

CO-OP FINDER

JOB RADAR

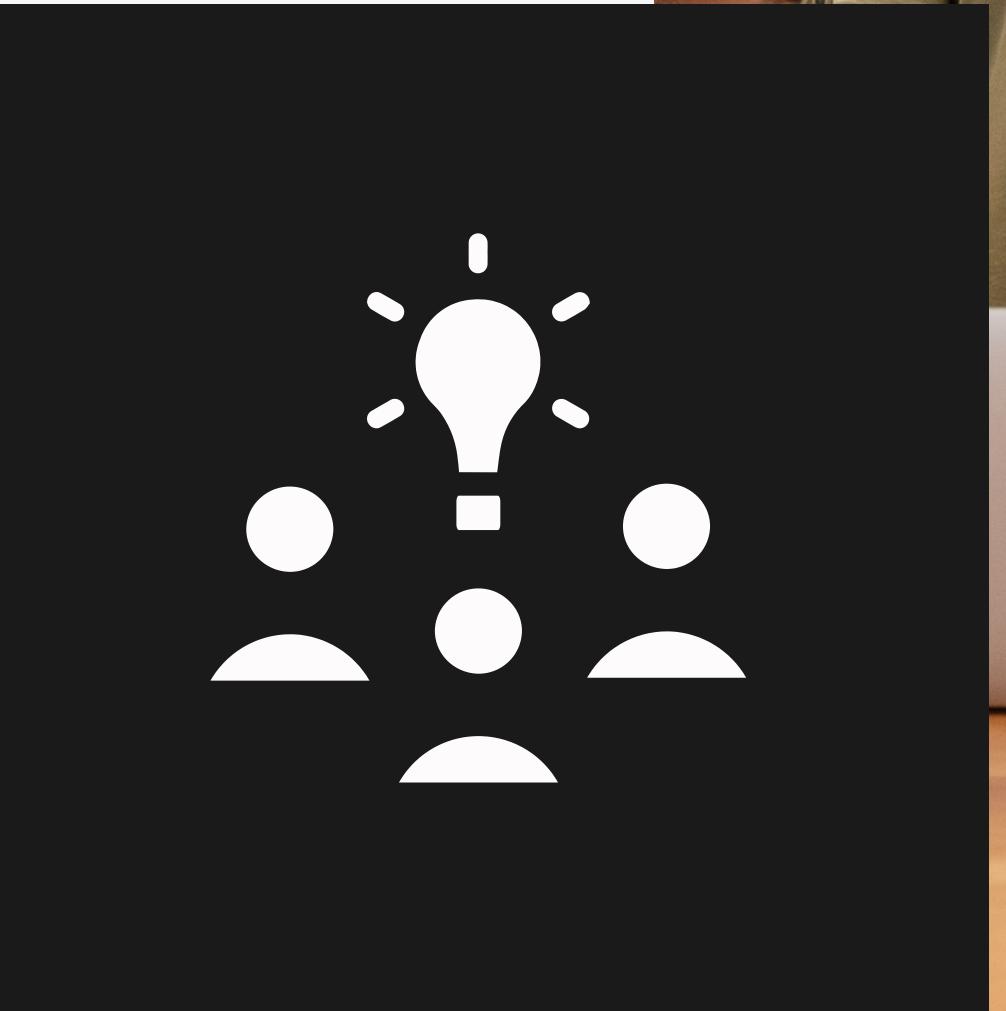
INSTRUCTOR: DR. OLENA SYROTKINA

GROUP 6

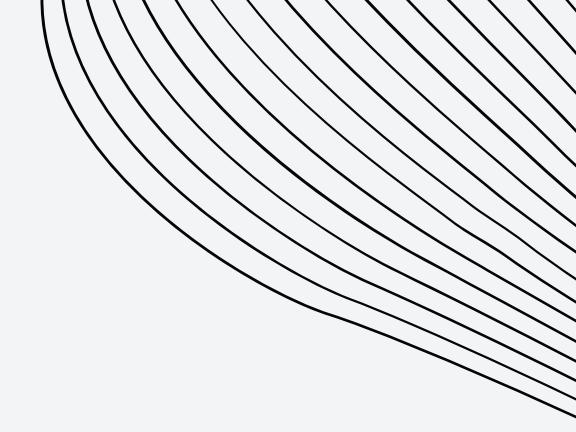
MEET PATEL (110126852)
VIKRAMSINGH BEDI (110125886)
KISHAN MODI (110131833)
MITI TAMAKUVALA (110126404)
APOORV SHAMMA (110125728)

PROJECT DESCRIPTION

We have created a unified platform that consolidates CO-OP job listings from prominent sources such as LinkedIn, Indeed, and Workopolis. This centralized hub serves as a convenient and efficient space for individuals to discover and apply for their CO-OP opportunities effortlessly.



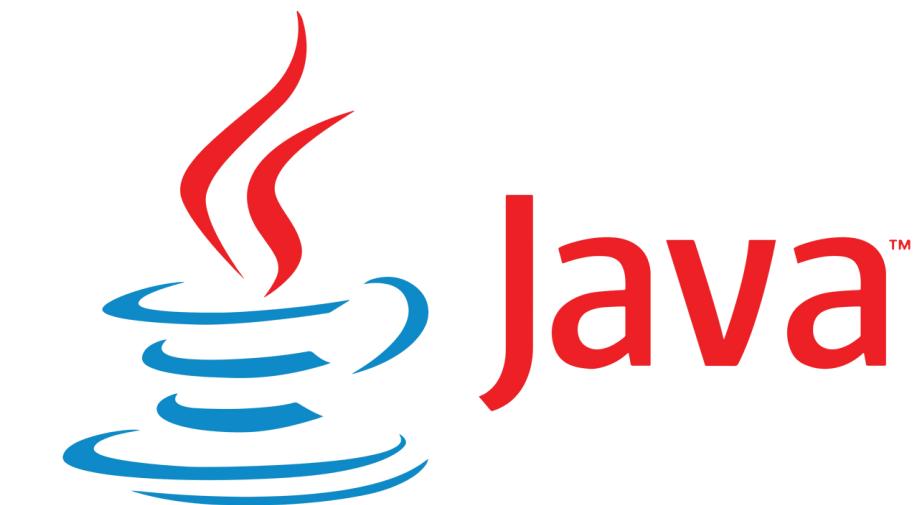
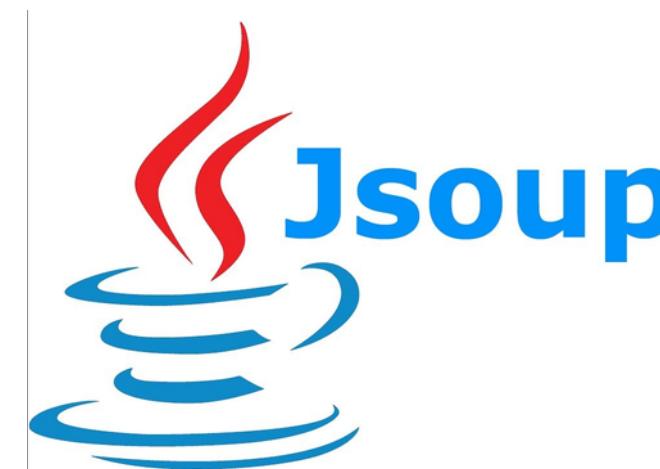
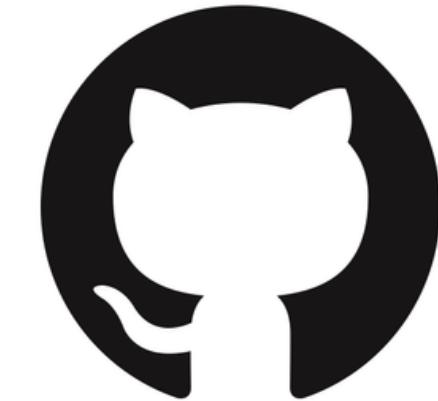
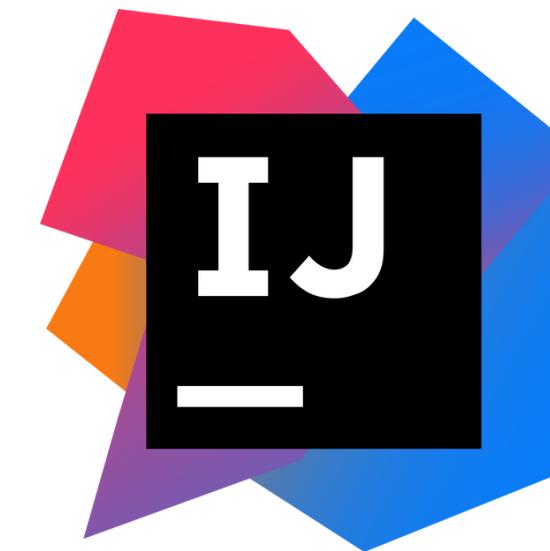
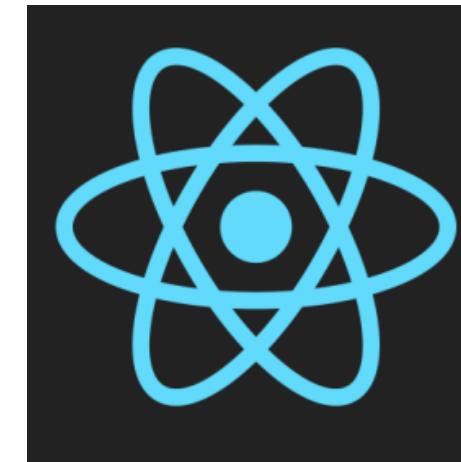
KEY FEATURES

- 
- 
- 01 WEB CRAWLER
 - 02 DATA VALIDATION USING REGEX
 - 03 HTML PARSER
 - 04 INVERTED INDEXING
 - 05 FREQUENCY COUNT
 - 06 PAGE RANKING
 - 07 WORD COMPLETION
 - 08 SPELL CHECKING
 - 09 SEARCH FREQUENCY
 - 10 FINDING PATTERNS USING REGEX

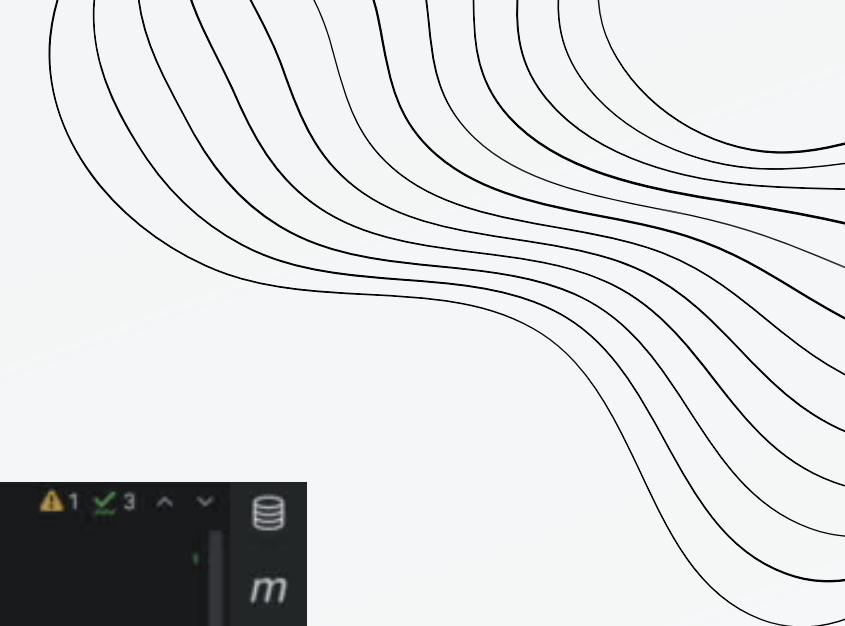
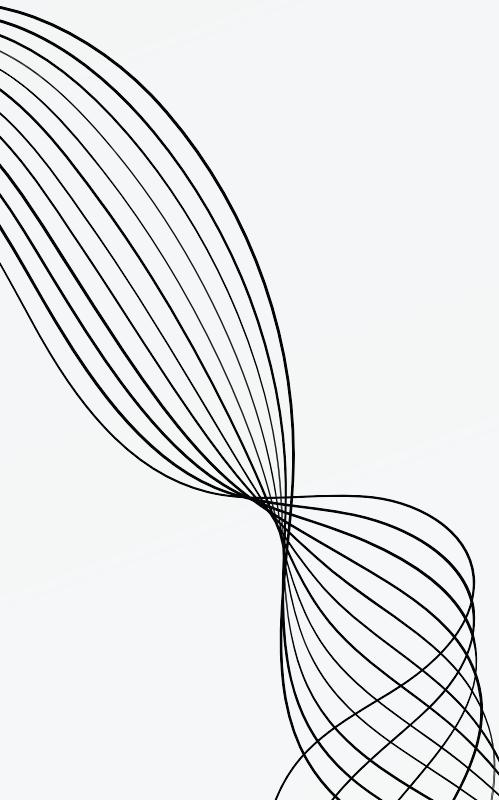
FEATURES OF TEAM MEMBERS

Team Members	Feature 1	Feature 2
Miti Tamakuvala	HTML Parser	Word Completion
Vikram Singh Bedi	Web Crawler	Inverted Index
Meet Patel	Data validation	Finding Pattern using REGEX
Apoorv Shamma	Page Ranking	Frequency Count
Kishan Modi	Spell Checking	Search Frequency

TOOLS AND TECHNOLOGY USED



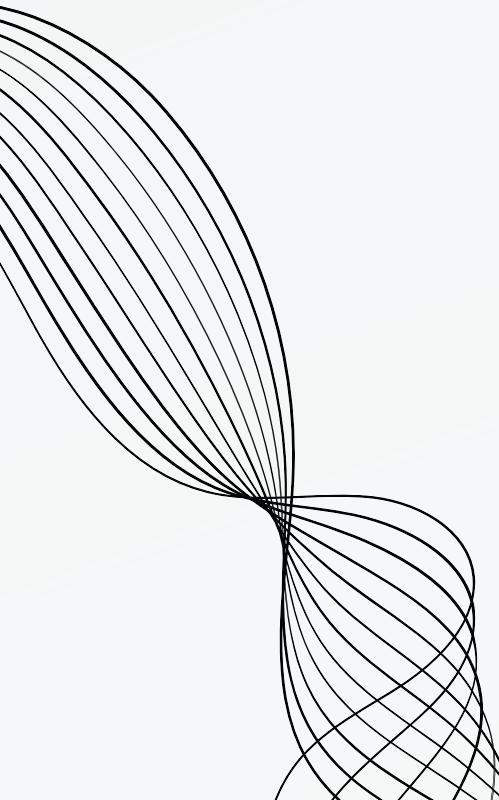
IMPLEMENTATION: WEB CRAWLER



```
1 package com.acc.jobradar.crawler;
2
3 import com.acc.jobradar.filehandler.FileHandler;
4 import lombok.NoArgsConstructor;
5 import org.openqa.selenium.By;
6 import org.openqa.selenium.WebElement;
7 import org.openqa.selenium.chrome.ChromeDriver;
8 import org.springframework.stereotype.Service;
9
10 import java.io.File;
11 import java.time.Duration;
12 import java.util.ArrayList;
13 import java.util.List;
14
15 2 usages ▾ Vikram Singh Bedi *
16 @NoArgsConstructor
17 @Service
18 public class WorkopolisCrawler {
19     1 usage
20     private static final String URL = "https://www.workopolis.com/en";
21     private final ChromeDriver chromeDriver;
22     private final FileHandler fileHandler;
23
24     2 usages ▾ Vikram Singh Bedi *
25     public void getJobPosting(String jobTitle, String location) throws InterruptedException {
26         chromeDriver.get(URL);
27
28         // Get the input box for keyword and location
29         WebElement keywordInput = chromeDriver.findElement(By.id("query-input"));
30         WebElement locationInput = chromeDriver.findElement(By.id("location-input"));
31
32         // Insert job title and location provided by user into the input box
33         keywordInput.sendKeys(jobTitle);
34
35         // Click the search button
36         WebElement searchButton = chromeDriver.findElement(By.className("search-button"));
37         searchButton.click();
38
39         // Wait for the results page to load
40         Duration duration = Duration.ofSeconds(10);
41         WebDriverWait wait = new WebDriverWait(chromeDriver, duration);
42         wait.until(ExpectedConditions.presenceOfElementLocated(By.className("job-list")));
43
44         // Extract job titles and locations from the results page
45         List<JobPosting> jobPostings = extractJobPostings(chromeDriver);
46
47         // Save the job postions to a file
48         fileHandler.saveJobPostings(jobPostings);
49
50     }
51 }
```

Workopolis

IMPLEMENTATION: WEB CRAWLER

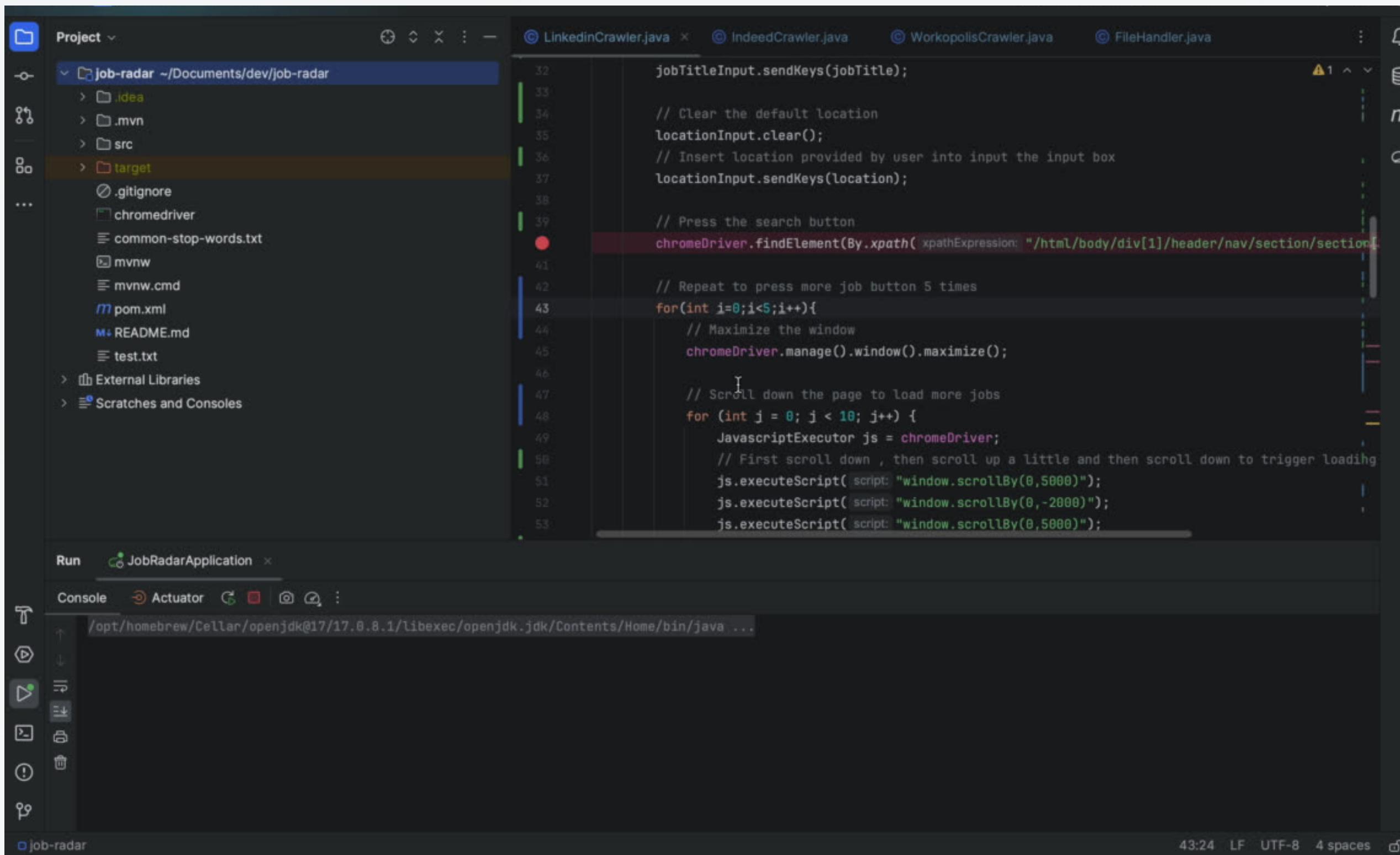


```
J job-radar  main
IndeedCrawler.java

1 package com.acc.jobradar.crawler;
2
3 import com.acc.jobradar.filehandler.FileHandler;
4 import lombok.AllArgsConstructor;
5 import org.openqa.selenium.By;
6 import org.openqa.selenium.NoSuchElementException;
7 import org.openqa.selenium.WebElement;
8 import org.openqa.selenium.chrome.ChromeDriver;
9 import org.springframework.stereotype.Service;
10
11 import java.io.File;
12 import java.time.Duration;
13 import java.util.ArrayList;
14 import java.util.List;
15
16 2 usages  ± Vikram Singh Bedi *
17 @AllArgsConstructor
18 @Service
19
20 public class IndeedCrawler {
21     1 usage
22     private static final String URL = "https://ca.indeed.com/jobs?";
23     private final ChromeDriver chromeDriver;
24     private final FileHandler fileHandler;
25
26     2 usages  ± Vikram Singh Bedi *
27     public void getJobPosting(String jobTitle, String location) throws InterruptedException {
28         // Open Indeed with job title and location provided by the user embedded in the URL
29         chromeDriver.get(URL + "q=" + jobTitle + "&l=" + location);
30
31         // List to store job link
32         List<String> jobLinks = new ArrayList<>();
33
34         // Find all job links on the page
35         List<WebElement> jobListings = chromeDriver.findElements(By.cssSelector("div.job_seen_list"));
36
37         for (WebElement listing : jobListings) {
38             String jobLink = listing.findElement(By.cssSelector("a.job_seen_link")).getAttribute("href");
39             jobLinks.add(jobLink);
40         }
41
42         // Write job links to file
43         fileHandler.writeJobLinks(jobLinks);
44     }
45
46     public void writeJobLinks(List<String> jobLinks) {
47         fileHandler.writeJobLinks(jobLinks);
48     }
49 }
```

Indeed

IMPLEMENTATION: WEB CRAWLER



The screenshot shows an IDE interface with the following components:

- Project View:** Shows the project structure for "job-radar" located at "/Documents/dev/job-radar". It includes folders like ".idea", ".mvn", "src", "target", and files like ".gitignore", "chromedriver", "common-stop-words.txt", "mvnw", "mvnw.cmd", "pom.xml", "README.md", and "test.txt".
- Code Editor:** Displays the content of `LinkedInCrawler.java`. The code implements a crawler for LinkedIn, using Selenium WebDriver to interact with the browser. It handles job title input, location input, and search button presses. It also uses JavaScriptExecutor to scroll down the page to load more jobs.
- Run Tab:** Shows a run configuration for "JobRadarApplication".
- Console Tab:** Displays the output of the application's console, which includes the path to the Java executable: "/opt/homebrew/Cellar/openjdk@17/17.0.8.1/libexec/openjdk.jdk/Contents/Home/bin/java".
- Status Bar:** Shows the time as 43:24, the encoding as LF, the file format as UTF-8, and the code style as 4 spaces.

Linkedin

HTML PARSER

Purpose:

To parse and extract the job data from the crawled LinkedIn, Indeed, and Workopolis HTML files

Data Extracted:

Job Roles, locations, Company Names, Job Descriptions, Application Links

HTML parsing is done using **Jsoup** Java Library.

Used **ArrayLists** to store details of jobs extracted during the iteration through HTML files.



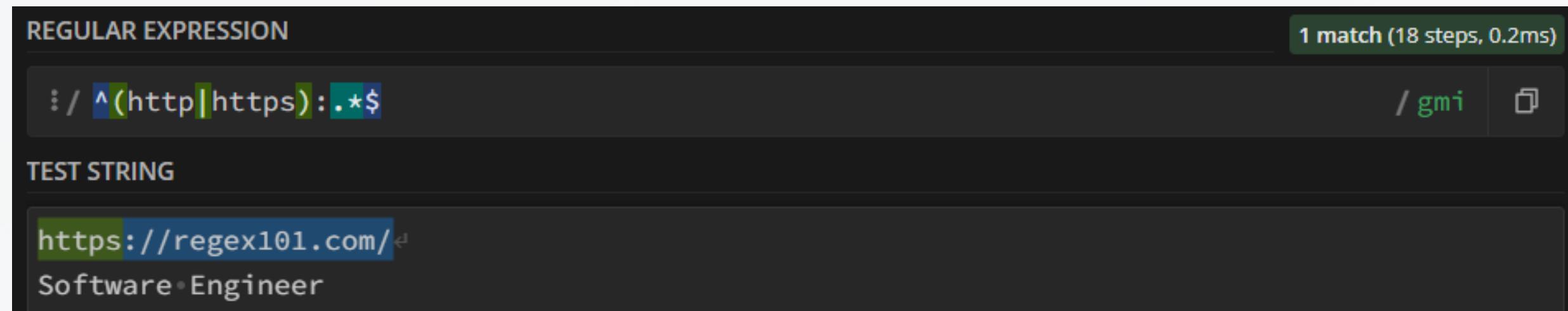
HTML PARSER

```
Run htmlParserIndeed x : -  
↻ | ☕ + ⌂ :  
"C:\Program Files\Java\jdk-20\bin\java.exe" "-javaagent:C:\Program Files\JetBrains\IntelliJ IDEA 2023.2.5\lib\idea_rt.jar=59969:C:\Program Files\JetBrains\IntelliJ IDEA 2023.2.5  
Indeed Job details ==> Job Role:Intelligent Grid Solutions Intern | Company Name:TRC Companies, Inc. | Company Location:Toronto, ON | Job Description:About Us: Groundbreaker. G  
Indeed Job details ==> Job Role:Data & Insights Canada Summer Intern | Company Name:Johnson & Johnson | Company Location:Toronto, ON | Job Description:At Johnson & Johnson, we i  
Indeed Job details ==> Job Role:Senior Back End (Java) Software Engineer | Company Name:Marler Search Group | Company Location:1000 Innovation Dr, Kanata, ON | Job Description:  
Indeed Job details ==> Job Role:Senior Software & PLC Engineer | Company Name:I.B.C. International Bar Coding Systems & Consulting Inc. | Company Location:Penticton, BC | Job De  
Indeed Job details ==> Job Role:Diagnostic Communication Programmer | Company Name:Expleo | Company Location:Valcourt, QC | Job Description:Overview: EXPLEO CANADA est à la recr  
Indeed Job details ==> Job Role:Software Engineer Intern | Company Name:ATPCO | Company Location:700 West Georgia Street, Vancouver, BC | Job Description:Job Description As a S  
Indeed Job details ==> Job Role:Senior Programmer | Company Name:Applied Électronique Limitée || Applied Electronics Limited | Company Location:#5200 Rue Saint-Patrick, Montréal  
Indeed Job details ==> Job Role:CNC Machinist/Programmer | Company Name:Hawk Machine Works Ltd | Company Location:202 6 Street NW, Linden, AB | Job Description:Here at Hawk Mac  
Indeed Job details ==> Job Role:Software Development Engineer | Company Name:Mastercard | Company Location:Vancouver, BC | Job Description:Our Purpose We work to connect and po  
Indeed Job details ==> Job Role:Full-Stack Software Engineer | Company Name:Manulife | Company Location:Waterloo, ON | Job Description:We are a leading financial services provider  
Indeed Job details ==> Job Role:Senior Software Engineer | Company Name:Microsoft | Company Location:Ireland, NS | Job Description:Summary Do you want to be part of a growing g  
Indeed Job details ==> Job Role>New Grad: AI Software Engineer | Company Name:Parallelz | Company Location:Toronto, ON | Job Description:Summary Parallelz enables developers and  
Indeed Job details ==> Job Role:software testing engineer | Company Name:Bloombase Canada Inc. | Company Location:Vancouver, BC | Job Description:Education: Secondary (high) sci  
Indeed Job details ==> Job Role:Junior Developer | Company Name:KPMG | Company Location:Quebec City, QC | Job Description:Overview: At KPMG, you'll join a team of diverse and d  
Indeed Job details ==> Job Role:software developer | Company Name:Border Connect Inc | Company Location:Windsor, ON | Job Description:Work Term: Permanent Work Language: English  
Indeed Job details ==> Job Role:Digital Marketing Intern | Company Name:Fun Time Print | Company Location:Remote | Job Description:Ensure you meet student application criteria :  
Indeed Job details ==> Job Role:CNC Equipment Programmer | Company Name:Systemair Commercial AHU Ltd. | Company Location:10 Rouse Street, Tillsonburg, ON | Job Description:CNC |  
Indeed Job details ==> Job Role:Software Engineer - Full-Stack | Company Name:Microsoft | Company Location:Vancouver, BC | Job Description:Microsoft Teams is the hub for teamwo  
  
Output file - 'Indeed_Files_output.txt' was created!  
Indeed Job Details added to the 'Indeed_Files_output.txt' file successfully!
```

DATA VALIDATION USING REGEX

Purpose: The purpose of the provided Java file (`DataValidator`) is to offer methods for validating `JobPosting` data. The `validateData` method examines essential fields within a `JobPosting` object, checking for default or "not found" values. Additionally, it verifies the validity of the website link associated with the job posting. The `isValidWebsiteLink` method specifically checks if a given string represents a valid website link by matching it against a regular expression pattern. The overall goal is to ensure the integrity of essential data fields and the correctness of the website link in a `JobPosting` object.

Utility Used: Regular Expressions

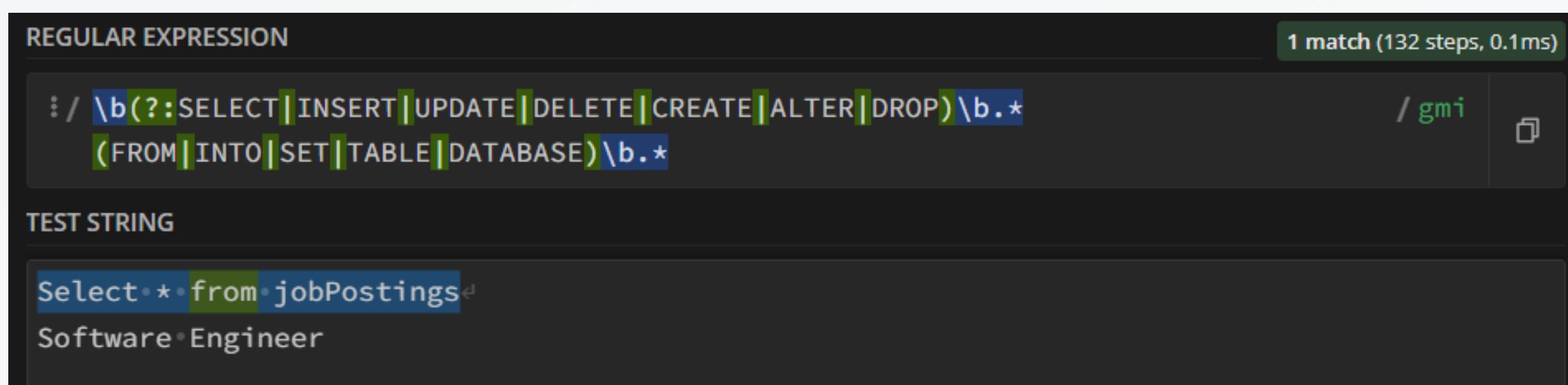


The screenshot shows a dark-themed interface for testing regular expressions. In the top left, under 'REGULAR EXPRESSION', is the pattern `:^(http|https):.*$`. To its right, a green button indicates `1 match (18 steps, 0.2ms)`. Below the expression is a section labeled 'TEST STRING' containing the URL `https://regex101.com/` followed by the text 'Software Engineer'. On the far right of the interface are two small icons: a green checkmark and a clipboard.

FINDING PATTERNS USING REGULAR EXPRESSIONS

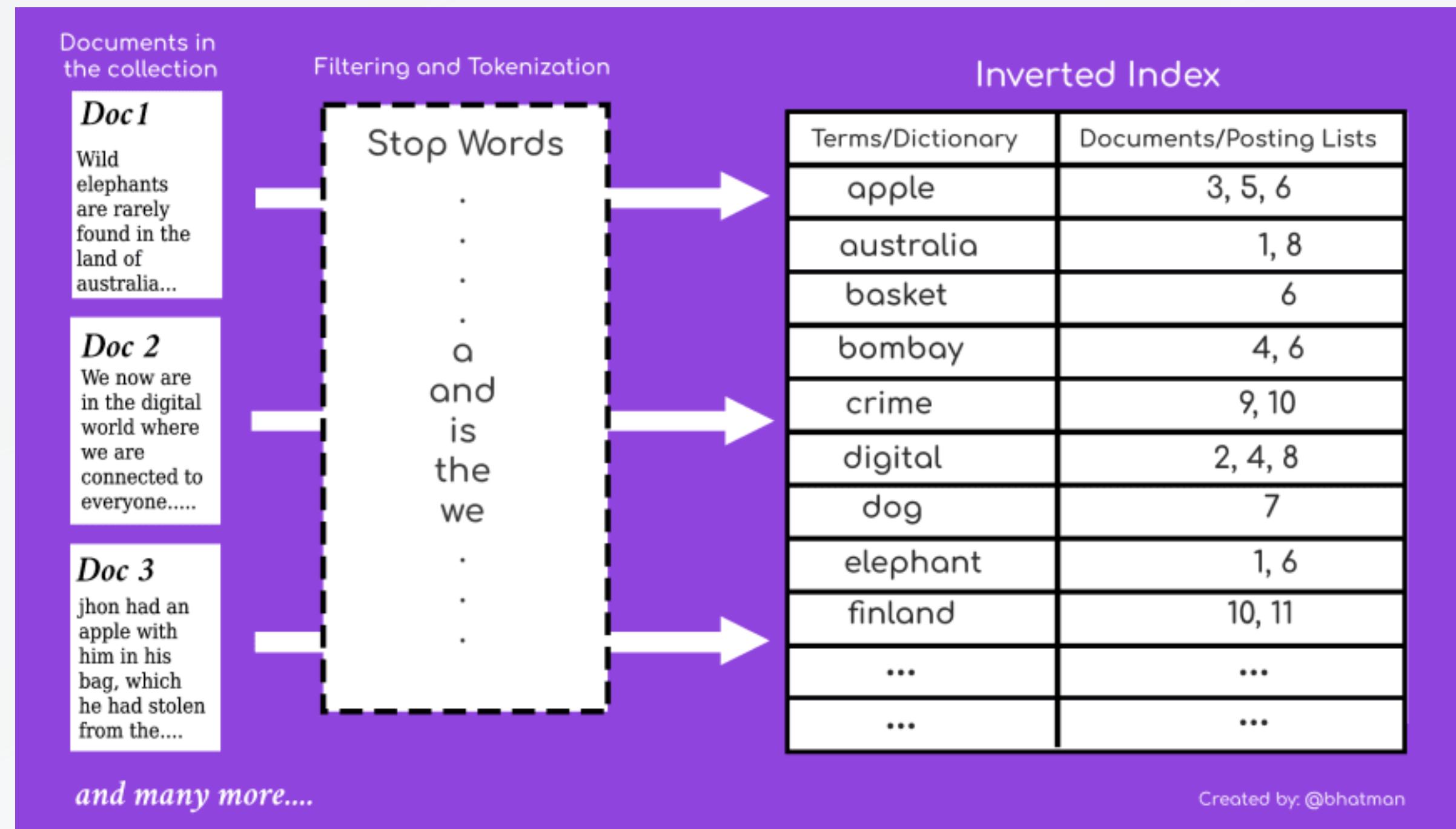
Purpose: The purpose of the provided Java file (SearchQueryValidator) is to offer a set of static methods for validating strings based on specific criteria. The file is primarily focused on data validation for search queries in a programming context. It includes methods for checking whether a string comprises only symbols or numbers (isJustSymbolOrNumbers), contains only alphanumeric characters and spaces (isJustCharactersWithNumbers), and whether it contains SQL keywords and associated clauses (containsSqlKeywords). The overall goal is to ensure that user inputs adhere to specific character sets and to prevent potential security issues related to SQL injection. Regular expressions are employed as the primary data structure for pattern matching in these validation methods. The validateString method combines these individual validations to provide a comprehensive check for a given input string.

Utility: Regular Expressions



The screenshot shows a regular expression tester interface. The top section is labeled "REGULAR EXPRESSION" and contains the following regex pattern:
`:/ \b(?:SELECT|INSERT|UPDATE|DELETE|CREATE|ALTER|DROP)\b.*
(FROM|INTO|SET|TABLE|DATABASE)\b.*`
The bottom section is labeled "TEST STRING" and contains the following input:
`Select * from jobPostings
Software Engineer`
A green status bar at the top right indicates "1 match (132 steps, 0.1ms)".

INVERTED INDEXING



INVERTED INDEXING

Purpose: The inverted index stores words and job postings that contain those words. Given a word, our inverted index will return a list of job postings with those given words.

Data Structures Used: Trie, HashMap, Array

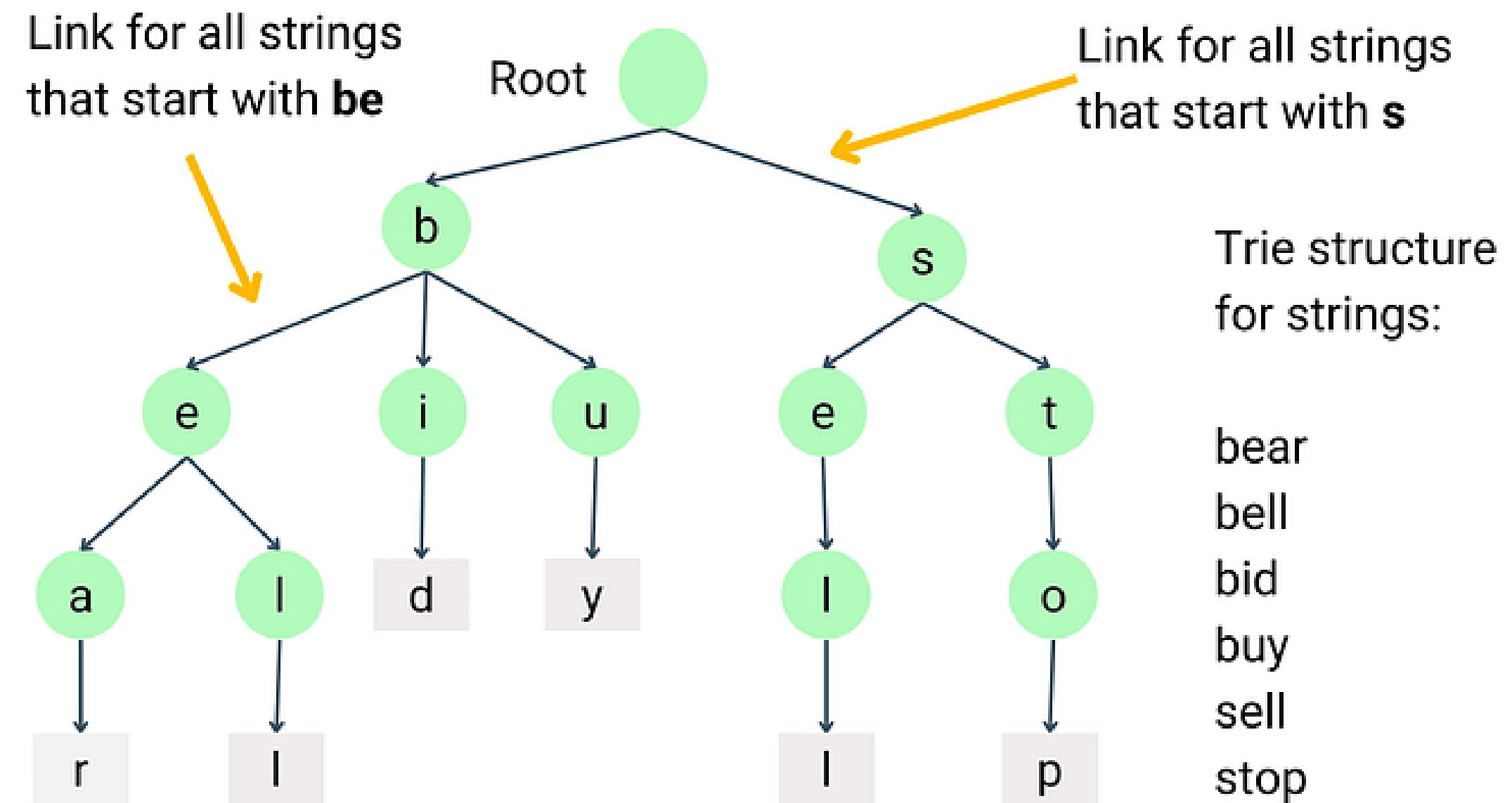
Underlying Data Structure: Trie is used as Tries excel in efficient prefix searches with space-efficient storage

Node Structure: Each node contains a map (children) to represent its child nodes, and a set (jobIds) to store the job IDs associated with the words formed so far.

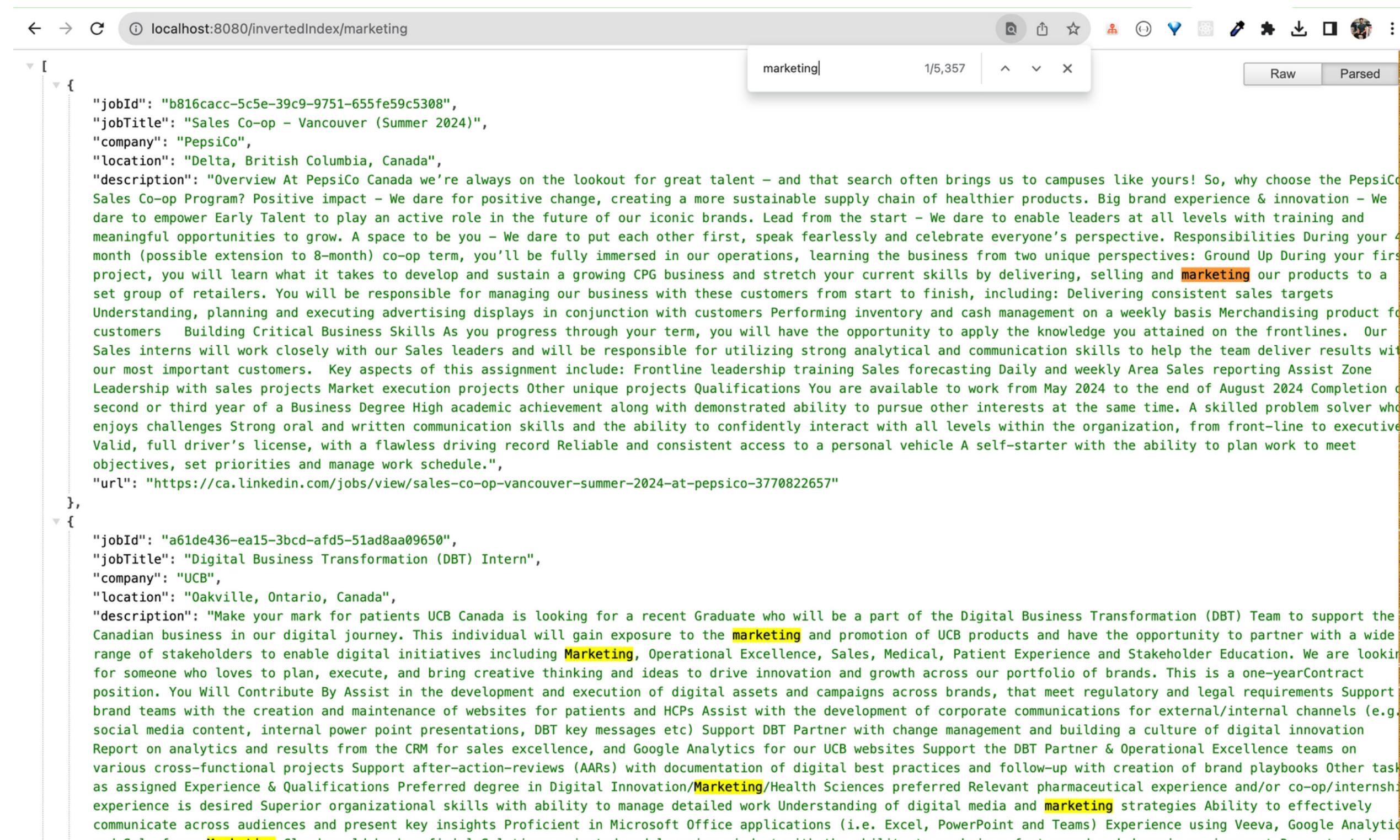
Insertion: Traverse the Trie for each character in the given word. A new TrieNode is created if the character is not already a child of the current node. The traversal continues, and the job ID is added to the jobIds set at each node to support partial matches

Searching: Traverse the trie for each character in the given word, and if the word is found, the associated job IDs are returned. If the word is not found, an empty list is returned.

INVERTED INDEXING



INVERTED INDEXING



localhost:8080/invertedIndex/marketing

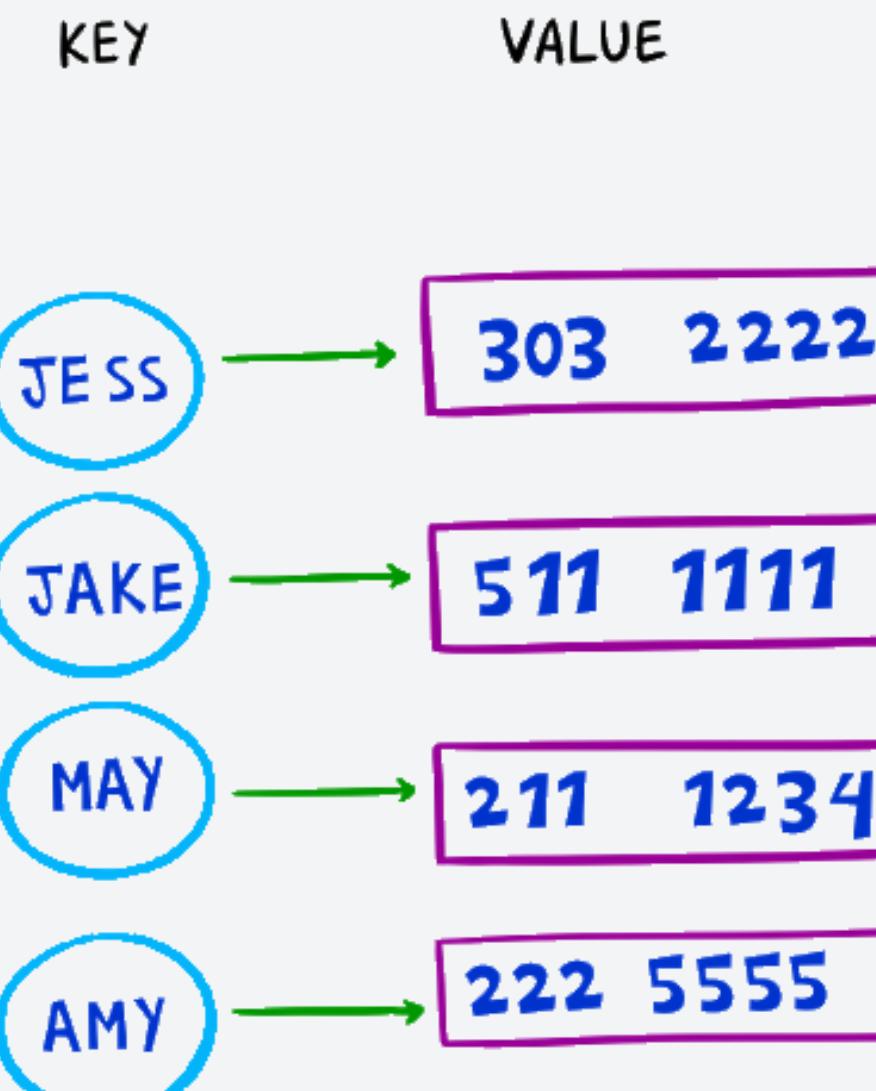
marketing | 1/5,357

Raw Parsed

```
[{"jobId": "b816cacc-5c5e-39c9-9751-655fe59c5308", "jobTitle": "Sales Co-op – Vancouver (Summer 2024)", "company": "PepsiCo", "location": "Delta, British Columbia, Canada", "description": "Overview At PepsiCo Canada we're always on the lookout for great talent – and that search often brings us to campuses like yours! So, why choose the PepsiCo Sales Co-op Program? Positive impact – We dare for positive change, creating a more sustainable supply chain of healthier products. Big brand experience & innovation – We dare to empower Early Talent to play an active role in the future of our iconic brands. Lead from the start – We dare to enable leaders at all levels with training and meaningful opportunities to grow. A space to be you – We dare to put each other first, speak fearlessly and celebrate everyone's perspective. Responsibilities During your 4 month (possible extension to 8-month) co-op term, you'll be fully immersed in our operations, learning the business from two unique perspectives: Ground Up During your first project, you will learn what it takes to develop and sustain a growing CPG business and stretch your current skills by delivering, selling and marketing our products to a set group of retailers. You will be responsible for managing our business with these customers from start to finish, including: Delivering consistent sales targets Understanding, planning and executing advertising displays in conjunction with customers Performing inventory and cash management on a weekly basis Merchandising product for customers Building Critical Business Skills As you progress through your term, you will have the opportunity to apply the knowledge you attained on the frontlines. Our Sales interns will work closely with our Sales leaders and will be responsible for utilizing strong analytical and communication skills to help the team deliver results with our most important customers. Key aspects of this assignment include: Frontline leadership training Sales forecasting Daily and weekly Area Sales reporting Assist Zone Leadership with sales projects Market execution projects Other unique projects Qualifications You are available to work from May 2024 to the end of August 2024 Completion of second or third year of a Business Degree High academic achievement along with demonstrated ability to pursue other interests at the same time. A skilled problem solver who enjoys challenges Strong oral and written communication skills and the ability to confidently interact with all levels within the organization, from front-line to executive Valid, full driver's license, with a flawless driving record Reliable and consistent access to a personal vehicle A self-starter with the ability to plan work to meet objectives, set priorities and manage work schedule.", "url": "https://ca.linkedin.com/jobs/view/sales-co-op-vancouver-summer-2024-at-pepsico-3770822657"}, {"jobId": "a61de436-ea15-3bcd-afd5-51ad8aa09650", "jobTitle": "Digital Business Transformation (DBT) Intern", "company": "UCB", "location": "Oakville, Ontario, Canada", "description": "Make your mark for patients UCB Canada is looking for a recent Graduate who will be a part of the Digital Business Transformation (DBT) Team to support the Canadian business in our digital journey. This individual will gain exposure to the marketing and promotion of UCB products and have the opportunity to partner with a wide range of stakeholders to enable digital initiatives including Marketing, Operational Excellence, Sales, Medical, Patient Experience and Stakeholder Education. We are looking for someone who loves to plan, execute, and bring creative thinking and ideas to drive innovation and growth across our portfolio of brands. This is a one-year contract position. You Will Contribute By Assist in the development and execution of digital assets and campaigns across brands, that meet regulatory and legal requirements Support brand teams with the creation and maintenance of websites for patients and HCPs Assist with the development of corporate communications for external/internal channels (e.g. social media content, internal power point presentations, DBT key messages etc) Support DBT Partner with change management and building a culture of digital innovation Report on analytics and results from the CRM for sales excellence, and Google Analytics for our UCB websites Support the DBT Partner & Operational Excellence teams on various cross-functional projects Support after-action-reviews (AARs) with documentation of digital best practices and follow-up with creation of brand playbooks Other tasks as assigned Experience & Qualifications Preferred degree in Digital Innovation/Marketing/Health Sciences preferred Relevant pharmaceutical experience and/or co-op/internship experience is desired Superior organizational skills with ability to manage detailed work Understanding of digital media and marketing strategies Ability to effectively communicate across audiences and present key insights Proficient in Microsoft Office applications (i.e. Excel, PowerPoint and Teams) Experience using Veeva, Google Analytics and Salesforce Marketing Cloud would be beneficial Solutions oriented and learning mindset with the ability to work in a fast-paced and dynamic environment Demonstrated"}, {"jobId": "a61de436-ea15-3bcd-afd5-51ad8aa09650", "jobTitle": "Digital Business Transformation (DBT) Intern", "company": "UCB", "location": "Oakville, Ontario, Canada", "description": "Make your mark for patients UCB Canada is looking for a recent Graduate who will be a part of the Digital Business Transformation (DBT) Team to support the Canadian business in our digital journey. This individual will gain exposure to the marketing and promotion of UCB products and have the opportunity to partner with a wide range of stakeholders to enable digital initiatives including Marketing, Operational Excellence, Sales, Medical, Patient Experience and Stakeholder Education. We are looking for someone who loves to plan, execute, and bring creative thinking and ideas to drive innovation and growth across our portfolio of brands. This is a one-year contract position. You Will Contribute By Assist in the development and execution of digital assets and campaigns across brands, that meet regulatory and legal requirements Support brand teams with the creation and maintenance of websites for patients and HCPs Assist with the development of corporate communications for external/internal channels (e.g. social media content, internal power point presentations, DBT key messages etc) Support DBT Partner with change management and building a culture of digital innovation Report on analytics and results from the CRM for sales excellence, and Google Analytics for our UCB websites Support the DBT Partner & Operational Excellence teams on various cross-functional projects Support after-action-reviews (AARs) with documentation of digital best practices and follow-up with creation of brand playbooks Other tasks as assigned Experience & Qualifications Preferred degree in Digital Innovation/Marketing/Health Sciences preferred Relevant pharmaceutical experience and/or co-op/internship experience is desired Superior organizational skills with ability to manage detailed work Understanding of digital media and marketing strategies Ability to effectively communicate across audiences and present key insights Proficient in Microsoft Office applications (i.e. Excel, PowerPoint and Teams) Experience using Veeva, Google Analytics and Salesforce Marketing Cloud would be beneficial Solutions oriented and learning mindset with the ability to work in a fast-paced and dynamic environment Demonstrated"}, {"jobId": "a61de436-ea15-3bcd-afd5-51ad8aa09650", "jobTitle": "Digital Business Transformation (DBT) Intern", "company": "UCB", "location": "Oakville, Ontario, Canada", "description": "Make your mark for patients UCB Canada is looking for a recent Graduate who will be a part of the Digital Business Transformation (DBT) Team to support the Canadian business in our digital journey. This individual will gain exposure to the marketing and promotion of UCB products and have the opportunity to partner with a wide range of stakeholders to enable digital initiatives including Marketing, Operational Excellence, Sales, Medical, Patient Experience and Stakeholder Education. We are looking for someone who loves to plan, execute, and bring creative thinking and ideas to drive innovation and growth across our portfolio of brands. This is a one-year contract position. You Will Contribute By Assist in the development and execution of digital assets and campaigns across brands, that meet regulatory and legal requirements Support brand teams with the creation and maintenance of websites for patients and HCPs Assist with the development of corporate communications for external/internal channels (e.g. social media content, internal power point presentations, DBT key messages etc) Support DBT Partner with change management and building a culture of digital innovation Report on analytics and results from the CRM for sales excellence, and Google Analytics for our UCB websites Support the DBT Partner & Operational Excellence teams on various cross-functional projects Support after-action-reviews (AARs) with documentation of digital best practices and follow-up with creation of brand playbooks Other tasks as assigned Experience & Qualifications Preferred degree in Digital Innovation/Marketing/Health Sciences preferred Relevant pharmaceutical experience and/or co-op/internship experience is desired Superior organizational skills with ability to manage detailed work Understanding of digital media and marketing strategies Ability to effectively communicate across audiences and present key insights Proficient in Microsoft Office applications (i.e. Excel, PowerPoint and Teams) Experience using Veeva, Google Analytics and Salesforce Marketing Cloud would be beneficial Solutions oriented and learning mindset with the ability to work in a fast-paced and dynamic environment Demonstrated"}]
```

FREQUENCY COUNT

HASHMAP



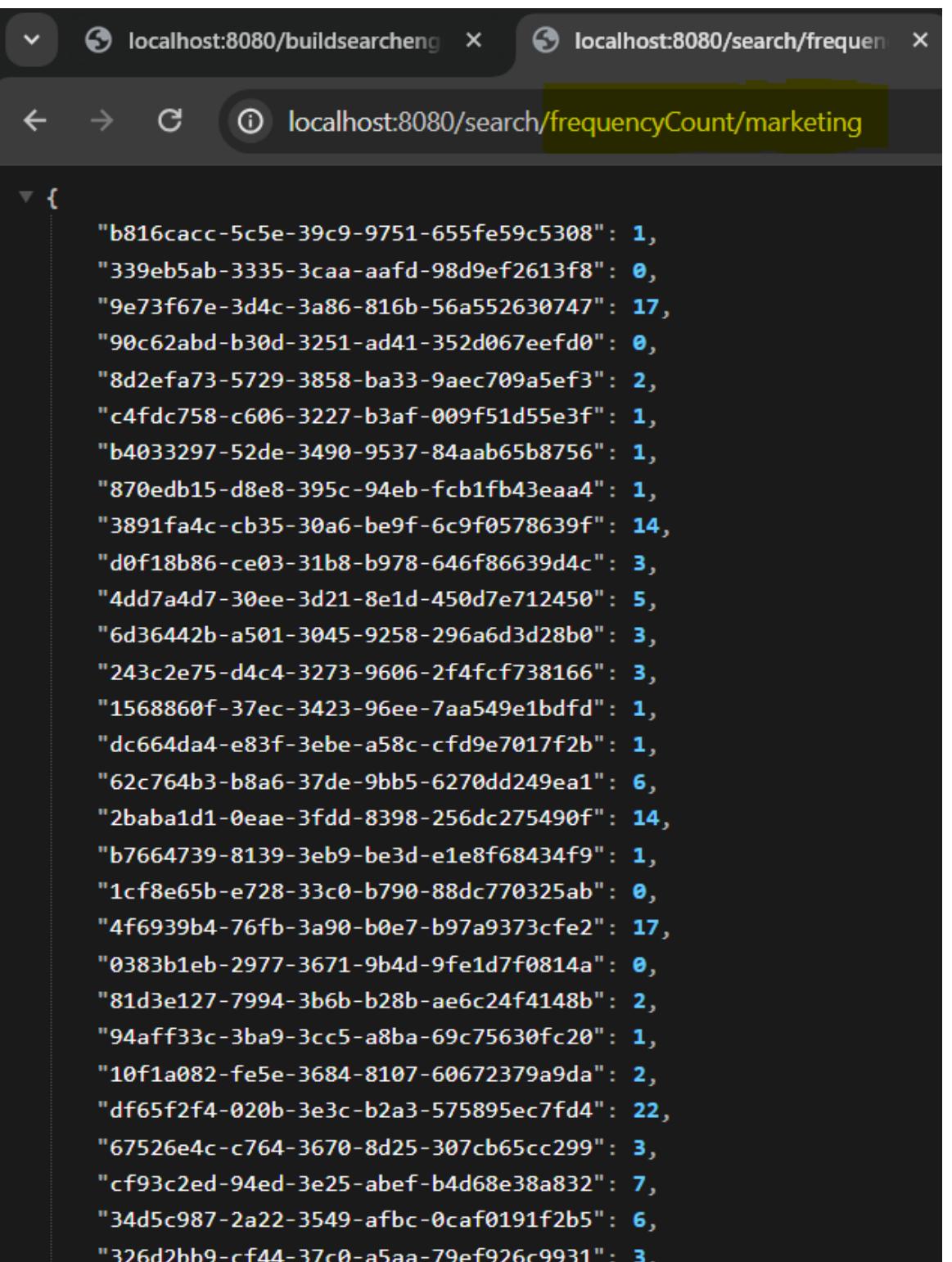
Purpose: The main purpose of this feature is to show the user the no. of times a word occurred in a given URL. For us, it will be a Job posting.

Data Structures Used: HashMap, ArrayList

Functionalities:

- The function will analyze all the Job attributes such as:
 - Job Title
 - Company name
 - Job description
 - Job location.
- All the manipulation will be case insensitive.

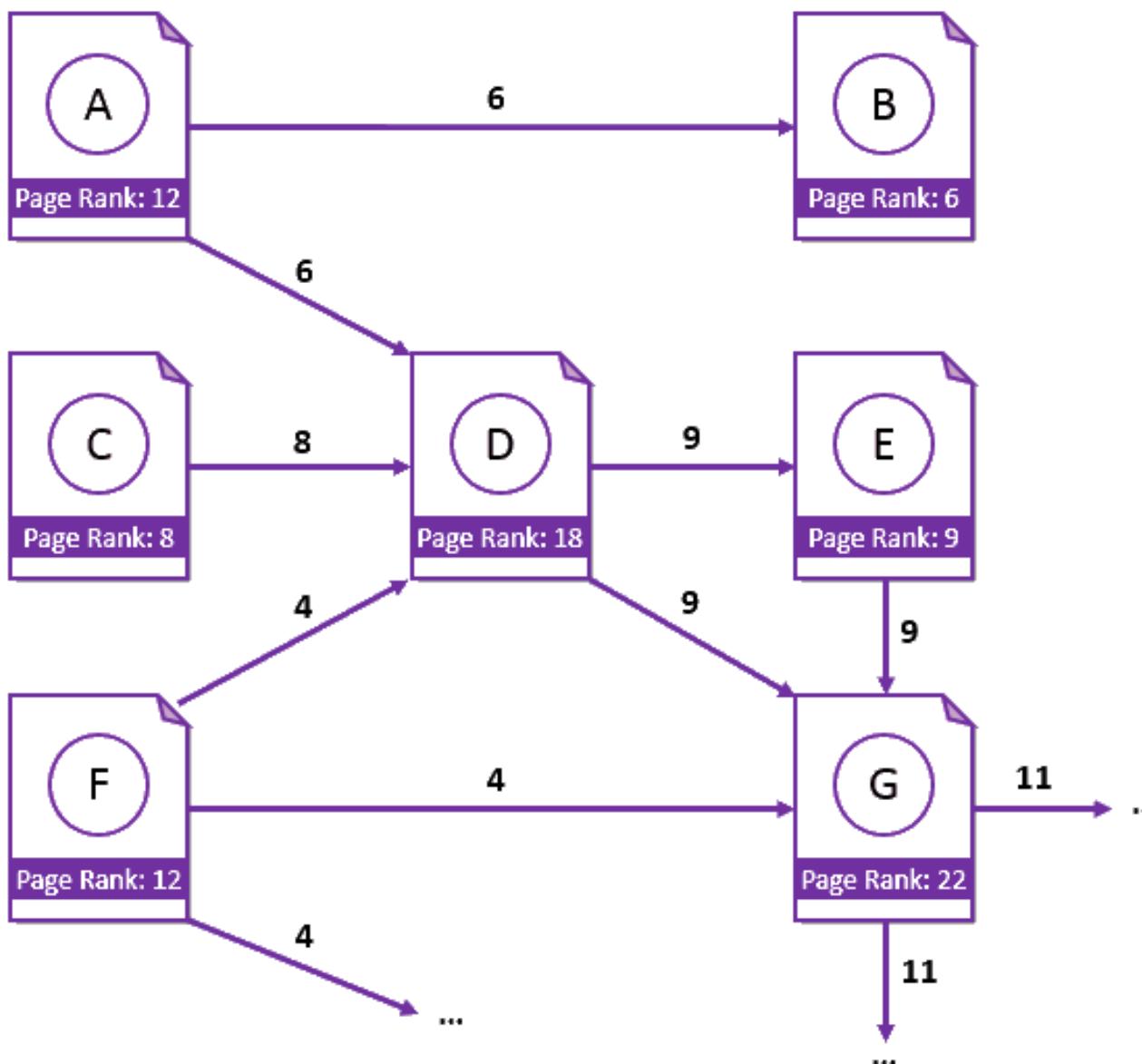
FREQUENCY COUNT



A screenshot of a web browser window displaying a JSON object representing a frequency count. The URL in the address bar is `localhost:8080/search/frequencyCount/marketing`. The JSON data shows various UUIDs as keys and their corresponding frequency counts as values.

```
{  
    "b816cacc-5c5e-39c9-9751-655fe59c5308": 1,  
    "339eb5ab-3335-3caa-aaf9-98d9ef2613f8": 0,  
    "9e73f67e-3d4c-3a86-816b-56a552630747": 17,  
    "90c62abd-b30d-3251-ad41-352d067eef0": 0,  
    "8d2efa73-5729-3858-ba33-9aec709a5ef3": 2,  
    "c4fdc758-c606-3227-b3af-009f51d55e3f": 1,  
    "b4033297-52de-3490-9537-84aab65b8756": 1,  
    "870edb15-d8e8-395c-94eb-fcb1fb43eaa4": 1,  
    "3891fa4c-cb35-30a6-be9f-6c9f0578639f": 14,  
    "d0f18b86-ce03-31b8-b978-646f86639d4c": 3,  
    "4dd7a4d7-30ee-3d21-8e1d-450d7e712450": 5,  
    "6d36442b-a501-3045-9258-296a6d3d28b0": 3,  
    "243c2e75-d4c4-3273-9606-2f4fcf738166": 3,  
    "1568860f-37ec-3423-96ee-7aa549e1bdfd": 1,  
    "dc664da4-e83f-3ebe-a58c-cfd9e7017f2b": 1,  
    "62c764b3-b8a6-37de-9bb5-6270dd249ea1": 6,  
    "2babab1d1-0eae-3fdd-8398-256dc275490f": 14,  
    "b7664739-8139-3eb9-be3d-e1e8f68434f9": 1,  
    "1cf8e65b-e728-33c0-b790-88dc770325ab": 0,  
    "4f6939b4-76fb-3a90-b0e7-b97a9373cfe2": 17,  
    "0383b1eb-2977-3671-9b4d-9fe1d7f0814a": 0,  
    "81d3e127-7994-3b6b-b28b-ae6c24f4148b": 2,  
    "94aff33c-3ba9-3cc5-a8ba-69c75630fc20": 1,  
    "10f1a082-fe5e-3684-8107-60672379a9da": 2,  
    "df65f2f4-020b-3e3c-b2a3-575895ec7fd4": 22,  
    "67526e4c-c764-3670-8d25-307cb65cc299": 3,  
    "cf93c2ed-94ed-3e25-abef-b4d68e38a832": 7,  
    "34d5c987-2a22-3549-afbc-0caf0191f2b5": 6,  
    "326d2bb9-cf44-37c0-a5aa-79ef926c9931": 3  
}
```

PAGE RANKING



Purpose: The main purpose of this feature is to rank web pages based on user search strings.

Data Structures Used: PriorityQueue (Max Heap), HashMap

Functionalities:

- Rank Pages based on the relevance of User search input.
- Pages sorted by search term frequency.
- Priority Queue is used to maintain a descending order of sorting.
- Page Ranking is unweighted.

PAGE RANKING

```
localhost:8080/search/pageRanking/marketing
"df65f2f4-020b-3e3c-b2a3-575895ec7fd4",
"f4cd5cf0-69d3-3523-8e83-c77b955aef19",
"4f693b4-76fb-3a90-b0e7-b97a9373cfe2",
"9e73f67e-3d4c-3a86-816b-56a552630747",
"2babad1-0eae-3fdd-8398-256dc275490f",
"3891fa4c-cb35-30a6-be9f-6c9f0578639f",
"cb4feec2-a8b2-3470-b93c-ac184fd36e06",
"f95cf179-701b-3be3-a483-3dd783ba1196",
"8654e7f7-3040-3a33-a7bb-1d2d055e0c6a",
"365e1746-36cb-3287-9d0f-fa30201748b0",
"98280fd-6e33-367c-bd08-92d7bc56264a",
"132fe67f-fb68-3716-8af2-71326f7922b4",
"87b76aec-5407-32cf-89f4-f8302f311758",
"78d15914-def5-31cd-b2cc-98de49e5f344",
"9442a99a-b5f5-3e43-b317-730632b1be89",
"360533d1-69e2-37b2-a6e5-fa85df36bd0f",
"d0d1278f-5d61-37cf-819f-5f2e2fe4396f",
"5e477040-23b7-3ef6-874a-49828d480dd3",
"b7628c46-5d8b-3cfa-812c-00c0cf49b483",
"cf93c2ed-94ed-3e25-abef-b4d68e38a832",
"b265d358-c6be-371c-8a3e-5b1db319ac7d",
"0bdcc806-ce61-3048-978c-05caf5c50197",
"62c764b3-b8a6-37de-9bb5-6270dd249ea1",
"e3d6e479-956d-39ed-aa70-574d1f96748f",
"34d5c987-2a22-3549-afbc-0caf0191f2b5",
"e2d4c68e-e030-3e04-9c04-d749926169a6"
```

Job Radar

Marketing

Marketing Co-Op

Location: Vancouver, BC (Remote) What's it like here? Copperleaf's software helps some of the world's largest energy firms make better strategic decisions. Join our team as we continue to enable our clients to build more resilient and sustainable infrastructure! We have a track record of delivering award-winning,...

Marketing Co-Op

Location: Vancouver, BC (Remote) What's it like here? Copperleaf's software helps some of the world's largest energy firms make better strategic decisions. Join our team as we continue to enable our clients to build more resilient and sustainable infrastructure! We have a track record of delivering award-winning,...

Marketing Internship

Creating Peace of Mind by Pioneering Safety and Security At Allegion, we help keep the people you know and love safe and secure where they live, work and visit. With more than 30 brands, 11,000+ employees globally and products sold in 130 countries, we specialize in security around the doorway and beyond....

Marketing Intern

We are seeking a highly motivated and passionate Marketing Intern to join our team. As a Marketing Intern, you will collaborate with the marketing team to...

```
localhost:8080/jobPosting/df65f2f4-020b-3e3c-b2a3-575895ec7fd4
{
  "jobId": "df65f2f4-020b-3e3c-b2a3-575895ec7fd4",
  "jobTitle": "Marketing Co-Op",
  "company": "Copperleaf Technologies Inc.",
  "location": "2920 Virtual Way, Suite 140, Vancouver, BC",
  "description": "Location: Vancouver, BC (Remote) What's it like here? Copperleaf's software helps some of the world's largest energy firms make better strategic decisions. Join our team as we continue to enable our clients to build more resilient and sustainable infrastructure! We have a track record of delivering award-winning,..."}
```

WORD COMPLETION

Trie Data Structure:

A Trie is employed to efficiently store and manage words extracted from job postings.

Inserting words in trie:

Each node in the Trie represents a single character of the input word.

Word Suggestions:

Suggestions based on user input prefixes. (also supports multi-word input)

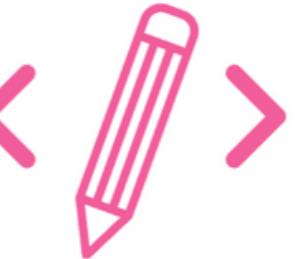
Matching prefix:

Trie traversal, returning words matching the input prefix with a specified maximum edit distance.

Input Handling and Lowercasing:

All characters are converted to lowercase during Trie operations, ensuring case-insensitive matching.

WORD COMPLETION

<  > Job Radar

Software Devel

software develop

software develops

software developer

software developers



SPELL CHECKING

Purpose: To check and provide the correct spelling of search query entered by the User.

Data Structures Used: LinkedHashSet, HashMap, ArrayList

Algorithm Used: Edit Distance, Jaccard Similarity

Edit Distance: Determines the minimum number of changes needed to transform string 1 into string 2. The vocabulary, constructed from parsed data, is employed to assess the edit distance between the entered word and words in the vocabulary. A threshold of 3 is considered as the maximum allowable edit distance between two words.

Jaccard Similarity: Computes the similarity score between words in the vocabulary and a misspelled word using the Jaccard index. The Jaccard Index is calculated as the length of the intersection of characters of two words divided by the union of characters of the same two words. It produces a score between 0 and 1, where 1 indicates high similarity, and 0 indicates low similarity.

Normalizing the Score: $\text{MergedScore} = \text{EditDistanceScore} + (10 * (1 - \text{JaccardSimilarityScore}))$

SPELL CHECKING

Softwxare engineinrg

Did you mean "[software engineering](#)" ? Search instead for [Softwxare engineinrg](#)".

Software Development Intern

SEARCH FREQUENCY

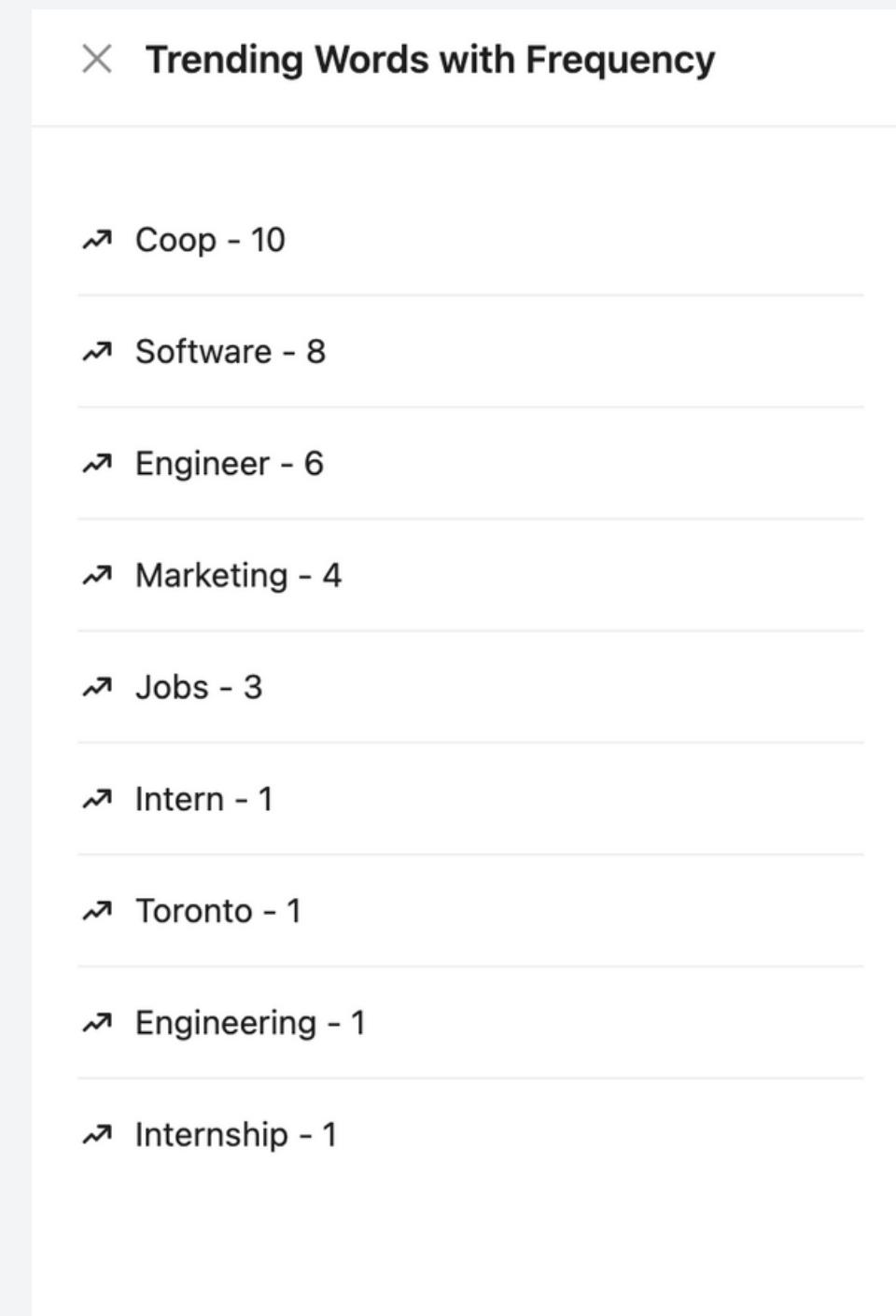
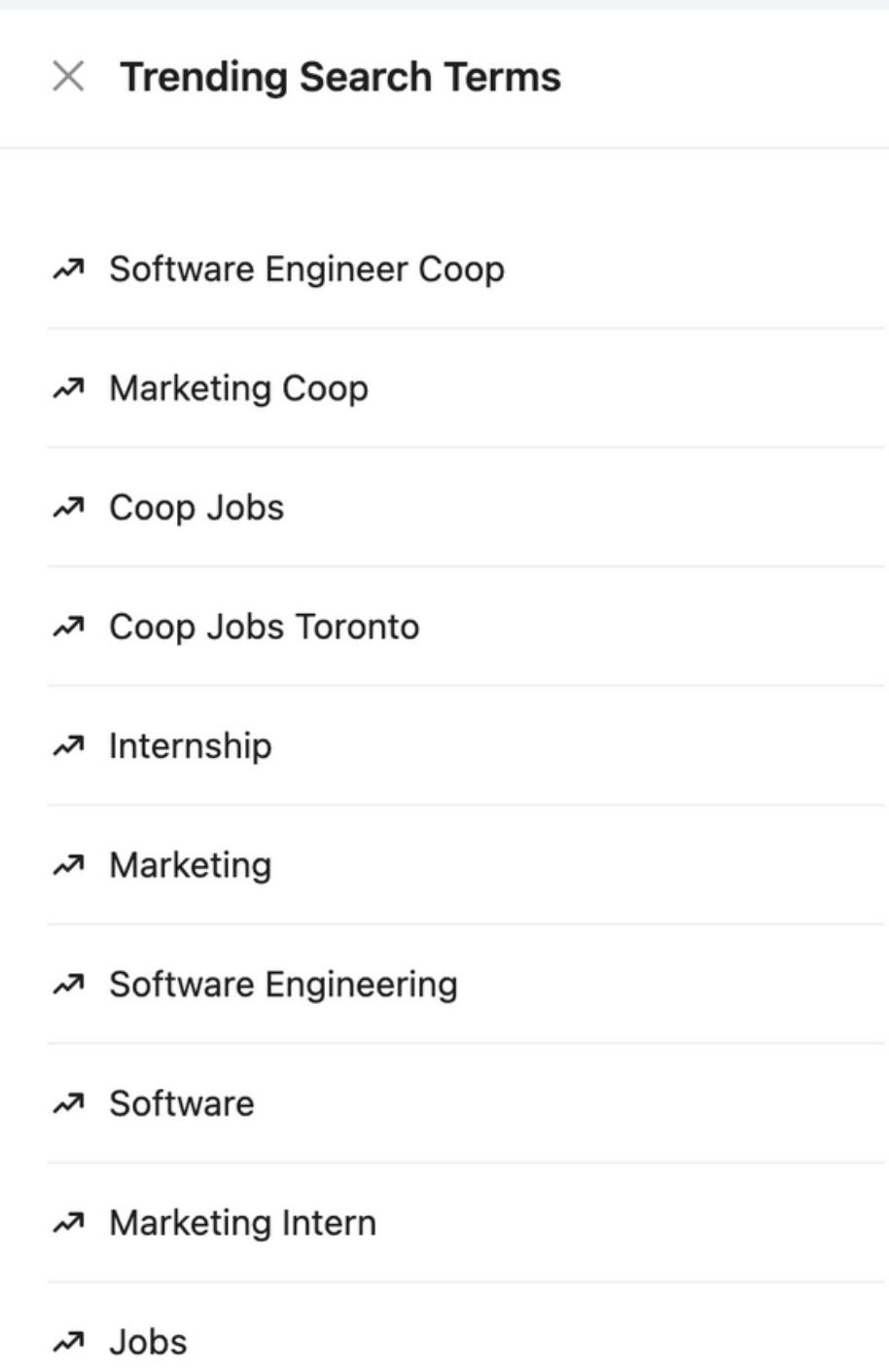
Purpose: To store the search frequency of search terms as well as the words entered by user alongside the frequency of the words.

Data Structures Used: Priority Queue - Max Heap, ArrayList, HashMap

Word Frequency: The engine processes search terms by splitting them and applies preprocessing before storing the words in a HashMap. Each word is associated with its frequency, and the frequency is incremented if the word has been searched previously. This information is intended for retrieval and display on the front-end, particularly on the homepage.

Trending Search Term: The top trending search terms are maintained in a Priority Queue implemented as a Max Heap. The Max Heap ensures efficient retrieval of the most frequently searched terms. The system keeps track of the top 10 trending search terms, which can be retrieved for presentation on the front-end, offering users insights into the most popular search terms.

SEARCH FREQUENCY



CRON JOB

Purpose: The purpose of Java file (`MyScheduledTasks`) uses Spring's `@Scheduled` annotation to define a scheduled task named `crawlTask`. This task, scheduled to run at midnight every day, sends an HTTP GET request to a specified endpoint (`crawlEndpointUrl`) using Spring's `RestTemplate`. The purpose is to automate a daily crawling operation, such as web scraping or data synchronization. The class is annotated with `@Component` for Spring component registration.

Feature: Cron Job

Requirement: The objective of this feature is to automate the daily extraction of job posting data from three major job search websites—Indeed, LinkedIn, and Workopolis.

REFERENCES

J. HEDLEY, “JSOUP: JAVA HTML PARSER, BUILT FOR HTML EDITING, CLEANING, SCRAPING, AND XSS SAFETY,” JSOUP: JAVA HTML PARSER. [HTTPS://JSOUP.ORG/](https://jsoup.org/)

GEEKSFORGEEKS, “INTRODUCTION TO TRIE DATA STRUCTURE AND ALGORITHM TUTORIALS,” GEEKSFORGEEKS, MAY 08, 2023. [HTTPS://WWW.GEEKSFORGEEKS.ORG/INTRODUCTION-TO-TRIE-DATA-STRUCTURE-AND-ALGORITHM-TUTORIALS/](https://www.geeksforgeeks.org/introduction-to-trie-data-structure-and-algorithm-tutorials/)

GEEKSFORGEEKS, “AUTO COMPLETE FEATURE USING TRIE,” GEEKSFORGEEKS, FEB. 20, 2023. [HTTPS://WWW.GEEKSFORGEEKS.ORG/AUTO-COMPLETE-FEATURE-USING-TRIE/](https://www.geeksforgeeks.org/auto-complete-feature-using-trie/)

“JAVA HASHMAP.” [HTTPS://WWW.W3SCHOOLS.COM/JAVA/JAVA_HASHMAP.ASP](https://www.w3schools.com/java/java_hashmap.asp)

WIKIPEDIA CONTRIBUTORS, “EDIT DISTANCE,” WIKIPEDIA, SEP. 12, 2023. [HTTPS://EN.WIKIPEDIA.ORG/WIKI/EDIT_DISTANCE](https://en.wikipedia.org/wiki/Edit_distance)

WIKIPEDIA CONTRIBUTORS, “JACCARD INDEX,” WIKIPEDIA, DEC. 03, 2023. [HTTPS://EN.WIKIPEDIA.ORG/WIKI/JACCARD_INDEX](https://en.wikipedia.org/wiki/Jaccard_index)

BAELDUNG, & BAELDUNG. (2023, MAY 6). A GUIDE TO JAVA REGULAR EXPRESSIONS API | BAELDUNG. BAELDUNG. [HTTPS://WWW.BAELDUNG.COM/REGULAR-EXPRESSIONS-JAVA](https://www.baeldung.com/regular-expressions-java)

SCHEDULING TASKS. (N.D.). GETTING STARTED | SCHEDULING TASKS. [HTTPS://SPRING.IO/GUIDES/GS/SCHEDULING-TASKS/](https://spring.io/guides/gs/scheduling-tasks/)

W. XING AND A. GHORBANI, “WEIGHTED PAGERANK ALGORITHM,” PROCEEDINGS. SECOND ANNUAL CONFERENCE ON COMMUNICATION NETWORKS AND SERVICES RESEARCH, 2004., 2004. DOI:10.1109/DNSR.2004.1344743

THANK YOU

Group - 6

Section - 3

COMP 8547 - Advanced Computing

Concepts

Instructor: Dr. Olena Syrotkina

