Discussion 5

Quiz 2 Review

Welcome! Information Retrieval - Discussion 4

Itinerary October 20, 2021

- Quiz 2 Review

Deadlines

- Quiz 2 on Friday
- Testing period, Assignment 2 is this weekend until Oct. 31 9pm
- Assignment 2 due Friday, Nov. 5

Definitions (1 pt each)

1. Crawler trap
2. Frontier of a web crawler
3. Search engine optimization
4. Politeness
5. Cache server
6. Deep web

True or False (1 pt each)

1. Crawlers should wait between requests to the same web site
2. If crawlers hit the same website as fast as possible, this will make crawling overall faster

What’s that HTTP Status code? (1 pt each)

2xx

3xx

4xx

a. Client error
b. Success
c. Redirection

Find me some doc! (1 pt each)

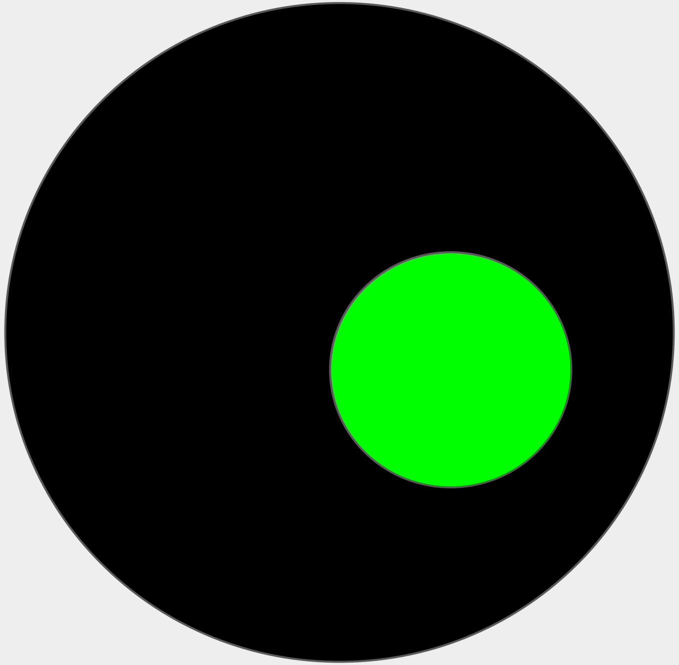
```
def is_valid(url):
   try:
       parsed = urlparse(url)
       if parsed.scheme not in set(["http",
"https"]):
           return False
       return re.match(r".*\.(doc)$",
parsed.path.lower())

   except TypeError:
       print ("TypeError for ", parsed)
       raise
```

Would this condition be sufficient for validating the URLs and documents?

UR What?? 1pt each

1. What is a **URL**?
    a. Give an example
2. What is a **URI**?
    a. Give an example

Which goes in the **inner** set, and which goes in the **outer** set?



Categorize each color 1pt each

```
https://www.example.com/forum/questions/?tag=networking&order=newest#top
```

Categorize each color 1pt each

Categorize each color of the URI.

https://www.brookerules.com/info/?month=birthday&m=05&d=26#bot

Next Week's Discussion

Tentative plan for next week's discussion based on upcoming course deadlines.

- Homework 2

Recommended Homework

To best prepare for next week's session, I recommend you do the following.

- **Make progress on web crawler homework this weekend!**