RK-J - Bug #2787

sitemap URL with non-alphanumeric chars is not working

12/30/2022 05:05 PM - Venmuhilan B

Status:	Closed	**Start date:**	12/30/2022
Priority:	Normal	**Due date:**	
Assignee:	Ayush Khandelwal	**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:	1-RK-J-1	**Spent time:**	0.00 hour

Description

I ingested the book(book name with special characters). There are no errors occurred during ingestion. But I checked the sitemap files for those books. Sitemap URL for those books are present in the public URL. But when we open it, it is showing error message. I think that is the issue.

A) sitemap file for the book where the bookName has special characters(like colon, semicolon, brackets, paranthesis) and alphanumeric:
(showing Error message when visiting the sitemap URL)
ex:
1. Tensorflow API Doc TF Module (Python)-PL-2022-10-30 16:23:38.140171-SPL
-
https://edutestdev-240612.appspot.com/sitemap/edutestdev_sitemap/sitemap-tensorflow-api-doc-tf-module-%28python%29-2022-11-12-14%3A27%3A00.927566-sg12nov22.xml
2.
https://edutestdev-240612.appspot.com/sitemap/edutestdev_sitemap/sitemap-python-python-for-beginners-%28full-course%29-dec01ak.xml

3.
https://edutestdev-240612.appspot.com/sitemap/edutestdev_sitemap/sitemap-deep-learning-andrew-ng%2C-coursera-course-dsnov24ko.xml
B) sitemap file for the book where the book name has only alphanumeric characters:
(Sitemap URL works fine)
ex:
1. https://edutestdev-240612.appspot.com/sitemap/edutestdev_sitemap/sitemap-python-3-tutorial-prd.xml
2. https://edutestdev-240612.appspot.com/sitemap/edutestdev_sitemap/sitemap-python-3-language-reference-prd.xml
3. https://edutestdev-240612.appspot.com/sitemap/edutestdev_sitemap/sitemap-ticket-2738-fix.xml

also check more URLs in edutestdev_sitemap - public URL:
https://storage.cloud.google.com/edutestdev_sitemap/sitemap.xml

History

#1 - 12/30/2022 05:07 PM - Venmuhilan B

- File Screenshot from 2022-12-30 22-36-35.png added

#2 - 12/30/2022 05:19 PM - Venmuhilan B

findings(Current implementation):
sitemap filename - sitemap-subject-book-version.xml:
sitemap-python-3.9-library-reference-(dt)-prd.xml

- After UTF-8 encoding in backend:*
  sitemap-python-3.9-library-reference-%28dt%29-prd.xml

- sitemap filename for the book uploaded to edutestqa_sitemap bucket:*
  sitemap-python-3.9-library-reference-%28dt%29-prd.xml
  - In sitemap-python-3.9-library-reference-%28dt%29-prd.xml file,
  public URL - https://storage.googleapis.com/edutestqa_sitemap/sitemap-python-3.9-library-reference-%2528dt%2529-prd.xml
  - It is working and it is double encoded by the GCP.
  (iv)
  sitemap entry for the book present in sitemap.xml:
  - https://edutestqa.appspot.com/sitemap/edutestqa_sitemap/sitemap-python-3.9-library-reference-%28dt%29-prd.xml (not working)

Tried double encoding the sitemap filename for above URL:
- https://edutestqa.appspot.com/sitemap/edutestqa_sitemap/sitemap-python-3.9-library-reference-%2528dt%2529-prd.xml (not working)

Tried encoding the double encoded filename:(triple encoding)
- https://edutestqa.appspot.com/sitemap/edutestqa_sitemap/sitemap-python-3.9-library-reference-%252528dt%252529-prd.xml (working)

For URL that contains non-alphanumeric characters(which are double encoded)
- URL starts with https://storage.googleapis.com/edutestqa_sitemap/ is working
- URL starts with https://edutestqa.appspot.com/sitemap/edutestqa_sitemap/ is not working

conclusion with ex:
This URL(https://edutestqa.appspot.com/sitemap/edutestqa_sitemap/sitemap-python-3.9-library-reference-%28dt%29-prd.xml) is present in sitemap entry for the book in sitemap.xml. In the current implementation, it is doing UTF-8 encoding for filename. The sitemap URL is not working for the filename that has non-alphanumeric characters.
So, If we want the URL with non-alphanumeric characters to work, we have to encode the double encoded filename and use that URL for sitemap.xml.
i.e https://edutestqa.appspot.com/sitemap/edutestqa_sitemap/sitemap-python-3.9-library-reference-%252528dt%252529-prd.xml

#3 - 12/30/2022 05:22 PM - Venmuhilan B

Fix:

In the current implementation filename is stored in the bucket as:
sitemap-vb2583-tutorial-sitemap-%28test%29-2747.xml
So, we have to do the triple encoding to make the URL work i.e (sitemap-vb2583-tutorial-sitemap-%252528test%252529-2747.xml). This will fix the issue. But if the filename has many special characters it will increase the no of characters in URL. It is not efficient

So, To fix this:
we need to save the filename without any encoding to bucket :
sitemap-vb2583-tutorial-sitemap-(test)-2747.xml

and we do the UTF-8 encoding on filename only for the sitemapURL
https://edutestdev-240612.appspot.com/sitemap/edutestdev_sitemap/sitemap-vb2583-tutorial-sitemap-%28test%29-2747.xml

In this way, we can solve the issue.It is working.

#4 - 12/30/2022 05:22 PM - Venmuhilan B

- Status changed from New to In Progress

- % Done changed from 0 to 90

#5 - 12/30/2022 05:22 PM - Venmuhilan B

- Status changed from In Progress to Resolved

- % Done changed from 90 to 100

#6 - 06/14/2023 01:32 PM - Venmuhilan B

- Status changed from Resolved to Feedback

- Assignee changed from Venmuhilan B to Ayush Khandelwal

#7 - 07/10/2023 08:13 AM - Ayush Khandelwal

- Status changed from Feedback to Closed

Working as expected

Files

Screenshot from 2022-12-30 22-36-35.png	104 KB	12/30/2022	Venmuhilan B