Started on	Monday, 14 February 2022, 9:12 AM
State	Finished
Completed on	Monday, 14 February 2022, 9:31 AM
Time taken	19 mins 39 secs
Marks	7.42/10.00
Grade	**74.17** out of 100.00

Question **1**

Correct

Mark 1.00 out of 1.00

One of the information retrieval cornerstones is a concept of relevance ✔ . This term describes subjective ✔ metric: if search result satisfies user's information need ✔ , then we consider this result relevant ✔ . Typical scale for relevance ✔ values is 0..1 ✔ , where maximum ✔ means the best result.

Your answer is correct.

Relevance is a subjective measure, while accuracy and precision do not depend on a human.

"contains" and "is close" do not match with any of proposed metrics.

Relevance is usually assessed with non-negative numbers, including 0 for non-relevant.

The correct answer is:
One of the information retrieval cornerstones is a concept of [relevance]. This term describes [subjective] metric: if search result [satisfies user's information need], then we consider this result [relevant]. Typical scale for [relevance] values is [0..1], where [maximum] means the best result.

Question 2

Correct

Mark 1.00 out of 1.00

Compute **Precision@K** for **K=8** if we have the following relevance assessments:

1. 1
2. 0
3. 1
4. 0
5. 1
6. 1
7. 1
8. 1
9. 0
10. 0

Write your answer as a **decimal number** using point symbol

.

Answer: 0.75 ✔

The correct answer is: 0.75

Question 3

Correct

Mark 1.00 out of 1.00

You are sniffing WiFi traffic in a room to capture a password.

You found that someone established a TCP connection to port 80 with **192.34.57.61**

Then you see the following string in your logs:

```
GET /files/my.password HTTP/1.1\nhost: sprotasov.ru\n\n
```

What is the password?

- a. sprotasov.ru
- b. ;) ✔
- c. HTTP/1.1 404 Not Found
- d. :`(
- e. No such URL :(
- f. :)
- g. IDK
- h. HTTP/1.1 200 OK

Your answer is correct.

The correct answer is:
;)

Question **4**

Partially correct

Mark 0.75 out of 1.00

Please arrange the stages of browser work, trying to open **https://github.com/**:

First,	extracts hostname from the URL	✔
Second,	probes DNS service to get host IP	✔
Third,	performs TLS handshake	✖
Fourth,	performs TCP handshake with remote machine at port 443	✖
Fifth,	sends HTTP GET request	✔
Sixth,	recieves HTTP response with page text is data section	✔
Seventh,	constructs DOM	✔
Eighth,	renders the page	✔

Your answer is partially correct.

You have correctly selected 6.
The correct answer is:
First, → extracts hostname from the URL,

Second, → probes DNS service to get host IP,

Third, → performs TCP handshake with remote machine at port 443,

Fourth, → performs TLS handshake,

Fifth, → sends HTTP GET request,

Sixth, → recieves HTTP response with page text is data section,

Seventh, → constructs DOM,

Eighth, → renders the page

Question **5**

Incorrect

Mark 0.00 out of 2.00

Imagine you recieved response for a query with 5 documents.

Assessor one evaluates them as [Rel, ***NotRel***, Rel, Rel, Rel] (left is higher)

Assessor two evaluates them as [Rel, Rel, Rel, ***NotRel***, Rel]

What is the difference (**AP@5[1] - AP@5[2]**) of their assessments?

NB:

- **Use** "stepwise" **fix** for the resulting curve.
- **AP** values are from **0..1 range**.
- Result is a **decimal number**, not a fraction.
- Use **point** as decimal separator.
- **Signs** matter.

Answer: 0 ✖

```
First
Precision:0.25  0.25  0.50  0.75  1.00
Recall-1: 1.00  0.50  0.67  0.75  0.80
Recall-2: 1.00  0.80  0.80  0.80  0.80

Total = (1 + 0.8 * 3) / 4 = 0.85

Second
Precision:0.25  0.50  0.75  0.75  1.00
Recall-1: 1.00  1.00  1.00  0.75  0.80
Recall-2: 1.00  1.00  1.00  0.80  0.80

Total = (1 * 3 + 0.8) / 4 = 0.95


0.85 - 0.95 = -0.1
```

The correct answer is: -0.1

Question 6

Partially correct

Mark 0.67 out of 1.00

Which of the quality metrics, are **both**:

- can be applied to a *single search engine result page* (SERP)
- take item *ranks* into consideration

- [ ] a. IR-way defined recall (out or result)
- [ ] b. accuracy
- [ ] c. pFound
- [x] d. Mean reciprocal rank (MRR) ✔
- [x] e. discounted cumulative gain (DCG) ✔
- [ ] f. A/B test
- [ ] g. ML-defined defined recall (out of all positive)
- [ ] h. cumulative gain (CG)

Your answer is partially correct.

You have correctly selected 2.
accuracy, recall (any) are just ratios.

cumulative gain = sum of the graded relevance values of all results in a search result list.

A/B testing is a process, not a metric.

The correct answers are:
Mean reciprocal rank (MRR),

discounted cumulative gain (DCG),

pFound

Question 7

Correct

Mark 1.00 out of 1.00

Examine **robots.txt** file at the website of our city (**innopolis.ru**). It specifically prohibits to index CMS (content management system) page, which name is among the very first lines in the file. Guessed, which one? Open this prohibited page in browser and see where it redirects you. Copy the full URL of web page you were redirected to (**including https, #,** and all other things) into the answer field.

Answer: https://innopolis.ru/bitrix/admin/index.php#authorize ✔

The correct answer is: https://innopolis.ru/bitrix/admin/index.php#authorize

Question 8

Correct

Mark 1.00 out of 1.00

Which IP-addresses belong to Yandex (are serving *[*.]yandex.ru* domain requests)? Mark **all**. Use any tools which can help you to answer the question.

- [x] a. 5.255.255.77 ✔
- [ ] b. 192.34.57.61
- [x] c. 213.180.193.56 ✔
- [x] d. 77.88.55.77 ✔

Your answer is correct.

nslookup [IP]

for both Windows and Nix.

The correct answers are:
77.88.55.77,

213.180.193.56,

5.255.255.77

Question 9

Correct

Mark 1.00 out of 1.00

You want to crawl a given website, e.g. http://www.ir-course.com/, with python code. Please, select **all correct** statements:

- [ ] a. It is impossible to crawl a website, if it is accessible from the browser, but *requests.get()* returns HTTP status code **451** (protected for legal reasons).
- [ ] b. The file *http://www.ir-course.com/sitemap.xml* cannot help us to speed up crawling.
- [x] c. The file *http://www.ir-course.com/robots.txt* contains information about pages, that website owner does not want to be crawled. ✔ yes, this is one of the options which you can specify in robots.txt.
- [x] d. With no additional information or tool, we will start from the main page, and then visit every reachable page, following the links. ✔ Sure, if **no tools (sitemaps, external engines, ...) given**, you will do it like this.
- [ ] e. It is possible to obtain DOM using *requests.get()* method, then parse it using *BeatifulSoup*, and then extract information from a * * tag, which is rendered asyncronously by javascript (for example after a click).

Your answer is correct.

The correct answers are:
The file *http://www.ir-course.com/robots.txt* contains information about pages, that website owner does not want to be crawled.,

With no additional information or tool, we will start from the main page, and then visit every reachable page, following the links.

◄ 02 - Quality

Jump to...

IR05. Text documents indexing ►