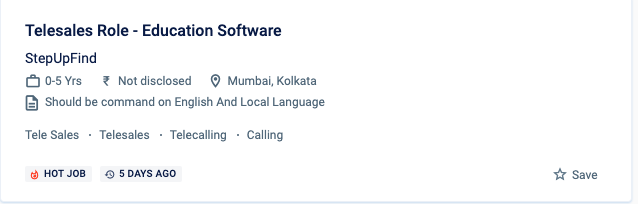
**Technical Assessment - 1**

The final output should have a file with fields: year, filename, count of high competition (with its synonyms) and count of technological competition

**Task 1:** Find the occurrences of the phrase “high competition” (and its synonyms) and “technology competition” (and its synonyms) from the entire 10-K. Synonyms can also include variants of a word like technological competitors, etc.

**Task 2:** Scrape all job postings by area from the links given in the file ‘link\_by\_areas.csv’. For every link, loop through all the job postings by page, collecting all the information for a given posting. The image below is an example of a single posting.



Create an output CSV file with job title, company, experience, salary, location, description, tags associated, function area, posting date, scraping date.

**Task 3**: Given is a list of textual descriptions for around 20,000 companies (provided in the file Company Descriptions). There are 2 types of descriptions – long and short. Use your discretion while using either.

3.1: Classify these companies based on their descriptions to only one of the industries from the industry labels given (provided in the file: Industry Segments – Top 10 Keywords).

Your output should contain the company name and the corresponding industry segment assigned by your algorithm.

3.2: Cluster the companies using only their description. Identify and justify the optimal number of clusters. Do not use the industry labels for this task.

Your output should contain the company name and the corresponding cluster label assigned by your algorithm.

**Please Note**

* These tasks are organised by their order of increasing difficulty and increasing priority (that is, task 3 will have higher importance while assessing your submission and is more difficult than the remaining).
* Data for Task 1: Sample of 30 company annual 10-K filings for years 2010-2017 (attached in the folder 10K).
* Data files for the remaining task can be found here: Technical Assessment
* For submission: send all intermediate CSVs generated (if relevant and useful), along with a well-commented and modularised code and the final output files.