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