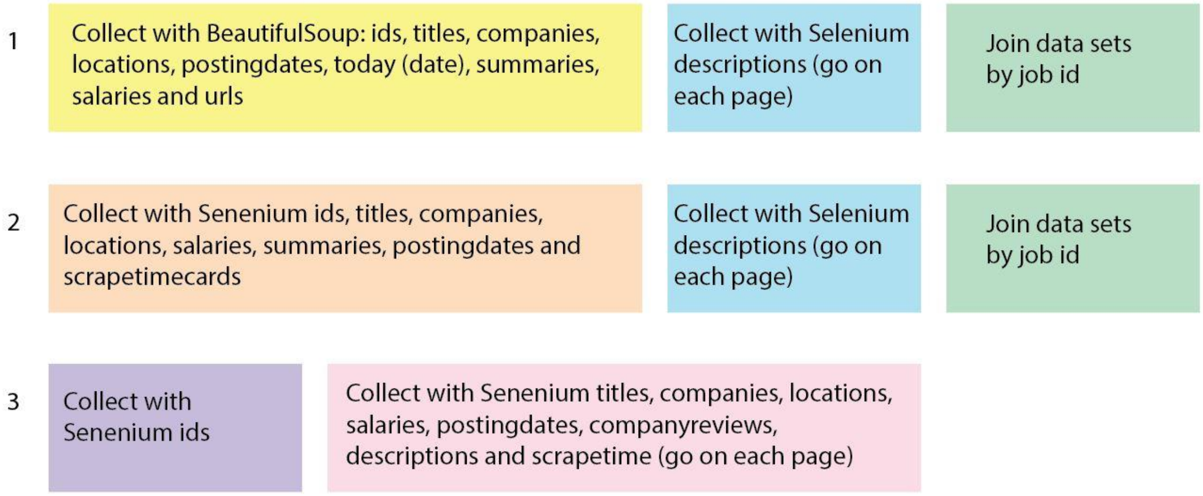
Scraper alternatives

There might be three alternative ways to collect the data from Indeed.com. Each method has its advantages and disadvantages in terms of collection time and quality of data collected.

Brief illustration of the three alternatives:



1

Collect with BeautifulSoup: ids, titles, companies, locations, postingdates, today (date), summaries, salaries and urls

Collect with Selenium descriptions (go on each page)

Join data sets by job id

2

Collect with Senenium ids, titles, companies, locations, salaries, summaries, postingdates and scrapetimecards

Collect with Selenium descriptions (go on each page)

Join data sets by job id

3

Collect with Senenium ids

Collect with Senenium titles, companies, locations, salaries, postingdates, companyreviews, descriptions and scrapetime (go on each page)

Alternative 1

First, use BeautifulSoup to collect, for each job card: id, title, company, location, postingdate, today (date), summary, salary and url. Then use Selenium to loop with the ids through the pages and collect the description. Finally, join the two datasets based on the id column.

- Advantages of this alternative: you get all the proper information from the job cards, and you enhance that data with the job descriptions collected separately.
- Disadvantage: not everything runs in one go, and you have to put it together later on.

Alternative 2

First, loop through the pages and collect from the current page the ids, titles, companies, locations, salaries, summaries, postingdates and scrapetimecards. Then use the ids to access the pages and collect the descriptions. Finally, join the two datasets by the id column.

- Advantage: you only use Selenium, so you are more consistent
- Disadvantages:
    - Many elements of most of the lists you create are returned empty; lists such as titles, companies, locations, postingdates and salaries (this is not a problem like the other examples, because it was already expected to contain missing data, as not all companies provide a salary range).
    - It takes a long time to scrape and only in the end you can assess the magnitude of the missing data. As such, this may be a less preferred alternative.

Alternative 3

First, loop through the pages and collect the ids from the current page. Then, use the ids to access each page and collect: titles, companies, locations, salaries, postingdates, companyreviews, descriptions and scrapetime.

- Advantage: you collect everything in one go.
- Disadvantages:
    - The location is difficult to be captured on the page, and so that column contains random information, such as company reviews.
    - It takes a long time to scrape and only in the end you can assess the magnitude of the missing data. As such, this may be a less preferred alternative as well.

After considering the three scraper alternatives, we decided to go with alternative 1. This method gives us the correct information and thus seems as the best option.