Data Science
2016/2017 Assignment

Introduction

If you are considering a career combining quantitative analysis and finance, there are few areas with as much promise as data science. It is allowing for rapid advancement in several major areas of financial services; key applications include:

- Improving credit scores

  Information included in traditional measures of creditworthiness such as FICO scores can be enriched with the addition of your online footprint – your professional background on LinkedIn, who your friends are on Facebook, etc.

- Macroeconomic analysis

  Initiatives like Premise and MIT's Billion Price Project use novel data sources to 'nowcast' key macroeconomic indicators like GDP, allowing economists to view them more or less in real-time.

- Fundamental research

  More and more equity and credit funds are turning to so-called 'alternative data' for an edge in their investment research. This category includes phone location data, transaction records, satellite imagery, and a wide variety of data sourced through web scraping.

For this assignment, your challenge is to find an opportunity to create edge by scraping web pages. Very few students are capable of citing specific alternative data projects in their interviews, so this would be a great way for you to stand out.

Your assignment

Background

Among the dozen-plus businesses selling scraped data to hedge funds, the best-known is YipitData. We would advise you to look through the list of companies they cover – Yipit is able to generate reasonable estimates of quarterly revenues for each of these through web-scraping. For some firms, this is quite straightforward:

- For each of the items it sells, Alibaba lists the quantity sold. Scraping this quantity daily allows you to determine daily sales volumes.
- Carmax, a used-car retailer, lists their inventory online. Tracking changes in inventories allows you to estimate sales.
- Lending Club, an online marketplace connecting retail borrowers and investors, allows you to track each of the loans on its website.

Other firms are more tricky. In some cases the data you want is not displayed publicly but can be found by going through the site source code.

Instructions

Your challenge is to build a scraper for gathering data which can be used to estimate revenues for a publicly listed business, i.e. for a company with its stock traded on a stock exchange. Once the scraper is built, run it **once**; your report will be a recap of that single scrape. **You are not expected to be able to assess company performance off one scrape** – doing so usually requires running a scraper daily for over a year.

A few general tips:

- Look for a direct measure

  It is tempting to try and estimate revenues with something like the number of reviews its stores / branches are receiving on Yelp or the numbers of likes / followers they are receiving on Facebook / Twitter. Such proxies tend to be too inaccurate to be helpful to fundamental analysts.

- Think about materiality

  It is fine to only track revenues for one brand or subsidiary of the company you are looking at, but professionals are unlikely to be impressed by your work if that brand / subsidiary does not meaningfully contribute to the performance of its parent business as a whole.

- Stay cognisant of ToS

  We do not in any way encourage or condone going against the Terms of Service of the website(s) you are scraping. If you are concerned you may be breaking the site(s)'s ToS, please note so in your report.

Feel free to scrape any one of the companies covered by YipitData or its competitors, although we would encourage you to try and find something original.

Your report

The output from your analysis should be a 1-page write-up with the following elements:

- Quick recap of how your data is sourced and how it can be used to estimate company revenues
- Summary stats for one-off run of your scraper, comprised of (1) a list of fields, (2) summary statistics for each field, (3) field descriptions.
- (Optional) Link to your source code on Github.

Summary statistics provided for quantitative fields should include average, max, and min. Please also include total row count.

Your analysis should be neither exhaustive nor complete – we are looking for preliminary findings only, and would encourage you to include a Next Steps section in your report. If you are approved for inclusion in our mentorship program, one of our industry mentors will help you build on your initial work.

We expect this assignment to take 3-5 hours to complete, but you are welcome to spend as much time on it as you deem appropriate. Feel free to contact us if you feel you have hit a dead end.

This is not an offer to sell or the solicitation of any offer to buy any securities.