Engineering Candidate Assignment - Scale

Before you begin, please reply to this email confirming you have received these instructions. Thank you!

We're excited for you to continue the interviewing process for the Software Engineering position here at BrightEdge! You have 48 hours from the time this email is sent to complete the challenge.

DEADLINE: <Deadline>

Deliverables: Please send me your files via email or public download location when you are finished. Please rename code extension to .txt if sending view email

Let's begin!

With a background in SEO, it's important to write good content and signal to the search engines what your article or page is about. One way to ensure that is by focusing your content on the core topics. In this assignment, you'll be responsible for designing a service that will take billions of URL and identify the metadata on the page, allow millions of requests on the content, and optimize for cost, performance and availability.

Assignment:

Part 1:

Develop a core crawler to crawl the content of a given URL.

Given any page (URL), be able to classify the page, and return a list of relevant topics. We'd like to have you build it generically, but for testing purposes, please consider the following URLs:

http://www.amazon.com/Cuisinart-CPT-122-Compact-2-Slice-Toaster/dp/B009GQ034C/ref=sr_1_1?s=kitchen&ie=UTF8&qid=1431620315&sr=1-1&keywords=toaster

http://blog.rei.com/camp/how-to-introduce-your-indoorsy-friend-to-the-outdoors/

http://www.cnn.com/2013/06/10/politics/edward-snowden-profile/

Input: Any URL (for testing purposes, please consider the URLs above)

Output: Meta Data of the HTML, such as title, description, body and etc.

Pick your favorite choice of language.

Demo code available on a public service, such as AWS, Azure or GCP

Part 2:

Provide design documentation to operationalize the collection of billions of URLs using the code developed. Propose the next steps, how to further optimize for cost, reliability, performance, and scale.

Input:

List of billions of URLs send in via a text file and/or in MySQL for a given year month

e.g. billions of URLs for amazon.com, walmart.com, bestbuy.com etc for July

Output: Design storage of the metadata and content. Design for unified data schema. Design for configurability, politeness and respect Robots.txt. Define SLOs and SLAs. Elaborate on the key monitoring metrics and tools you would employ to effectively track the system's progress.

Part 3:

Break down on how to proceed with engineering to Proof of Concept. What are the list of potential blockers. What are known and trivial and what are the estimates arrival time. Implementation schedules. How to have a successful and highly quality release.

Output: Documentations on how to proceed to next steps. Documentations of how to evaluate the proof of concept. Describe your plan for distributing ownership and responsibilities of this service among team members. Release plan. Resources and time estimations.

What we are evaluating from the assignment:

- Design for reliability, performance, scale and cost
- Methods, services, and frameworks used in the design
- Methods on evaluation of the design, implementations, and operations.
- Break Down Proof of Concept
- Break Down on estimates
- Documentation

What's allowed: You are allowed to crawl/parse a page with 3rd party libraries.

What's not allowed: You are not permitted to use a 3rd party services that offer the same functionality.

FAQs:

Q: Can I use external libraries for HTML parsing?

A: Yes, you can.

Q: Can I use external libraries?

A: Yes you can.

Q: Can I use external services?

A: Yes you can uses services offered by GCP, AWS or Azure.

Questions about the assignment: jobs.questions@brightedge.com