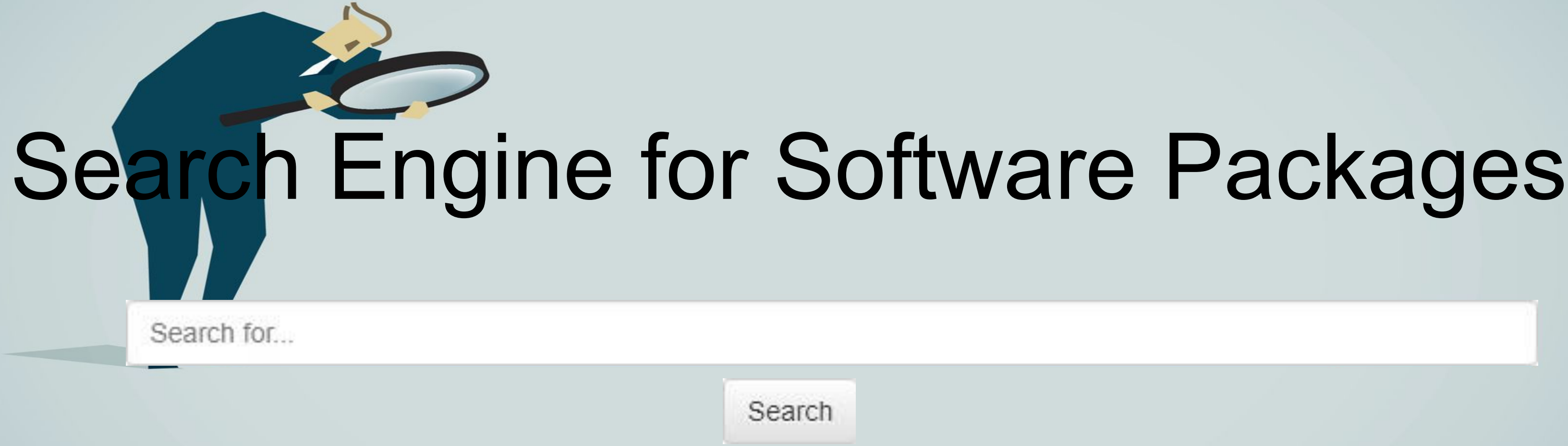
Search Engine for Software Packages



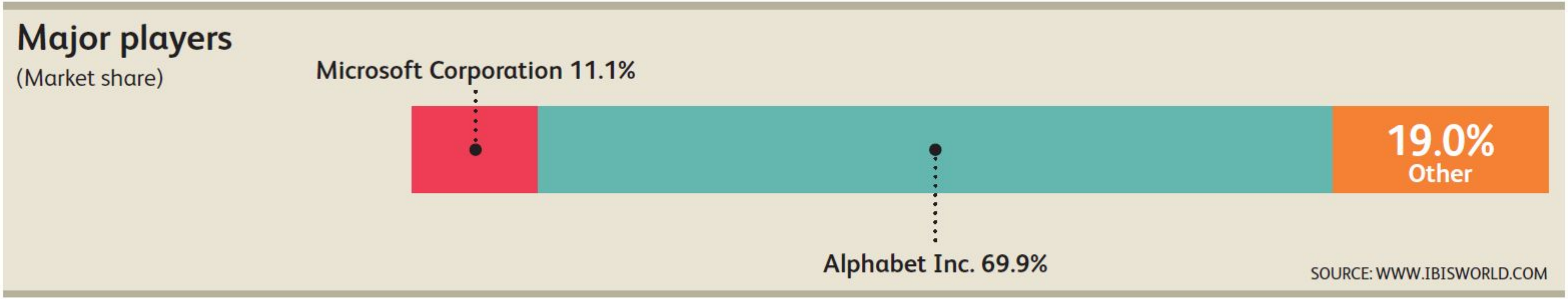
Final Presentation - May 2nd, 2018

Joshua Choo, Mengshi Feng, Vivian Liu, Avery Nisbet, Yidan Zhang

Agenda

- Project Overview
  - Industry Analysis
  - User Interviews
  - Current Solution vs Proposed Solution
- Project Components and Integration
- Demo
- Next Steps

Search Engine Industry: Market Share (US)

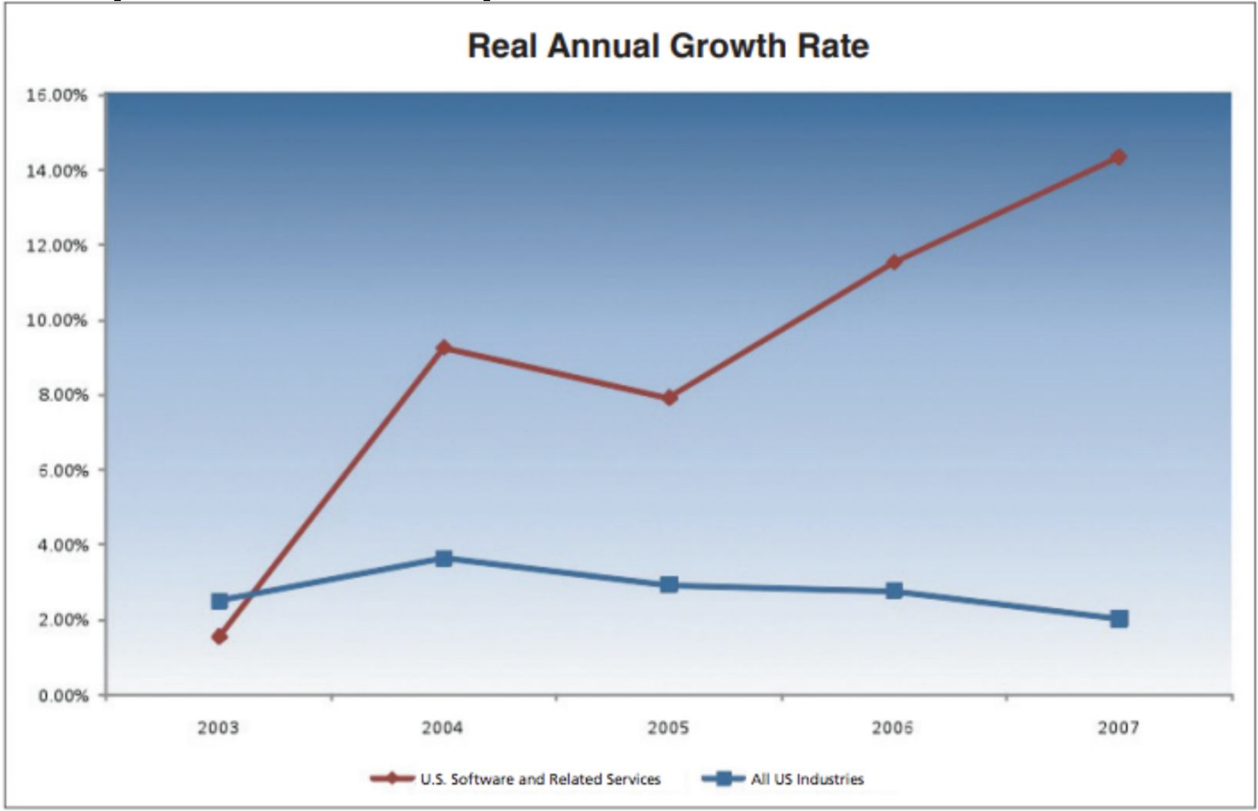


Google has a lion’s share of the market

Search Engines in the US, IBISWorld Industry Report 51913a. (2017). IBISWorld, pp.4-34. Retrieved November 2, 2017, from http://clients1.ibisworld.com/reports/us/industry/default.aspx?entid=1982

Why is this problem important?

- 1 million
- 12 out o stration



Business Software Alliance. (n.d.). Software Industry Facts and Figures. Retrieved November 4, 2017, from http://www.bsa.org/country/public%20policy/~/media/files/policy/security/general/sw_factsfigures.ashx

Why is Software Search Difficult?

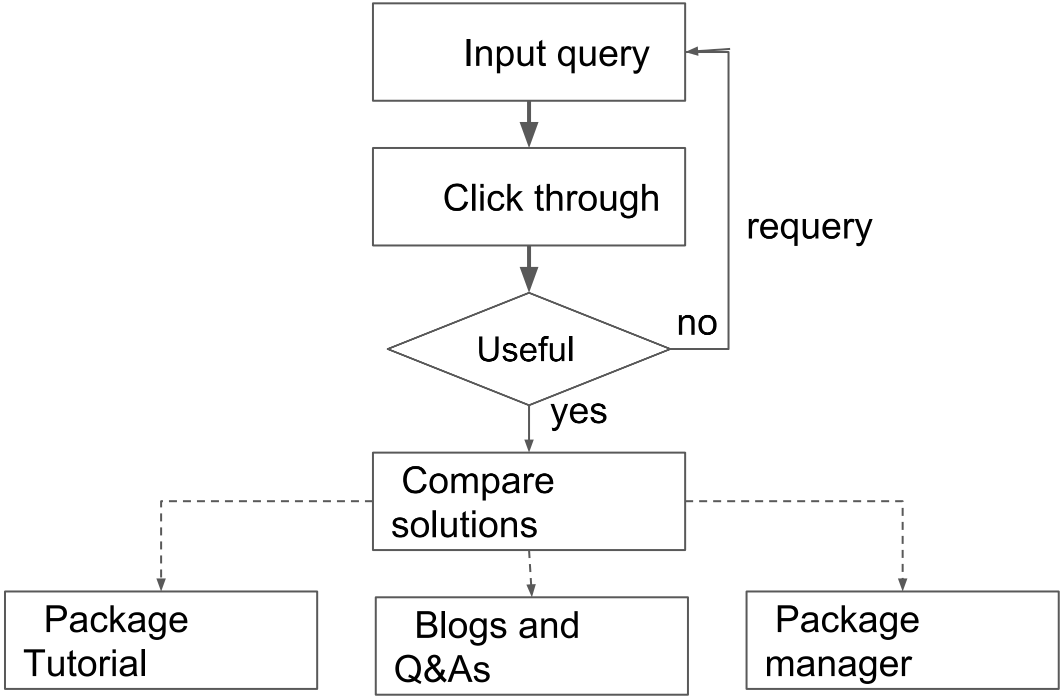
- Difficult to formulate precise queries
- Misleading or outdated documentation
- Time consuming to iteratively evaluate and aggregate useful information across links

Software Package Search: A Problem Left to be Solved

Google

Bing

Not a universal solution!



Input query

Click through

Useful

yes

no

requery

Compare solutions

Package Tutorial

Blogs and Q&As

Package manager

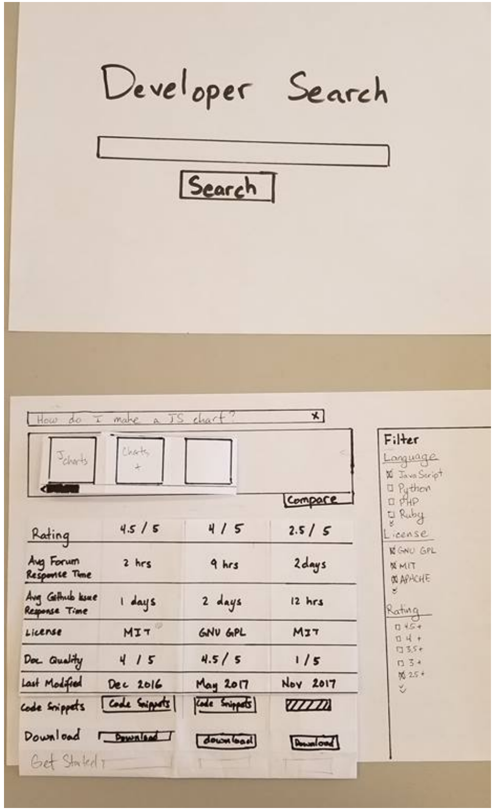
Setting Course: Narrowing Down the Solutions with User Interviews

Package evaluation depends on who searched, and what the search was

List of evaluation criteria created from interviews

(Gallardo-Valencia and Sim, 2011)

Setting Course: Narrowing Down the Solutions with User Interviews



Rating	4.5 / 5	4 / 5	2.5 / 5
Avg Forum Responce Time	2 hrs	9 hrs	2days
Avg Github Issue Response Time	1 days	2 days	12 hrs
License	MIT	GNU GPL	MIT
Doc. Quality	4 / 5	4.5 / 5	1 / 5
Last Modified	Dec 2016	May 2017	Nov 2017
Code Snippets	Code Snippets	Code Snippets	
Download	Download	download	Download
Get Started?			

Filter

Language
- ☒ JavaScript
- ☐ Python
- ☐ PHP
- ☐ Ruby

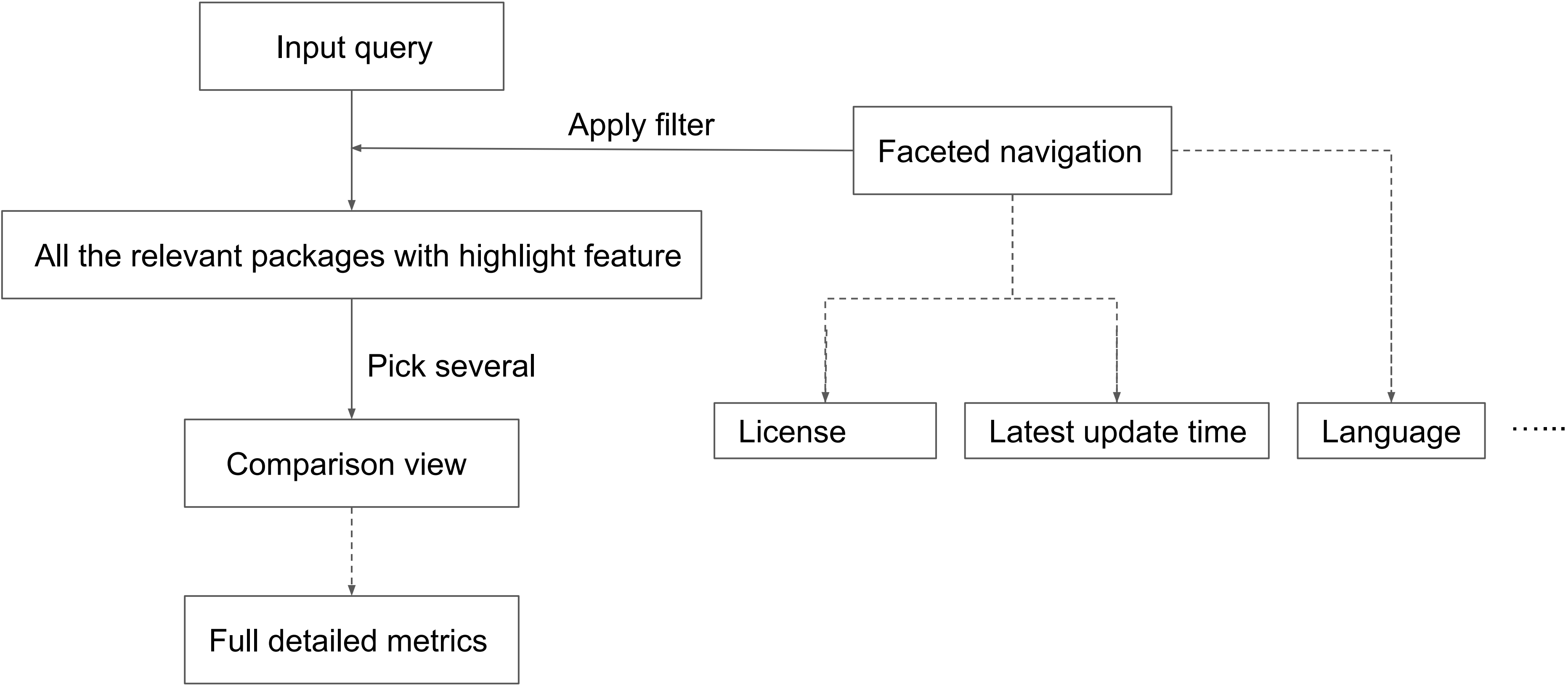
License
- ☒ GNU GPL
- ☒ MIT
- ☒ APACHE

Rating
- ☐ 4.5+
- ☐ 4+
- ☐ 3.5+
- ☐ 3+
- ☒ 2.5+

Testing a low-fidelity prototype:

- Paper Prototypes made from Search Engine Design Features
- Users run through search experience
- Could easily draw a new feature to test

How will people search package using our system?



Input query

Apply filter

Faceted navigation

All the relevant packages with highlight feature

Pick several

License

Latest update time

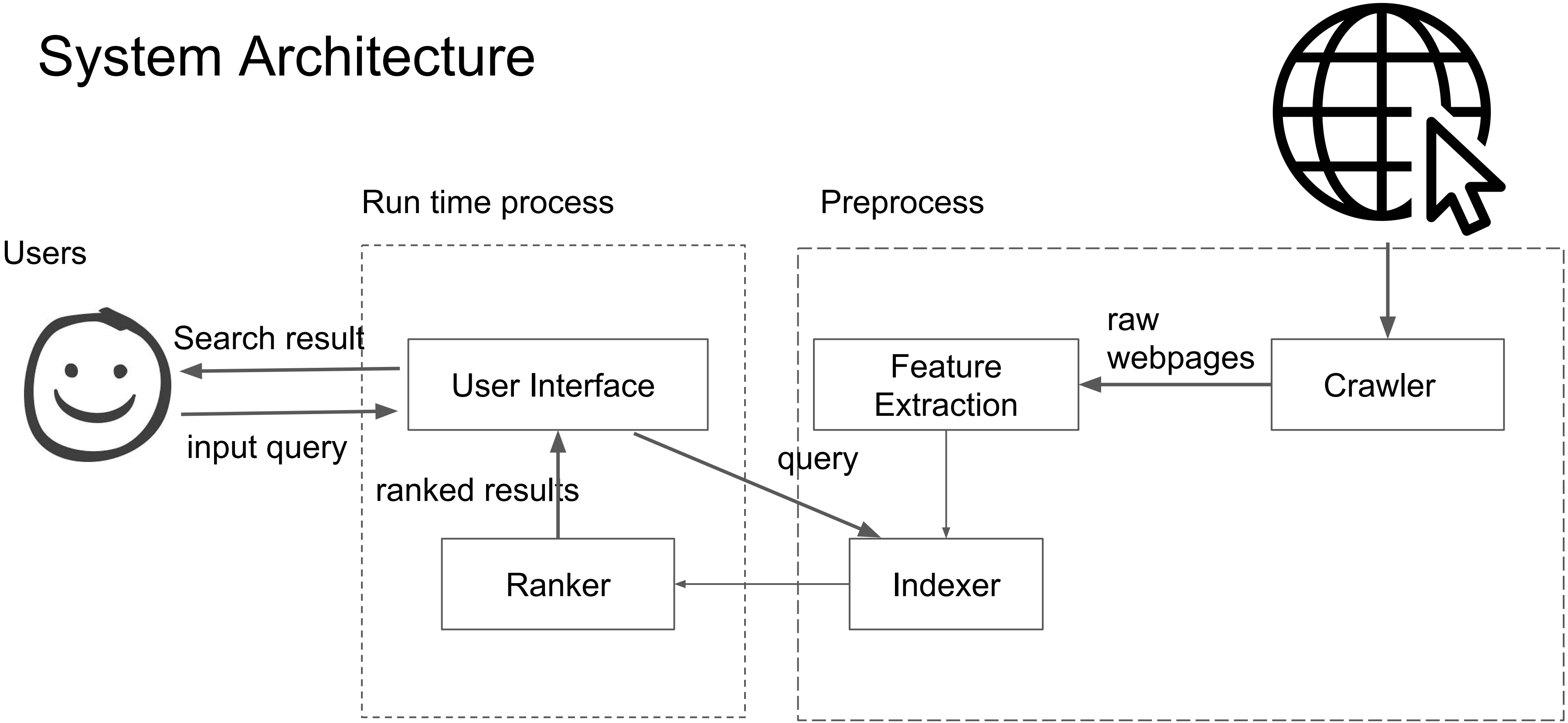
Language

……

Comparison view

Full detailed metrics

System Architecture



Run time process

Preprocess

Users

Search result

input query

User Interface

ranked results

Ranker

query

Feature Extraction

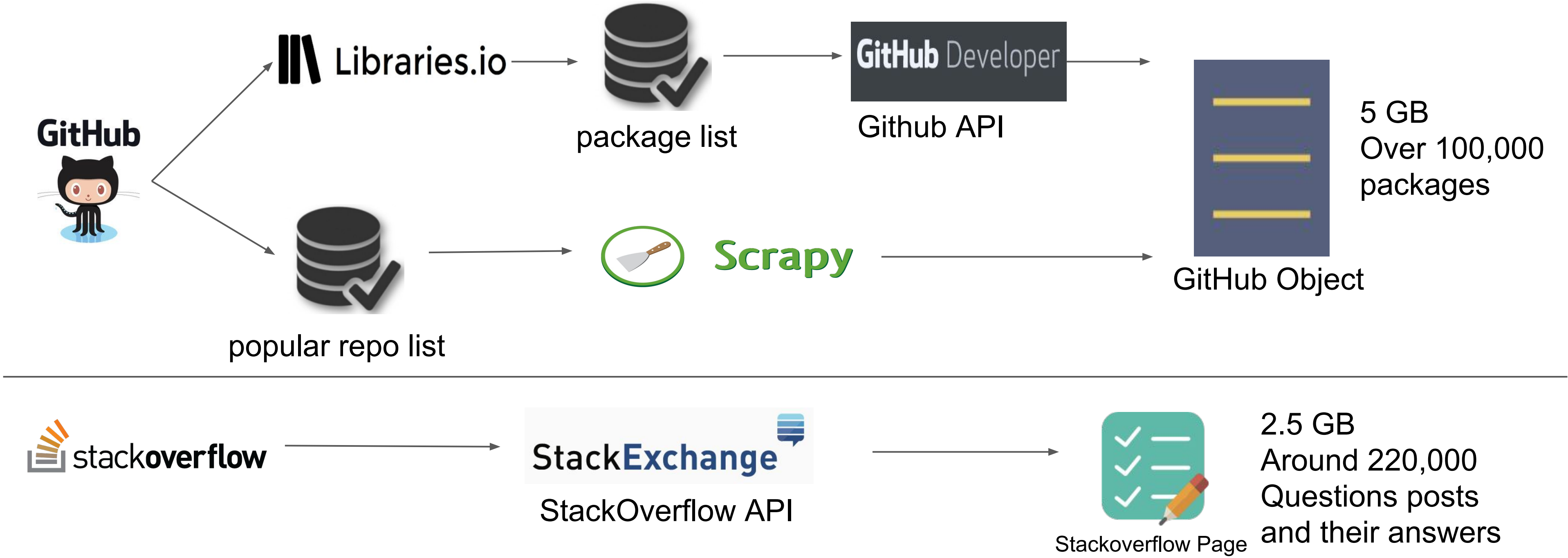
raw webpages

Crawler

Indexer

Gathering Data → Extracting Features → Indexing → Ordering → Displaying

Crawler: Gathering Information



package list

Github API

popular repo list

GitHub Object

5 GB
Over 100,000
packages

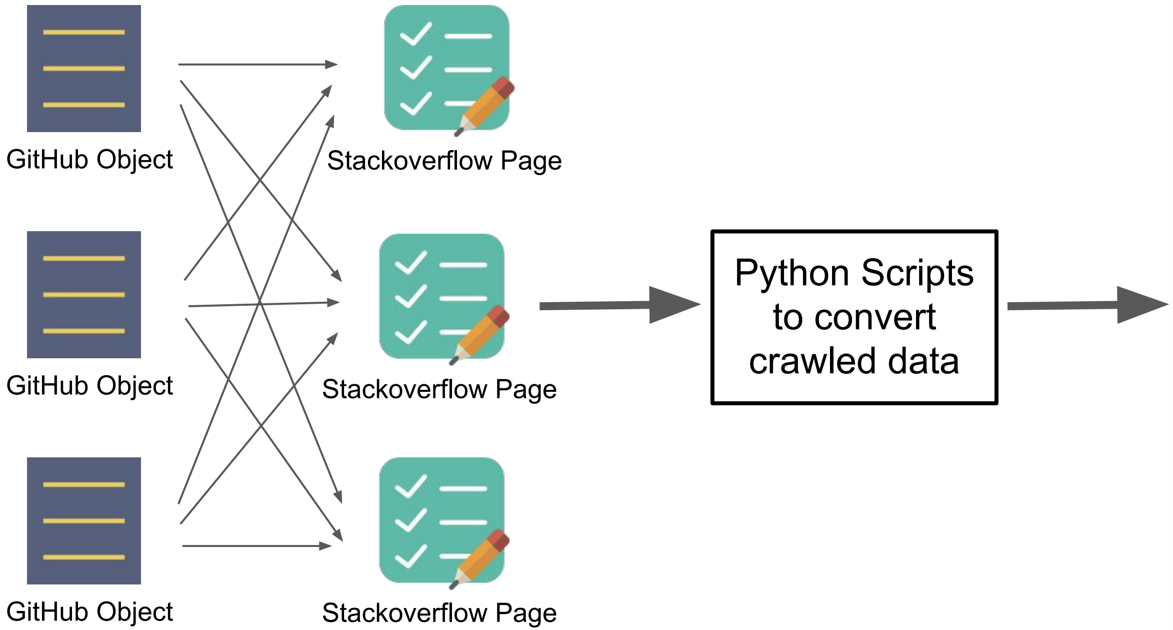
StackOverflow API

Stackoverflow Page

2.5 GB
Around 220,000
Questions posts
and their answers

Gathering Data → Extracting Features → Indexing → Ordering → Displaying

Feature Extraction: Interpreting Unstructured Data



GitHub Object

GitHub Object

GitHub Object

Stackoverflow Page

Stackoverflow Page

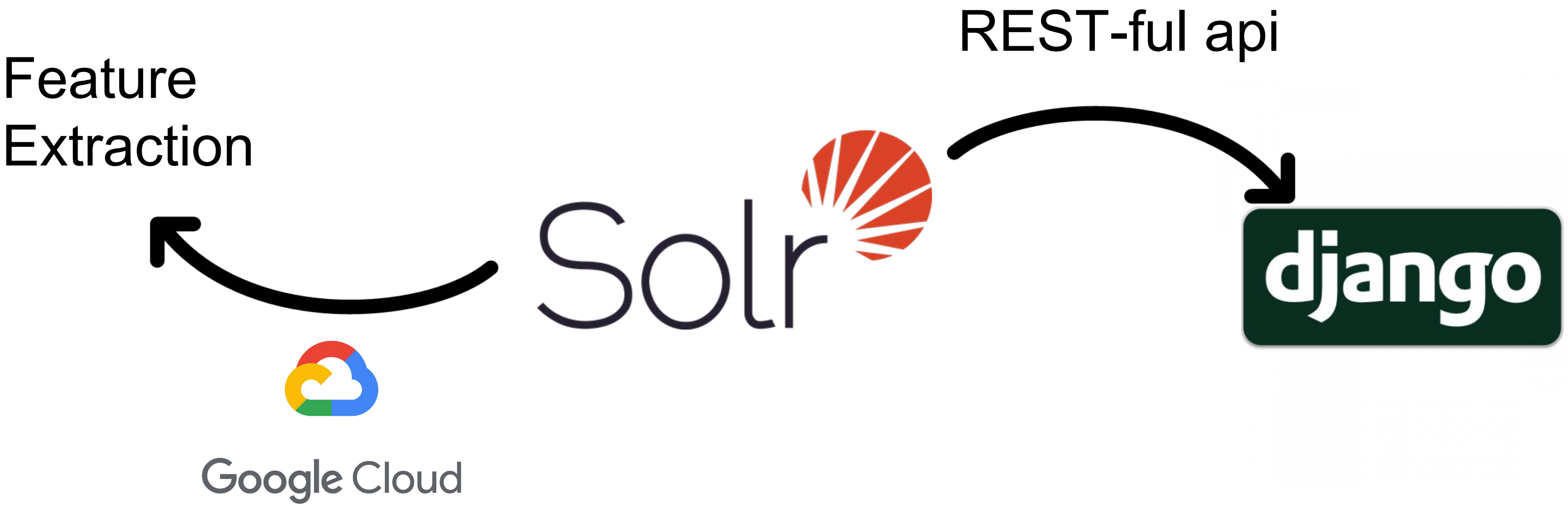
Stackoverflow Page

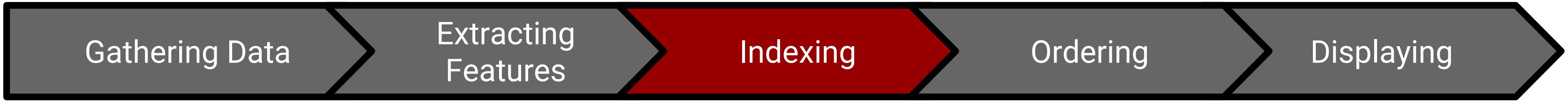
Python Scripts to convert crawled data

```
{
    "Name": "angular",
    "_childDocuments_": [
        {
            "Body_markdown": ".....",
            "Title": ".....",
            "Link": ".....",
            "Code_snippets": ".....",
            "Answers": ".....",
        },
       {
            "Body_markdown": ".....",
            "Title": ".....",
            "Link": ".....",
            "Code_snippets": ".....",
            "Answers": ".....",
        },
        ………...
    ]
}
```

Gathering Data | Extracting Features | Indexing | Ordering | Displaying

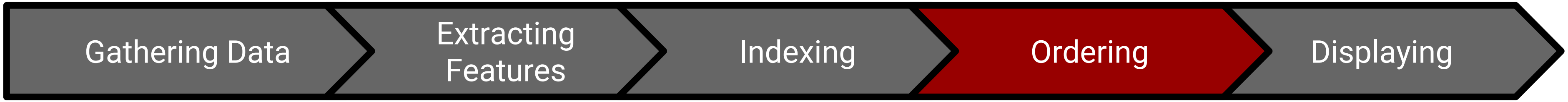
Indexer: Providing Quick Look-up on Data





Ranker: Reordering indexed results

- Score relevance of packages based on Okapi BM25 similarities of package metadata (Github metadata, Stackoverflow documents and Slant queries) and query phrase.
- Each package is tagged with relevant Slant queries and similarity between them and the user’s query are considered.
- Implemented slop parameters to ensure that relevant documents without the exact phrase query would be considered, while not inflating the scores of irrelevant documents that contains the individual words in the phrase query.



Gathering Data

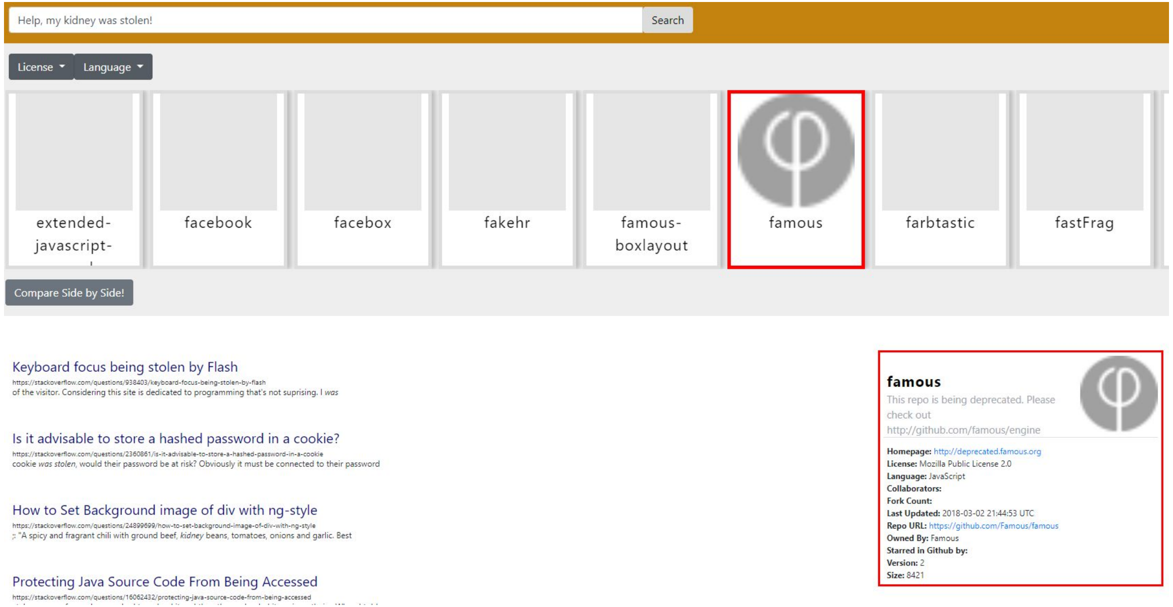
Extracting Features

Indexing

Ordering

Displaying

User Interface: Displaying Information Intuitively



Keyboard focus being stolen by Flash
https://stackoverflow.com/questions/938403/keyboard-focus-being-stolen-by-flash
of the visitor. Considering this site is dedicated to programming that's not suprising. I *was*

Is it advisable to store a hashed password in a cookie?
https://stackoverflow.com/questions/2360861/is-it-advisable-to-store-a-hashed-password-in-a-cookie
cookie *was stolen*, would their password be at risk? Obviously it must be connected to their password

How to Set Background image of div with ng-style
https://stackoverflow.com/questions/24899699/how-to-set-background-image-of-div-with-ng-style
;: "A spicy and fragrant chili with ground beef, *kidney* beans, tomatoes, onions and garlic. Best

Protecting Java Source Code From Being Accessed
https://stackoverflow.com/questions/16062432/protecting-java-source-code-from-being-accessed

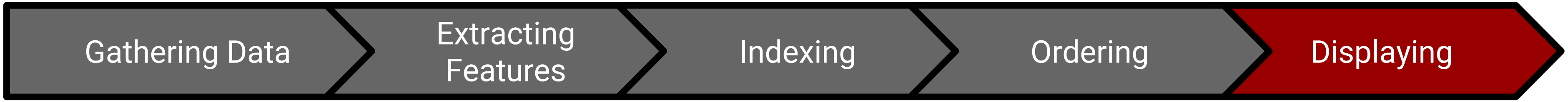
famous

This repo is being deprecated. Please check out http://github.com/famous/engine

Homepage: http://deprecated.famous.org
License: Mozilla Public License 2.0
Language: JavaScript
Collaborators:
Fork Count:
Last Updated: 2018-03-02 21:44:53 UTC
Repo URL: https://github.com/Famous/famous
Owned By: Famous
Starred in Github by:
Version: 2
Size: 8421

Features:

- **Carousel** for packages.
- **Details** for the package appear on package selection.
- **Filters** help you narrow down your search.
- **Keywords** from the query are accepted as filters.
- **Compare** packages side by side.
- **Stackoverflow** results help you get opinions of other developers.


Gathering Data → Extracting Features → Indexing → Ordering → Displaying

Demo

http://35.230.66.167/

Next Steps

Continuous and Exhaustive Crawling

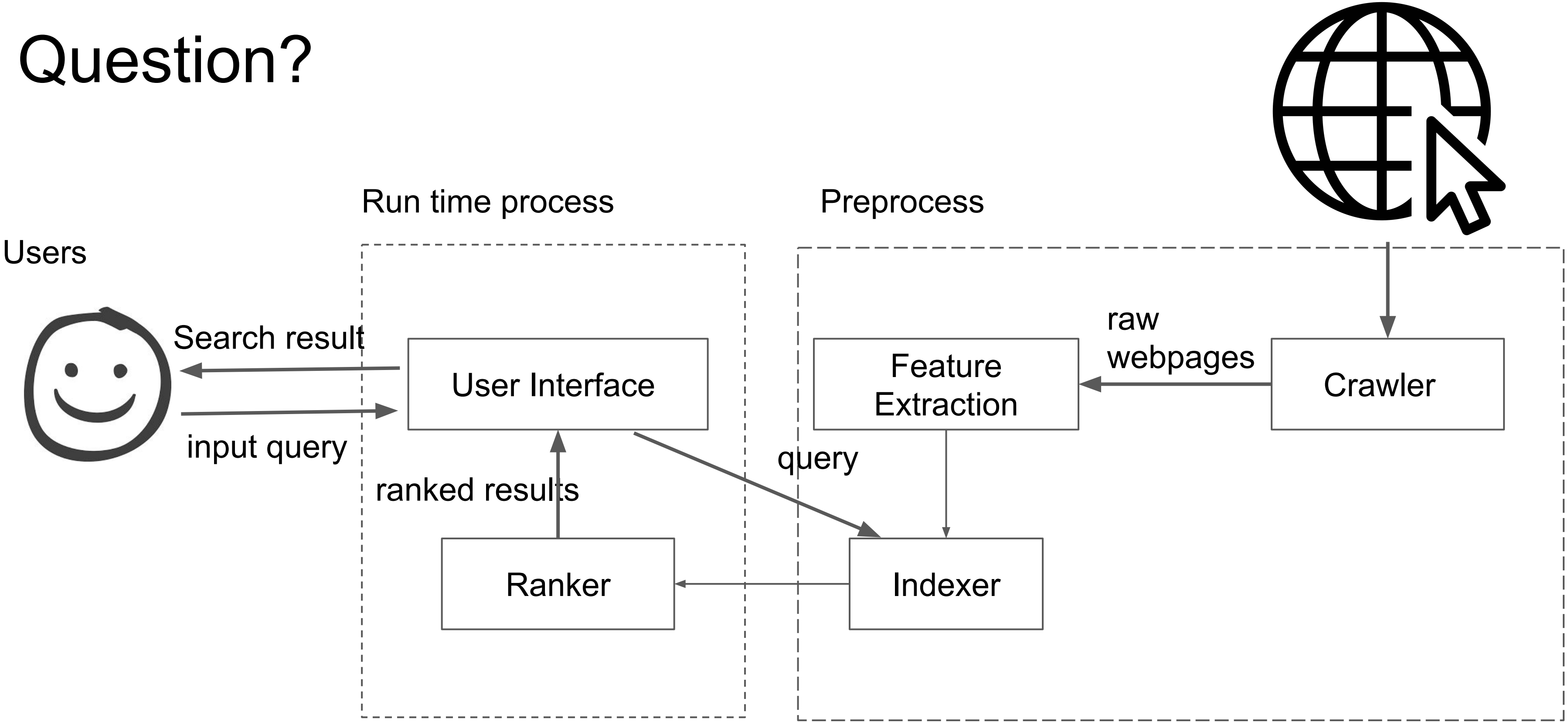
Data Pipeline

Recognition of User Queries

Extraction of More Metadata

Improved Ranking Algorithm

Question?


Run time process
Preprocess
Users
Search result
input query
User Interface
ranked results
Ranker
query
Feature
Extraction
raw
webpages
Crawler
Indexer

Appendix -- Solr Schema

1st level

```
{
          “name”: “Angular”,

          “path”: “1.git”,

          “homepage_url”: "https://angular.io",

          “headme”: “...”,

          ...
```

2nd level

```
“_childDocuments_”: {

          "Title": "Capture Video of Android&#39;s Screen",

          “path”: “2.stack”,

          “_childDocuments_”

                    “path”: “3.stack.answer”,

                    “Answer_id”: “x”

}

}
```

3rd level