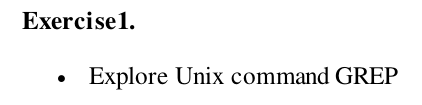
Experiment 1

**Information Retrieval**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Name:** | **:** | **Abhishek N N** | **Register No** | **:** | **20BCE1025** |
| Faculty | : | Alok Chauhan | Slot | : | L51-L52 AB1-605B |
| Course | : | Web Mining Lab | Code | : | CSE3024 |
| Programme | : | B.Tech CSE Core | Semester | : | Win – 22 - 23 |





**grep command in Unix/Linux**

The grep filter searches a file for a particular pattern of characters, and displays all lines that contain that pattern. The pattern that is searched in the file is referred to as the regular expression (grep stands for global search for regular expression and print out).

**Syntax:**

grep [options] pattern [files]

**Options Description:**

-c : This prints only a count of the lines that match a pattern

-h : Display the matched lines, but do not display the filenames.

-i : Ignores, case for matching

-l : Displays list of a filenames only.

-n : Display the matched lines and their line numbers.

-v : This prints out all the lines that do not matches the pattern

-e exp : Specifies expression with this option. Can use multiple times.

-f file : Takes patterns from file, one per line.

-E : Treats pattern as an extended regular expression (ERE)

-w : Match whole word

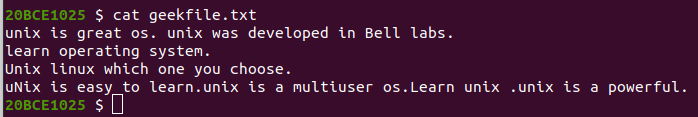
-o : Print only the matched parts of a matching line, with each such part on a separate output line.

-A n : Prints searched line and n lines after the result.

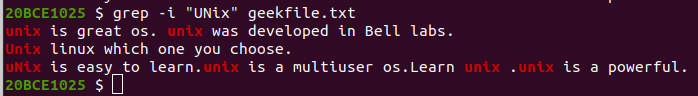
-B n : Prints searched line and n line before the result.

-C n : Prints searched line and n lines after before the result.

Consider the below file as an input.



1. **Case insensitive search:** The -i option enables to search for a string case insensitively in the given file. It matches the words like “UNIX”, “Unix”, “unix”.



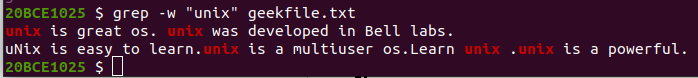
1. **Displaying the count of number of matches:** We can find the number of lines that matches the given string/pattern



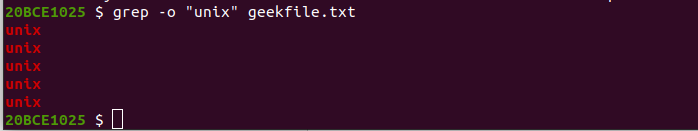
1. **Display the file names that matches the pattern:** We can just display the files that contains the given string/pattern.



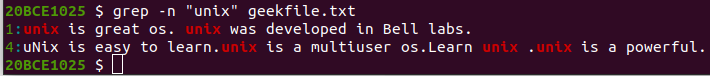
1. **Checking for the whole words in a file:** By default, grep matches the given string/pattern even if it is found as a substring in a file. The -w option to grep makes it match only the whole words.



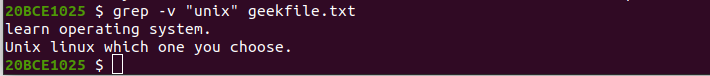
1. **Displaying only the matched patter :** By default, grep displays the entire line which has the matched string. We can make the grep to display only the matched string by using the -o option.



1. **Show line number while displaying the output using grep -n:** To show the line number of file with the line matched.

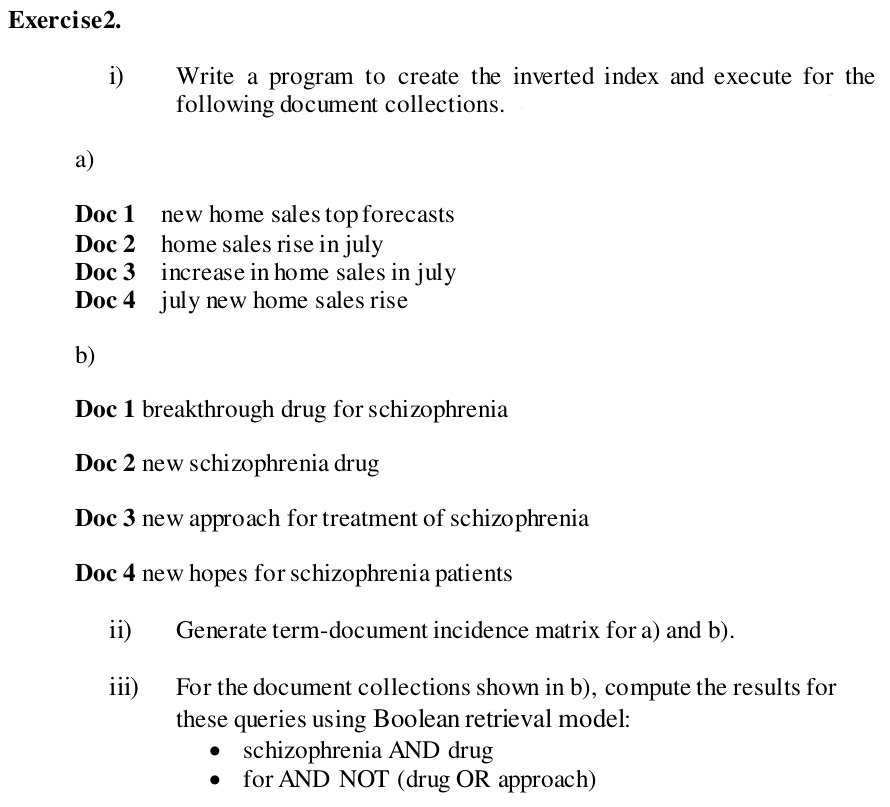


1. **Inverting the pattern match:** You can display the lines that are not matched with the specified search string pattern using the -v option.



1. **Matching the lines that start with a string:** The ^ regular expression pattern specifies the start of a line. This can be used in grep to match the lines which start with the given string or pattern.





**Inverted Indexing Algorithm:**

1. Map Data Structure is used for Indexing
   1. Initialize a map with key as string and value as list (doc.freq is list length)
2. For each document in given directory give one doc id and repeat step 3) to 5)
3. Split the doc into strings and for For each string repeat step 4) to 5)
4. If string is not present in map assign doc id as a list
5. Else append doc id to existing list
6. Save dictionary as file for future searching

**Time Complexity: O( n log(n) )**

Length of the total string is n and log(n) of time is required to insert 1 string

**Space Complexity: O( n^2 )**

O( n ) for term and doc.freq in map and for posting list O( n^2 )

**Term-document incidence matrix generation Algorithm:**

**Boolean retrieval Algorithm:**