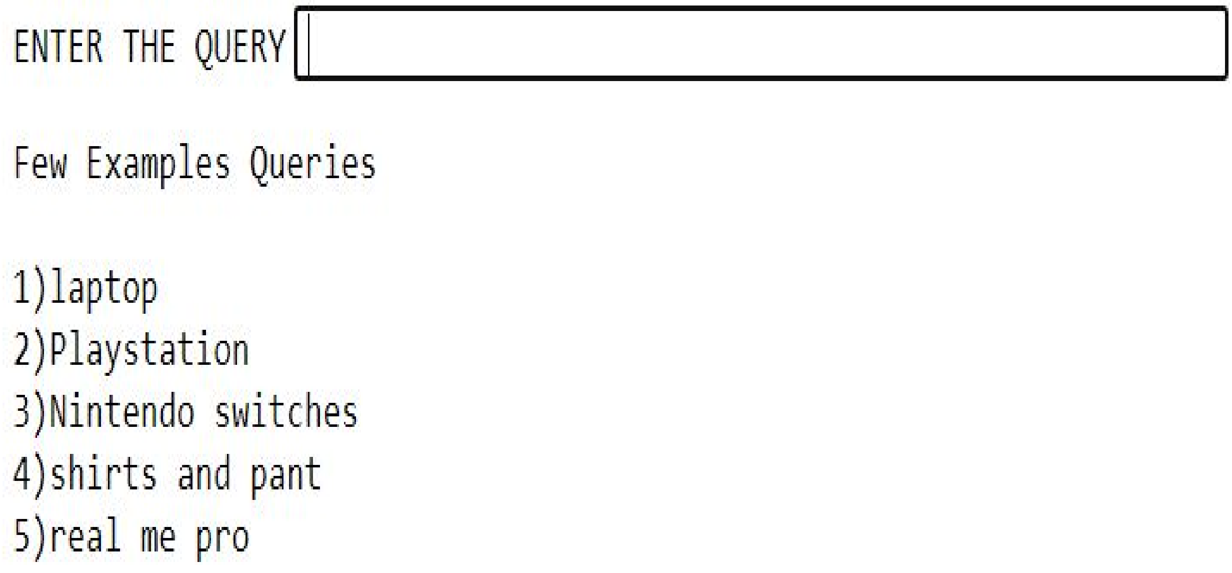
INFORMATION RETRIEVAL

PROJECT REPORT
SMART PRODUCT FILTER

Introduction

This project is mainly about searching for products based on title and recommending the best product on the basis of few factors. To achieve this, product information is crawled from the web(mostly from amazon.in). Based on the query given by the user the top K documents were retrieved using a ranked retrieval algorithm(Here K = 10 and ranked retrieval algorithm is VECTOR SPACE MODEL).Documents are retrieved based on Boolean retrieval algorithm.



```
ENTER THE QUERY

Few Examples Queries

1)laptop
2)Playstation
3)Nintendo switches
4)shirts and pant
5)real me pro
```

Collecting Data

- ❑ Dataset for this retrieval system is crawled from www.amazon.in using SELENIUM and Beautiful Soup as HTML Parser**.**
- ❑ Most searched keywords in the year 2020(132 keywords including searches in US and India) are taken as the keywords for searching the products.

- After getting into the product (PRODUCT TITLE,PRICE,RATINGS,REVIEW COUNT ORIGINAL PRODUCT URL) information is scraped and stored in local storage as text files.
- In total 48424 products are scraped and stored in 48424 different text files.
- Products which don't have ratings or reviews are taken as NULL.

SAMPLE DATA IN TEXT FILE -- (File name = '29289.txt')

RYLAN Multi-Purpose Laptop Desk for Study and Reading with Foldable Non-Slip Legs Reading Table Tray, Laptop Table, Laptop Stands, Laptop Desk, Foldable Study Laptop Table, Study Table (Black)

₹789

4.7 out of 5 stars

150

https://www.amazon.in//Multi-Purpose-Laptop-Reading-Foldable-Non-Slip/dp/B08KXSL9KY/ref=sr_1_45?dchild=1&keywords=computer+desk&qid=1608280549&sr=8-45

TASK-1

Index Creation for all the terms in titles in all the documents and retrieving the top K (K = 10) relevant documents.

- All the terms which are extracted are preprocessed using different preprocessing steps.
- By traversing all the documents an inverted index is created by extracting the terms in documents. (in this index document frequency and term frequency in document can be found.)
- Keys in an Inverted Index dictionary can be considered as the Vocabulary.
- Ranking documents can be done using the Vector Space Model.
- In the Vector Space Model cosine similarity is calculated between query and each document.

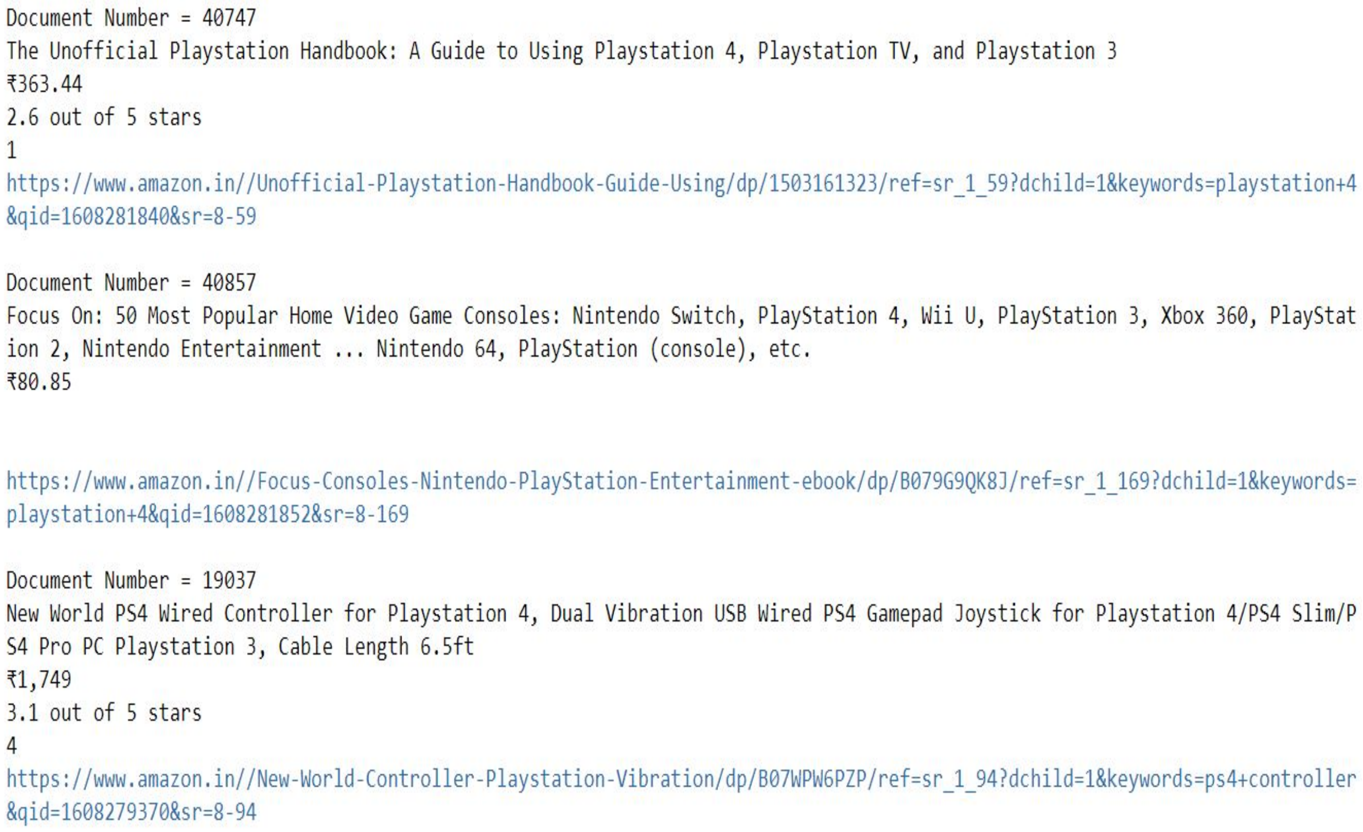
- Here tf-idf scores are considered as the coefficient of terms in the document vector.
  - **Tf-idf = (1+log(tf))*(log(size/df))**
  - For each term in Vocabulary tf-idf score is calculated.
- Terms in query are preprocessed and searched in inverted Index similar to Boolean retrieval accordingly cosine similarity between query and document is calculated.
  - **cosSIM(q,d) = (q.d)/(|q|*|d|)**
- After finding cosine similarity top 10 relevant documents are retrieved.
  - **Sample Query**

query = '**Playstation**'

Relevant Documents -> **['40747', '40857', '19037', '40788', '40839,**

'40854', '40936', '43223', '46360', '18939']

(<filename.txt> , filename = number)



```
Document Number = 40747
The Unofficial Playstation Handbook: A Guide to Using Playstation 4, Playstation TV, and Playstation 3
₹363.44
2.6 out of 5 stars
1
https://www.amazon.in//Unofficial-Playstation-Handbook-Guide-Using/dp/1503161323/ref=sr_1_59?dchild=1&keywords=playstation+4
&qid=1608281840&sr=8-59

Document Number = 40857
Focus On: 50 Most Popular Home Video Game Consoles: Nintendo Switch, PlayStation 4, Wii U, PlayStation 3, Xbox 360, PlayStat
ion 2, Nintendo Entertainment ... Nintendo 64, PlayStation (console), etc.
₹80.85


https://www.amazon.in//Focus-Consoles-Nintendo-PlayStation-Entertainment-ebook/dp/B079G9QK8J/ref=sr_1_169?dchild=1&keywords=
playstation+4&qid=1608281852&sr=8-169

Document Number = 19037
New World PS4 Wired Controller for Playstation 4, Dual Vibration USB Wired PS4 Gamepad Joystick for Playstation 4/PS4 Slim/P
S4 Pro PC Playstation 3, Cable Length 6.5ft
₹1,749
3.1 out of 5 stars
4
https://www.amazon.in//New-World-Controller-Playstation-Vibration/dp/B07WPW6PZP/ref=sr_1_94?dchild=1&keywords=ps4+controller
&qid=1608279370&sr=8-94
```

❑ Complexities :

- ❑ Inverted Index creating can be done in O(M*N) where M = Total No.of Documents and N is No.of terms in Vocabulary.
- ❑ Creation of inverted takes 10-15 minutes of time to complete all the iterations.
- ❑ Finding Document Frequency and finding Tf-idf score can be done in O(N)

  Where N is Vocabulary length.

- ❑ Searching results for a specific query takes 5-10 seconds.

TASK-2

In the retrieved documents recommending the best product based on few factors (based on rating and review count)

- ❑ After retrieving top 10 documents by extracting each document rating (out of 5) and review count (number of reviews) are extracted and stored in a dictionary with key as file name and values as ratings and review count.
- ❑ Here, a new factor is introduced which is,
  - ❑ Multiplication factor = x,

    where x is explicitly assigned value for a specific range of ratings. For example x = 1 if rating belongs to (4.5-5) or x = 0.9 if rating belongs to (4-4.5) ...

- ❑ Basis factor is calculated for each retrieved document
  - ❑ Basis factor = rating * review_count * multiplication_factor
- ❑ After finding Basis Factors for all retrieved documents the document which has maximum basis factor is selected as the best product.

KORE NITHISH KUMAR

S20180010086