**Web Analytics**

Assignment 2

**Question 1**

seed URL: <https://finance.google.com/finance>

|  |  |
| --- | --- |
| url  **Seed URL** | spider  **spider** |
| index  **Index** | sort  **Sort** |
| create search  **start service** | search  **search engine** |

**Question 2**

|  |
| --- |
| 1)How many webpages have you crawled? 212 |
| 212 |
| 2)How many unique terms have you indexed? 24099 |
| items |
| 3)How many mapping relations between term and document have you created? 152350 |
| pairs |

**Question 3**

I picked the ‘2’ term, ‘finance’, in the Terms.dat file. Its IDF is approximate to 1.1081.

IDF = log(the number of total pages/the number of pages that contain the term)



**Quesiton 4**

I Pick the 20th document in Items.dat file.

The most frequently appeared terms in that document is:’google’.

The term frequency for ‘google’ is 0.0357

|  |
| --- |
| mostfreq  **Solution 1: my own program** |
| q41  **Solution 2: [www.cleverstat.com](http://www.cleverstat.com)(The results are patially blocked due to some reasons)** |

**Question 5**

1. Domain: Shopping

2)Screenshots of searching ‘electronic toothbrush’ on 3 vertical search engines: Amazon, ebay and Jet respectively.

|  |
| --- |
| amazon |
| ebay |
| jet |

3)Comparison table for three search engines

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Collection Coverage | Search Precision | UI | Sort option |
| **Amazon** | Best | High | Simple, modern and attractive | Many options |
| **ebay** | Limited | Fair | Clean | Many options |
| **Jet** | Limited | Fair | Clean | Limited options available: only relevance and price |