|  |
| --- |
|  |
| **Team 2**  Abhishek Chintamani  Amit Shukla  Rupali Agarwal  Smaranika Sikdar  Varun Ananthula  Varun Ananthula |

# Data Collection & Pre-Processing

# Group Assignment

# Domain : Indian Stock Market

# 

|  |
| --- |
|  |
|  |

Table of Contents

[Data Collection & Pre-Processing 1](file:///C:\Personal\Google%20Drive\Next%20Level\ISB\Term%201\Data%20Processing\Assignment\DCPP-Assignment\DCPP%20Assignment%20Strategy%20Document.docx#_Toc103723969)

[Group Assignment 1](file:///C:\Personal\Google%20Drive\Next%20Level\ISB\Term%201\Data%20Processing\Assignment\DCPP-Assignment\DCPP%20Assignment%20Strategy%20Document.docx#_Toc103723970)

[Domain : Indian Stock Market 1](file:///C:\Personal\Google%20Drive\Next%20Level\ISB\Term%201\Data%20Processing\Assignment\DCPP-Assignment\DCPP%20Assignment%20Strategy%20Document.docx#_Toc103723971)

[1. Executive Summary 3](#_Toc103723972)

[2. Domain & Seed Sources 4](#_Toc103723973)

[3. Structured & Unstructured Sources 5](#_Toc103723974)

[4. Data Collection 6](#_Toc103723975)

[5. Data Format Conversion 7](#_Toc103723976)

[6. Data Cleaning & Pre-Processing 8](#_Toc103723977)

[7. Observation & Insights 9](#_Toc103723978)

[8. Enhancing data with Crowd Sourcing Methods 10](#_Toc103723979)

[9. References & Sources 11](#_Toc103723980)

## Executive Summary

###### Problem Statement

###### Proposed Solution

###### Challenges

## Domain & Seed Sources

###### Chosen Domain

###### Seed Sources

## Structured & Unstructured Sources

###### Structure Sources:

* Nifty 50 stocks data - Investing.com company page
  + The data source for the Nifty stocks is Investing.com.
  + This source was chosen as the website has the Nifty indices performance data of previous years for the comparison.
  + We needed the Nifty 50 Index components historical data to see the trend of the stocks in Nifty 50 index.
  + The other sources which were considered were nseindia.com, finance.yahoo.com
  + These other sources didn’t have the data we needed
  + These websites only had the current data for the components in Nifty 50, the historical data for the components was not available
  + Some websites had the Nifty 50 index historical data but not at the components level.

###### Unstructured Sources:

* DuckDuckGo Search
  + We found DuckDuckGo to have a good mix of returning **correct results** (hit rate) and supporting **bulk web scraping**
  + Needed to search for the company name on Moneycontrol website to get the company’s Moneycontrol link
  + Tried the following options
    - Moneycontrol Website Search
      * Tried manually searching for company name on their website.
      * The hit rate was low (around 50%) and even for matched cases there was no simple url for replicating this search through code
    - Google Search
      * Google search had a very good hit rate (~98%) when we limit the searches using the ‘site:’ option to the specific Moneycontrol section / link.
      * But Google started blocking my web scraping requests after the first 50 – 70 calls and it was taking more than a day to unblock – hence we could not scale it for the 4.5K companies to be searched
      * Note that we did not explore rotating proxy solutions – as they were too costly and a bit advanced for this exercise.
    - Bing / Microsoft Search
      * Bing’s hit rate was lower (~80%)
      * And even it was blocking web scraping requests within 50 – 70 calls
    - Less popular search engines
      * Tried various less popular search engines such as Dogpile, Ask.com and DuckDuckGo
      * Found a respectable hit rate with DuckDuckGo manual search (~95%)
    - DuckDuckGo
      * Using DuckDuckGo manual search we were able to get a decent hit rate of ~95% - had to do a bit of trial and error and tweak the search string
      * Also we were able to run higher number of requests without getting blocked during our testing phase.
* Moneycontrol Company page
  + Moneycontrol company pages had a good mix of **depth of info** and **ease of scraping** in terms of website structure and allowing bulk scraping.
  + We needed a source for online links for each company – specifically the company homepage (for links to their social handles) and their contact info (for future crowdsourcing steps)
  + Moneycontrol is one of the most popular consumer websites in India for information about public listed companies – so we started with Moneycontrol.
  + Since we were able to get the info we were looking for (and also there was lot more info available for future use) – we did not explore other options
  + Note that we are also storing each company’s Moneycontrol page link in our data – so that in future if we need more details or updated stock market info we can directly scrape it from the respective Moneycontrol pages
* Company home pages
  + As mentioned in the Problem Statement – we wanted to include Twitter info in our data. This is to ensure that in addition to the traditional stock market data we also include some of the metrics that might be impacting the market in these modern times
  + Company home pages ended up being the ideal sources for the company’s twitter handles due to **minimal errors**
  + Tried the following options
    - Twitter search
      * Direct twitter search using company name resulted in too many results for each search and there was no easy way to identify which of the results is the actual handle
      * Especially when most of the results were handles of aggregators and media companies who were reporting news about the company we were searching for.
      * Hence we could not use twitter search
    - Google search
      * Faced similar issues by searching for twitter handles through Google search.
      * Though the quality of the results improved (in terms of company name matching) – but here too there were too many results from aggregator and media company handles
    - Aggregators (Moneycontrol, Valueresearch, etc.)
      * We tried to identify if any of the aggregators provide social handles info for these companies
      * But we could not find this info on popular websites such as Moneycontrol and Valueresearch
      * Our sense is that the traditional sources for stock market info have not yet started tracking modern metrics such as # of Twitter Followers or recent tweets from the company regarding their performance, etc.
    - Company Homepages
      * During this exercise we realized that not all companies were using twitter to share their company updates – especially true in case of small and mid-sized companies in B2B sectors
      * We also noticed that companies that have a twitter account have a high chance of having a good homepage and mentioning their twitter handle on the homepage
      * With a pilot manual search we found a hit rate of ~20% and more importantly for the remaining 80% we could not find an easy way to identify their twitter handles
* Twitter
  + We identified that Twitter is a platform with a good mix of **business relevant info** and **high consumer (and investor) reach**
  + As part of identifying modern metrics that could be correlated to the stock market we wanted to track relevant social metrics such as presence and rate of activity on social media.
  + Tried the following options
    - LinkedIn
      * Very high relevance in terms of business content
      * But it is primarily targeted to working professionals and not really designed to reach a high number of consumers and investors
    - Facebook
      * Very high reach among consumers and investors
      * But low relevance in terms of business content
    - Twitter
      * This is where we found Twitter to be the ideal mix between business relevant info and good reach among consumers and investors
      * Companies worldwide are already using Twitter as a modern route for business communication (in addition to the traditional routes of press releases, etc.)
      * Once we identify the Twitter handle of a company, the info available on this handle is highly relevant for that company – there is no noise or incorrect data

## Data Collection

###### Downloaded Data

* + Nifty 50 stocks data was downloaded from the website <https://in.investing.com/indices/s-p-cnx-nifty-components-fundamental>
    - Nifty scripts current price, historical performance and fundamentals were downloaded from the site.

###### Crawled Data

* DuckDuckGo Search
  + We first explored a python library for DuckDuckGo search – it had already integrated the technical aspects of making the search request and returning the required number of results.
  + but we noticed hit rate dropped to 70% vs 95% during manual testing. Same searches manually were giving correct results.
  + On further deep dive understood that the library is using DuckDuckGo API and it is clearly mentioned by DuckDuckGo that their API does not search the full web (by design)
  + So we used DuckDuckGo search url directly for web scraping
  + after a bit of trial and error and tweaking the search string we were able to get to the hit rate of ~95%
  + A few challenges we faced and the corrective measures taken to overcome them were
    - Got blocked in the first few attempts – realized that we had not setup the search header to ensure request is detected as that from a valid web browser. Had to wait for a day and re-run with the correct search header
    - We were getting blocked in Google and Bing searches after a few attempts – to ensure lower chance of blocking on DuckDuckGo we introduced a sleep timer of 30 secs after each search. I had to setup a dedicated system to run this for a few days for 4.5k companies

## Data Format Conversion

Nifty 50 Stocks data

* This data was in the structured format so there was no need of conversion to structured data fields
* To merge the data to the main sheet we needed the “ISIN” attribute for the Nifty 50 stocks which was missing.
* We added the ISIN attribute column in the Nifty 50 stocks sheet

## Data Cleaning & Pre-Processing

Nifty 50 Stocks data

* There are 50 stocks in the Nifty indices.
* One column was added to identify the Nifty stocks giving the Labels as ‘Yes’ or ‘No’ based on whether the stock is part of the Nifty 50 indices or not.
* The Nifty stocks data was then merged with the stocks main sheet data
* The merging was done on the attribute “ISIN”
* The performance parameters for the rest of the stocks were replaced with zeroes.

## Observation & Insights

## Enhancing data with Crowd Sourcing Methods

## References & Sources