CS 410: Text Information Systems
University of Illinois at Urbana-Champaign
Fall 2022

Course Project Proposal

Submitted by: Samuel Kirby
slkirby2@illinois.edu

October 22, 2022

Project Track: Free Topics

Project Title: Use of sentiment analysis on mutual fund quarterly commentaries

Team Members: Samuel Kirby slkirby2 Solo project

Project Description:

There are thousands of mutual funds available to investors, managed by hundreds of different asset management firms. Each quarter, many (if not most) of these managers produce quarterly commentary documents that provide information about how each of their funds performed, what contributed or detracted from performance, and their future outlook.

The analysis of these documents represents a tremendous opportunity to gauge the overall degree of market sentiment amongst a large, diversified group of professional investors, and it is a task that would be impossible to perform manually. This makes it an ideal task for machine learning and sentiment analysis techniques.

I propose the use of sentiment analysis to assess the individual and aggregate market outlook among a large cross-section of asset managers. My expected approach is:

1. Construct a simplistic crawler to download a suitable sample size (I would expect 50-100) quarterly commentary files for the quarter ending 2022-09-30. This crawler would use keywords typically found (or required to be included by regulation), and download the PDF files which is the standard format for such documents.

2. Employ a library/package to extract text from the PDFs. Note that this text will likely be "messy" and unstructured, so some measures may be required to cleanse or pre-process the text before performing sentiment analysis.

3. Perform sentiment analysis on the documents. Because I have not yet reached Week 11 where this topic is covered, I do not yet know the specific technique or algorithm that I will employ.

4. I will either use R (preferred, due to my familiarity), or python, depending upon the availability of required libraries/packages.

5. Evaluation of the work may be challenging. Again, since I have not yet covered Week 11 material I do not yet know the various approaches that could be used, but I would expect that some sort of manually "labelled" test documents could be a viable approach. Another, more complex but potentially useful approach would be to compare the fund manager's sentiment to the actual positioning of their portfolio; however this is likely out of scope for this project.

6. I expect that this project will easily reach the 20 hours threshold for a single-person team. Estimated time required for each step:

    1. Crawler/PDF downloader – 3 hours
    2. Extract text from PDFs and perform cleansing/standardization – 6 hours
    3. Evaluate sentiment analysis approaches – 2 hours
    4. Write R or Python program to perform sentiment analysis – 8 hours
    5. Manually label a set of documents for accuracy evaluation – 2 hours
    6. Report/presentation writing – 2 hours

    Total: 23 hours