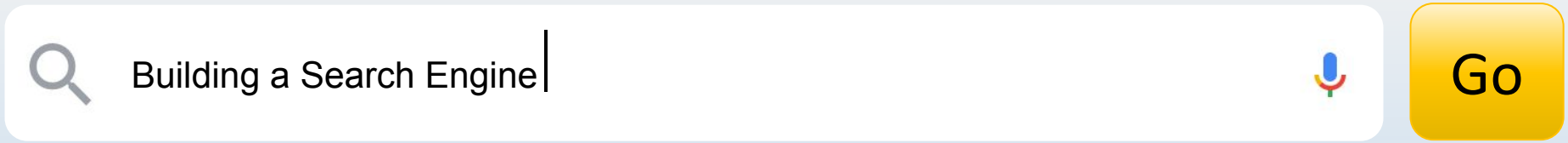
PROJECT

CSCE 2203

Analysis and Design of Algorithms Lab



Search Query

Your search engine takes a string query then displays the sorted results (based on the page score which is detailed in Slide #4) of the retrieved webpages. Your program shall accept search strings containing:

1. Quotations
    - "data structures" → search results will only include webpages containing keyword data structures in the same and the same case.
2. AND
    - data AND structures → search results will include webpages that have the keyword "data" and the keyword "structures".
3. OR
    - data OR structures → results will include webpages that have the keyword "data" or the keyword "structures"
4. A plain search string, like data structures (without quotes, AND or OR) will be treated as data OR structures

Ranking Webpages

PageRank video: https://www.youtube.com/watch?v=P8Kt6Abq_rM

CTR https://en.wikipedia.org/wiki/Click-through_rate

- Webpages have a number of keywords that are used to describe its content. When a search query is issued, a search in the index is initiated to match all the webpages that have keywords that match the query string.
- Then you are required to **sort** the retrieved webpages based on their importance (score), which depends on 2 components:

Part of the project is to research both. As a starting point, click on the links.

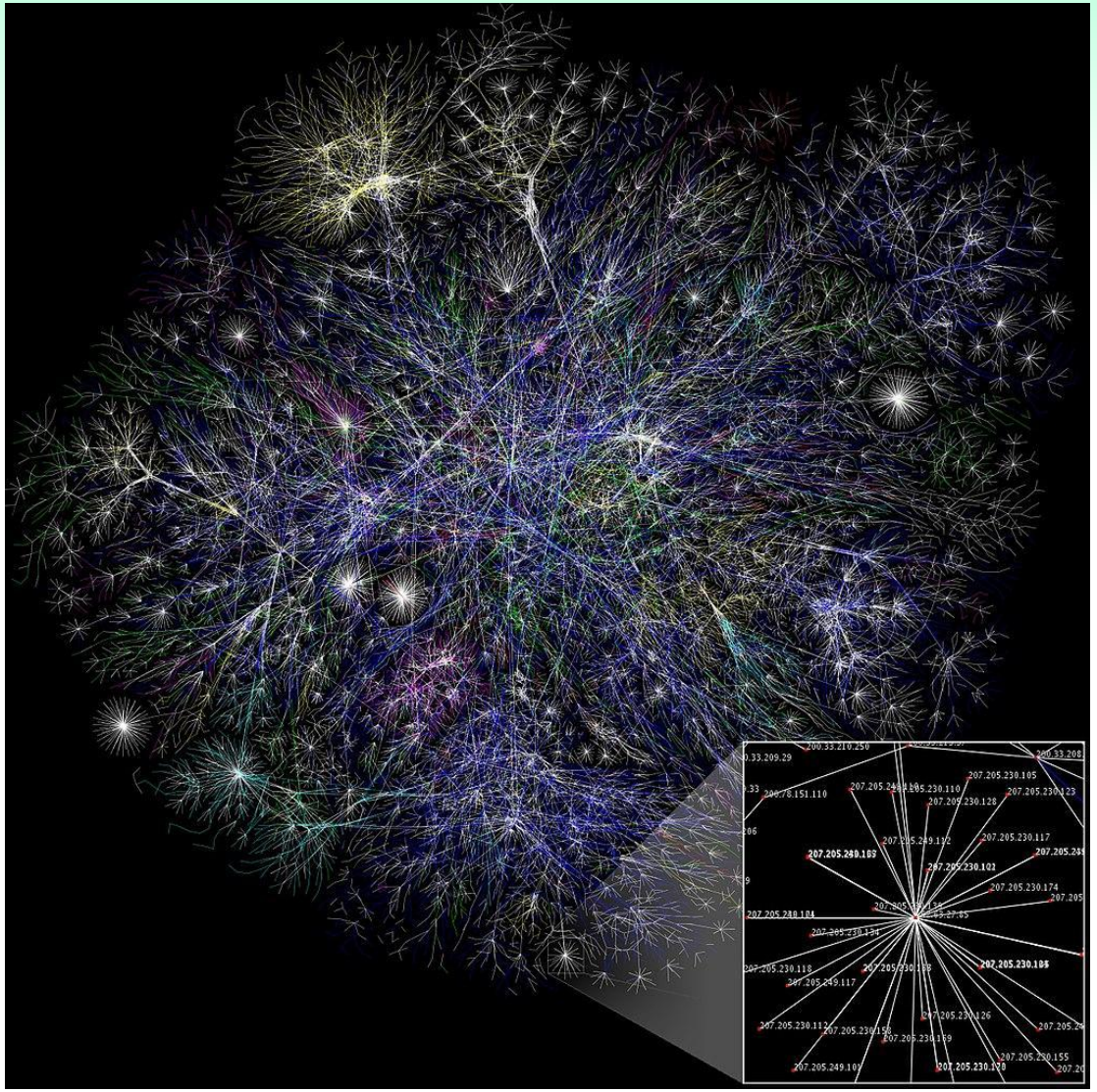
1. PageRank
   - PageRank algorithm is the initial rank that is given to the page when it doesn't have historical click data.
   - The initial importance computed by PageRank is based off the position of the page in the web graph, based on hyperlinks included in the page.
2. CTR
   - Click-Through-Rate (CTR) is the other component of the page score that relies on how users perceive it as important.
   - This metric is calculated based on how many times the page was displayed in search results (also known as *impressions*) and how many times it has been clicked.

PR_{norm} represents the normalized PageRank value across all webpages

$$\boldsymbol{score(page)} = 0.4 \times PR_{norm} + \left(\left(1 - \frac{0.1 \times impressions}{1 + 0.1 \times impressions}\right) \times PR_{norm} + \frac{0.1 \times impressions}{1 + 0.1 \times impressions} \times CTR\right) \times 0.6$$

Web Graph

A **web graph** is a directed **graph**, whose vertices correspond to webpages, and a directed edge connects page X to page Y if there exists a hyperlink on page X, referring to page Y.



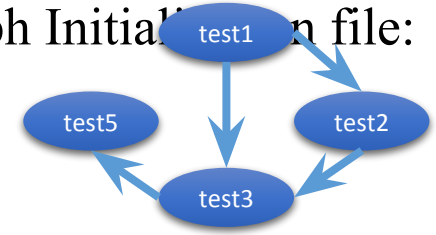
(Wikipedia) Partial map of the Internet in January 15, 2005

Program Initialization

Your program initialization should accept 3 input files:

1. **Web graph file** (in CSV format). Each line in the input file would have two URLs showing a link from the first page to the second page. Sample web graph Initialization file:

```
www.test1.com,www.test2.com
www.test2.com,www.test3.com
www.test3.com,www.test5.com
www.test1.com,www.test3.com
```



2. **Keyword file** (in CSV format), which contains the list of keywords for each webpage. Sample keyword initialization file:

```
www.test1.com,data,structures,complexity
www.test2.com,machine,learning
www.test3.com,programming,complexity,procedural,objects
```

3. **Number of impressions file** (in CSV format), which contains the initial number of times each webpage appeared in the search results (to compute CTR). Sample impressions initialization file:

```
www.test1.com,6
www.test2.com,20
www.test3.com,100
```

Update Number of Clicks

After your program displays the search results (list of relevant webpages sorted by score), the CTR for each webpage must be updated:

1. Your program shall update the number of impressions for the webpages that appeared in the results list. **This updates the 1st component of CTR**.

2. Your program shall allow the user to choose which webpage (among the results list) to open. **This updates the 2nd component of CTR**.

Note: The updated values must be saved onto a file and loaded when the program starts. This way, updates won't be lost when the program ends.

Program Menus

- When your program is initially started, you shall allow the user to either perform a search or exit the program.

- If the user chooses to search, a numbered results list (sorted by webpage score) shall appear to him/her, then he/she shall be allowed to:
  1. Open a webpage among the result by typing in it's number on the list
  2. Perform a new search
  3. Exit the program

- If the user chooses to open a webpage, you shall allow him/her to:
  1. Return to the results list and open a new webpage
  2. Perform a new search
  3. Exit the program

```
Welcome!
What would you like to do?
1. New search
2. Exit

Type in your choice: _
```

```
Search results:
1. www.test4.com
2. www.algorithms101.net
3. www.c_plus_plus_tutorials.org

Would you like to
1. Choose a webpage to open
2. New search
3. Exit

Type in your choice: _
```

```
You're now viewing www.test2.com.
Would you like to
1. Back to search results
2. New search
3. Exit

Type in your choice: _
```

What to Submit

Your submission must include

1. Source code (`.cpp` files)
2. An executable (`.exe`) file to run the engine
3. A report including:
    1) The pseudo-code for your indexing and ranking algorithms
    2) A time and space complexity analysis for your indexing and ranking algorithms
    3) The main data structures used by your algorithm
    4) Any design tradeoffs you made along with their justifications

Project Logistics

- The project carries **30%** of the course's grade.
- The deadline for submitting the project is **November 27, 2020 11:59 PM**.
- Please submit your work on time because no late submissions will be accepted.
- This is an **individual** project.
- AUC's Academic Integrity guidelines will be strictly enforced.
- Good luck!