

SWE 225 Information Retrieval Assignment 3 Milestone 1 Report

Flavor: Algorithms and Data Structures Developer

Team 9:

Pooja Bhatia, 14567073

Rutuja Pansare, 55443146

Ella Dodor, 31129450

Samyak Jhaveri, 13043185

Code

```
from bs4 import BeautifulSoup
from pymongo import MongoClient
import json
from nltk.tokenize import word_tokenize
from nltk.stem import SnowballStemmer
import glob

client = MongoClient()
mydb=client["latestDB"]
myCollection=mydb["tokenTable"]
ss = SnowballStemmer(language="english")

for name in glob.glob('/Users/poojabhatia/Documents/Info
Retrieval/Assignment 3/DEV/*/*'):
    f = open(name)
    title, body_text="", ""
    data = json.load(f)
    html=data['content']
    url=data['url']
    print(url)
```

```
soup = BeautifulSoup(html, "html.parser")
body_text += soup.get_text()
if soup.title == None and body_text == None:
    continue
if soup.title == None or soup.title.string == None:
    title = ""
if body_text == None:
    body_text = ""
if body_text == "" and title == "":
    continue;

all_tokens = word_tokenize(title+body_text)
only_words= [w.lower() for w in all_tokens if w.isalnum()]
#print(only_words)
f.close()
curr_file_tokens=dict()
for word in only_words:
    word=ss.stem(word)
    if word in curr_file_tokens:
        curr_file_tokens[word]+=1
    else:
        curr_file_tokens[word]=1

url_escaped=url.replace(".", "\\")
for token, count in curr_file_tokens.items():
    myCollection.update_one({"_id": token}, {"$set":
{url_escaped:count}}, upsert=True)
```

Index Analytics

Number of Indexed Documents	55386
Number of Unique Tokens	737314
Total size in kB of index on Disk	193000 kB

Screenshot

The screenshot displays the latestDB.tokenTable interface. At the top, a breadcrumb shows the path 'latestDB.tokenTable' and 'Documents'. Below this, a summary bar provides key metrics: DOCUMENTS 737.3k, STORAGE SIZE 193.0MB, AVG. SIZE 1.2KB, INDEXES 1, TOTAL SIZE 18.4MB, and AVG. SIZE 18.4MB. The interface includes tabs for Documents, Aggregations, Schema, Explain Plan, Indexes, and Validation. A search bar with a filter '{ field: 'value' }' and buttons for OPTIONS, FIND, RESET, and a refresh icon is present. Below the search bar, there are buttons for ADD DATA, VIEW, and a list of view options. A status bar indicates 'Displaying documents 1 - 20 of 737314' with navigation arrows and a REFRESH button. The main content area shows a list of documents, each with a unique identifier and a URL. The first document is a class page, and the subsequent documents are various game-related pages, including board games, mobile games, and game projects.

latestDB.tokenTable

DOCUMENTS 737.3k STORAGE SIZE 193.0MB AVG. SIZE 1.2KB INDEXES 1 TOTAL SIZE 18.4MB AVG. SIZE 18.4MB

Documents Aggregations Schema Explain Plan Indexes Validation

FILTER { field: 'value' } OPTIONS FIND RESET REFRESH

ADD DATA VIEW

Displaying documents 1 - 20 of 737314 REFRESH

```
{
  "_id": "class",
  "https://transformativeplay\\ics\\luci\\edu\\classes/": 17,
  "https://transformativeplay\\ics\\luci\\edu\\juliet-norton/": 2,
  "https://transformativeplay\\ics\\luci\\edu\\classes/ics-163-mobile-ubiquitous-games/": 20,
  "https://transformativeplay\\ics\\luci\\edu\\events/board-games-reclaimed/": 2,
  "https://transformativeplay\\ics\\luci\\edu\\aaron-trammell/": 2,
  "https://transformativeplay\\ics\\luci\\edu/": 2,
  "https://transformativeplay\\ics\\luci\\edu\\michael-a-cowling/": 2,
  "https://transformativeplay\\ics\\luci\\edu\\arvr-theater-syllabus/": 41,
  "https://transformativeplay\\ics\\luci\\edu\\research/publications/": 2,
  "https://transformativeplay\\ics\\luci\\edu\\transformative-costumed-play/": 2,
  "https://transformativeplay\\ics\\luci\\edu\\inf-190-games-from-concept-to-pitchable-prototype/": 29,
  "https://transformativeplay\\ics\\luci\\edu\\tangible-storytelling/": 2,
  "https://transformativeplay\\ics\\luci\\edu\\events/evoking-transformative-play-day-with-cds-middle-schoolers/": 2,
  "https://transformativeplay\\ics\\luci\\edu\\inf-242-winter-2017/": 25,
  "https://transformativeplay\\ics\\luci\\edu\\bonnie-bo-ruberg/": 2,
  "https://transformativeplay\\ics\\luci\\edu\\ics-169-a-b-capstone-game-project-2019-20/": 29,
  "https://transformativeplay\\ics\\luci\\edu\\research/": 2,
  "https://transformativeplay\\ics\\luci\\edu\\costumes-and-wearables-as-game-controllers-at-tei-2015/": 2,
  "https://transformativeplay\\ics\\luci\\edu\\matt-knutson/": 2,
  "https://transformativeplay\\ics\\luci\\edu\\costumes-wearables-as-game-controllers-workshop-at-uci/": 2,
  "https://transformativeplay\\ics\\luci\\edu\\classes/in4matx-295-digital-media-games/": 27,
  "https://transformativeplay\\ics\\luci\\edu\\geoffrey-c-bowker/": 2,
  "https://transformativeplay\\ics\\luci\\edu\\classes/ics-169-capstone/#schedule": 24,
  "https://transformativeplay\\ics\\luci\\edu\\jeffrey-bryan/": 2,
  "https://transformativeplay\\ics\\luci\\edu\\capstone18-19/": 34,
  "https://transformativeplay\\ics\\luci\\edu\\global-game-jam-2018/": 2,
  "https://transformativeplay\\ics\\luci\\edu\\playful-fab/": 2,
  "https://transformativeplay\\ics\\luci\\edu\\classes/ics-169-ab-capstone-game-project-fall-2017/": 17,
```