evaluating various ESG parameters and deciding which to use for prediction. Gathering data should take no more than a few weeks. I will spend the rest of the semester building and testing the predictive model.

Deliverables

- Literature review
• Set of parameters to determine a firm’s ESG value

• Text dataset to use for ESG prediction

• Statistically-robust text-based predictive model

• Written report

References
