**SIT772 Database and Information Retrieval**

**Assignment-2**

**Name:** Shantanu Gupta

**Student ID:** 218200234

**Q1:**

a)

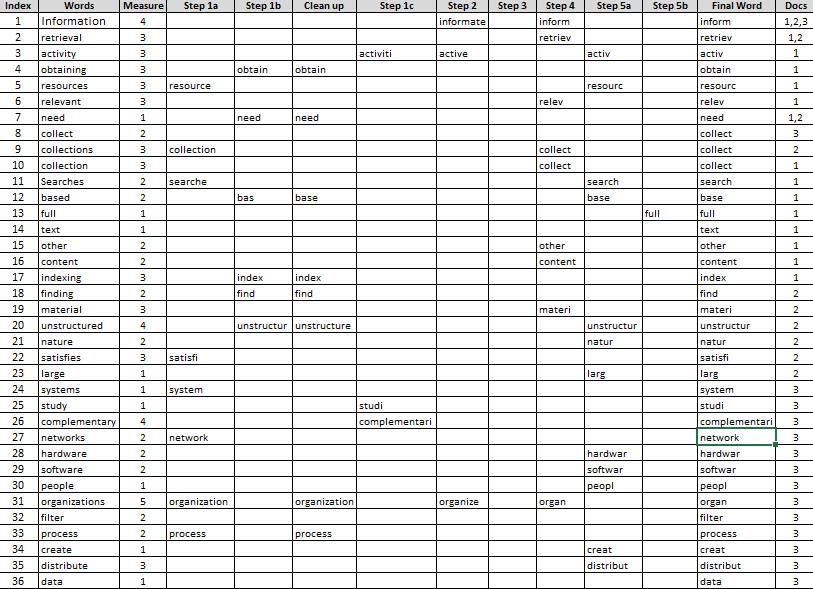
**Documents after removing the stop words are:**

Doc 1： Information retrieval activity obtaining information resources relevant information need collection information resources Searches based full text other content based indexing

Doc 2： Information retrieval finding material unstructured nature satisfies information need large collections

Doc 3： Information systems study complementary networks hardware software people organizations collect filter process create distribute data

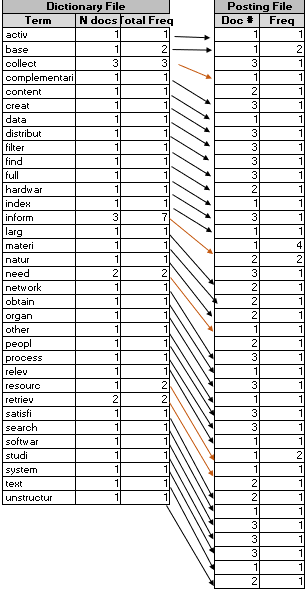
**Applying Porter stemming algorithm to the above mentioned documents**



b)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Step 1** | |  | **Step 2** | |  | **Step 3** | | |
| **Term** | **Doc #** |  | **Term** | **Doc #** |  | **Term** | **Doc #** | **Freq** |
| active | 1 |  | active | 1 |  | activ | 1 | 1 |
| base | 1 |  | base | 1 |  | base | 1 | 2 |
| base | 1 |  | base | 1 |  | collect | 3 | 1 |
| collect | 1 |  | collect | 1 |  | collect | 1 | 1 |
| content | 1 |  | collect | 2 |  | collect | 2 | 1 |
| full | 1 |  | collect | 3 |  | complementari | 3 | 1 |
| index | 1 |  | complementari | 3 |  | content | 1 | 1 |
| inform | 1 |  | content | 1 |  | creat | 3 | 1 |
| inform | 1 |  | create | 3 |  | data | 3 | 1 |
| inform | 1 |  | data | 3 |  | distribut | 3 | 1 |
| inform | 1 |  | distribut | 3 |  | filter | 3 | 1 |
| need | 1 |  | filter | 3 |  | find | 2 | 1 |
| obtain | 1 |  | find | 2 |  | full | 1 | 1 |
| other | 1 |  | full | 1 |  | hardwar | 3 | 1 |
| relev | 1 |  | hardwar | 3 |  | index | 1 | 1 |
| resourc | 1 |  | index | 1 |  | inform | 1 | 4 |
| resourc | 1 |  | inform | 1 |  | inform | 2 | 2 |
| retriev | 1 |  | inform | 2 |  | inform | 3 | 1 |
| search | 1 |  | inform | 3 |  | larg | 2 | 1 |
| text | 1 |  | larg | 2 |  | materi | 2 | 1 |
| collect | 2 |  | materi | 2 |  | natur | 2 | 1 |
| find | 2 |  | natur | 2 |  | need | 1 | 1 |
| inform | 2 |  | need | 1 |  | need | 2 | 1 |
| inform | 2 |  | need | 2 |  | network | 3 | 1 |
| larg | 2 |  | network | 3 |  | obtain | 1 | 1 |
| materi | 2 |  | obtain | 1 |  | organ | 3 | 1 |
| natur | 2 |  | organ | 3 |  | other | 1 | 1 |
| need | 2 |  | other | 1 |  | peopl | 3 | 1 |
| retriev | 2 |  | people | 3 |  | process | 3 | 1 |
| satisfy | 2 |  | process | 3 |  | relev | 1 | 1 |
| unstructur | 2 |  | relev | 1 |  | resourc | 1 | 2 |
| collect | 3 |  | resource | 1 |  | retriev | 1 | 1 |
| complementari | 3 |  | resource | 1 |  | retriev | 2 | 1 |
| create | 3 |  | retriev | 1 |  | satisfi | 2 | 1 |
| data | 3 |  | retriev | 2 |  | search | 1 | 1 |
| distribut | 3 |  | satisfy | 2 |  | softwar | 3 | 1 |
| filter | 3 |  | search | 1 |  | studi | 3 | 1 |
| hardwar | 3 |  | software | 3 |  | system | 3 | 1 |
| inform | 3 |  | studi | 3 |  | text | 1 | 1 |
| network | 3 |  | system | 3 |  | unstructur | 2 | 1 |
| organ | 3 |  | text | 1 |  |  |  |  |
| peopl | 3 |  | unstructur | 2 |  |  |  |  |
| process | 3 |  |  |  |  |  |  |  |
| softwar | 3 |  |  |  |  |  |  |  |
| studi | 3 |  |  |  |  |  |  |  |
| system | 3 |  |  |  |  |  |  |  |

c)



d)

**Information:** Information has been stemmed and reduced to its original form ‘inform’. The word information present in all three documents but with highest number of occurrence in Document 1. This means that first document is more relevant to our query than the other two documents.

|  |  |  |
| --- | --- | --- |
| **Term** | **Doc #** | **Frequency** |
| inform | 1 | 4 |
| inform | 2 | 2 |
| inform | 3 | 1 |

**System:** The word system has been stemmed and reduced to its original form ‘system’ and is present in document number 3.

|  |  |  |
| --- | --- | --- |
| **Term** | **Doc #** | **Frequency** |
| system | 3 | 1 |

**Index:** The word indexing has been stemmed and reduced to its original form ‘index’ and is present in only one document out of the three documents.

|  |  |  |
| --- | --- | --- |
| **Term** | **Doc #** | **Frequency** |
| index | 1 | 1 |

e)

**Query1: (Information and Systems) or Indexing**

The user is finding a document which provides with the words (information and systems) or indexing.

Information and Systems

D3

Information

D1

Information

D1

Information

D1

Systems

D3

Information and Systems

D3

Indexing

D1

Final Output

D1, D3

The document one and three are the output for the query. The final output is D1, D3 as the words are present in both the documents.

**Query2: Systems or Indexing**

Both words systems and indexing are present only in one document. The query requires to find the words in the documents present with us.

Systems

D3

Final Output

D1, D3

Indexing

D1

The final output is D1, D3 as the words are present in both the documents.

**Query3: Information and Systems**

The word information is present in all three documents and the word systems is present in the document number 3. Our query is to find a document which contains both information and systems in it.

Final Output D3

D3

Information

D1

Information

D3

Information

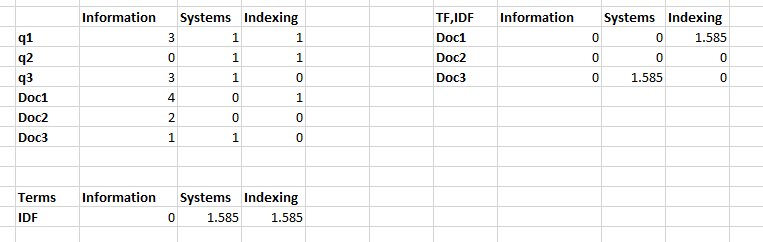
D2

Systems

D3

The final output is D3 which is document number 3 as the word information and system is present in these both documents.

f)



**Query1: (Information and Systems) or Indexing**

|  |  |
| --- | --- |
| **Sim(q1,Doc1)** | 0.43 |
| **Sim(q1,Doc2)** | 0.00 |
| **Sim(q1,Doc3)** | 0.43 |

From the cosine similarity present in the above table we can conclude that the most relevant document for the query 3 is document 1 and document 3.When compared with Boolean model we get the same documents (Doc 1 and Doc 3) as an output.

**Query2: Systems or Indexing**

|  |  |
| --- | --- |
| **Sim(q2,Doc1)** | 0.43 |
| **Sim(q2,Doc2)** | 0.00 |
| **Sim(q2,Doc3)** | 0.43 |

From the cosine similarity present in the above table we can conclude that the most relevant document for the query 2 is document 1 and document 3.When compared with Boolean model we get the same documents (Doc 1 and Doc 3) as an output.

**Query3: Information and Systems**

|  |  |
| --- | --- |
| **Sim(q3,Doc1)** | 0.00 |
| **Sim(q3,Doc2)** | 0.00 |
| **Sim(q3,Doc3)** | 0.279319 |

From the cosine similarity present in the above table we can conclude that the most relevant document for the query 1 is document 3.When compared with Boolean model we get the same document (Doc 3) as an output.

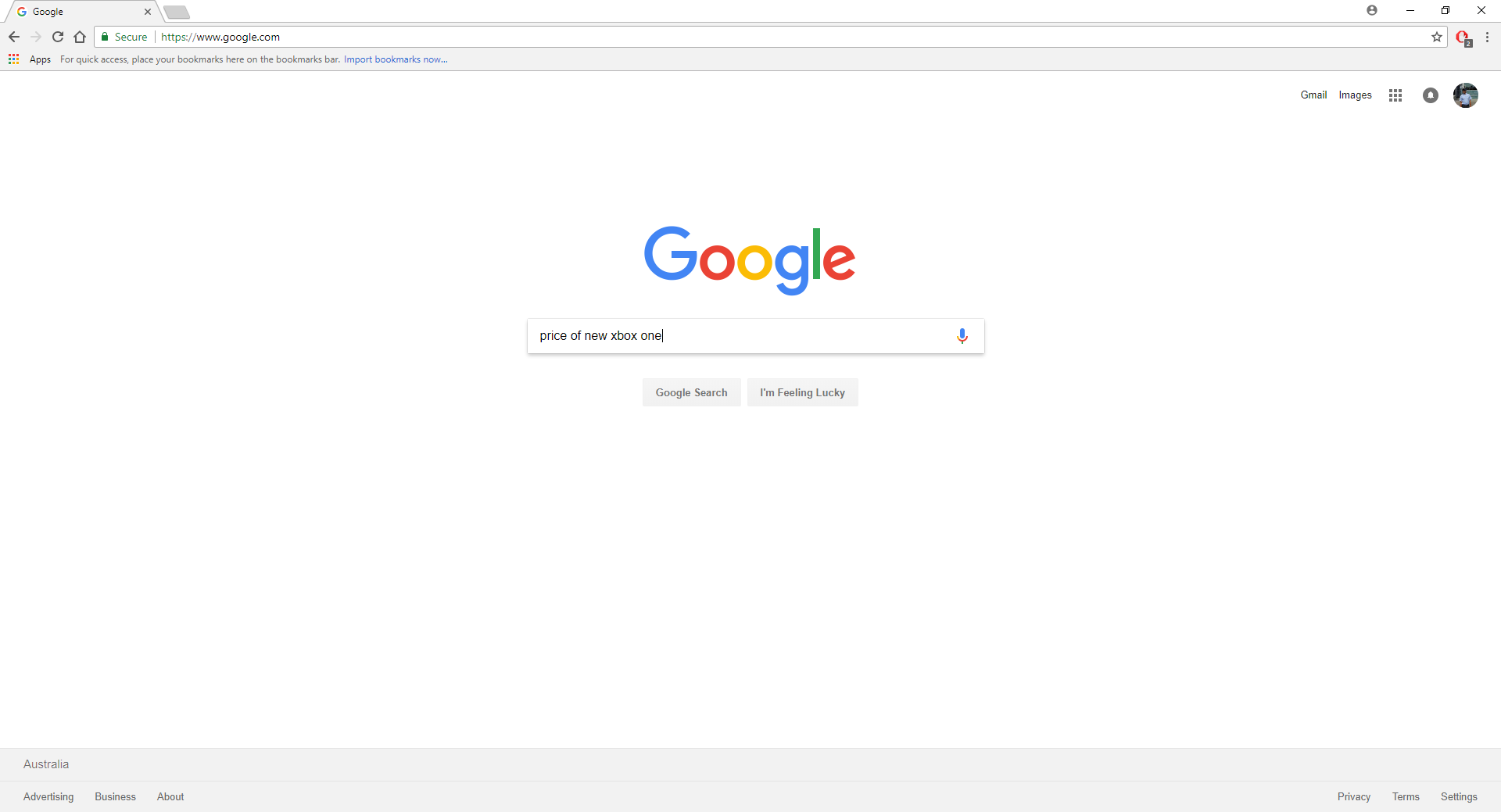
**Q2:**

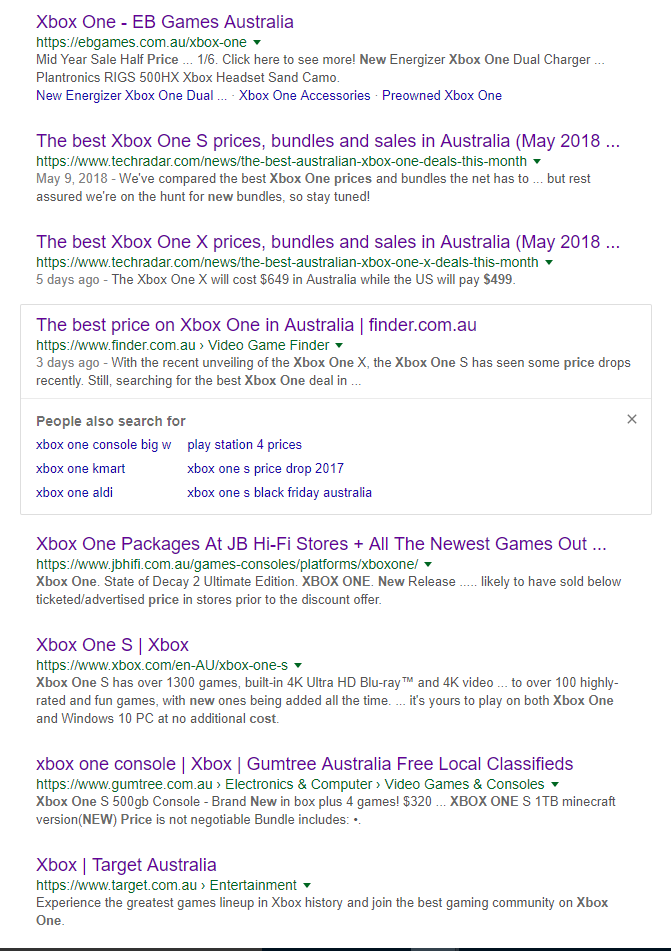
**Search Engine:** Google

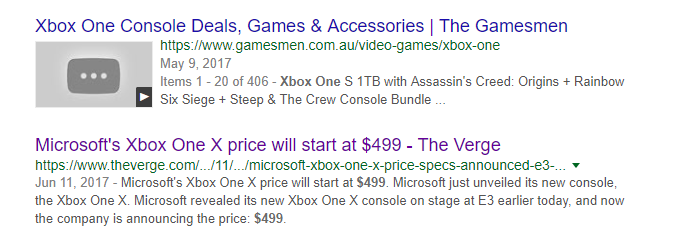
**Target 5:** obtain the price of new Xbox one.

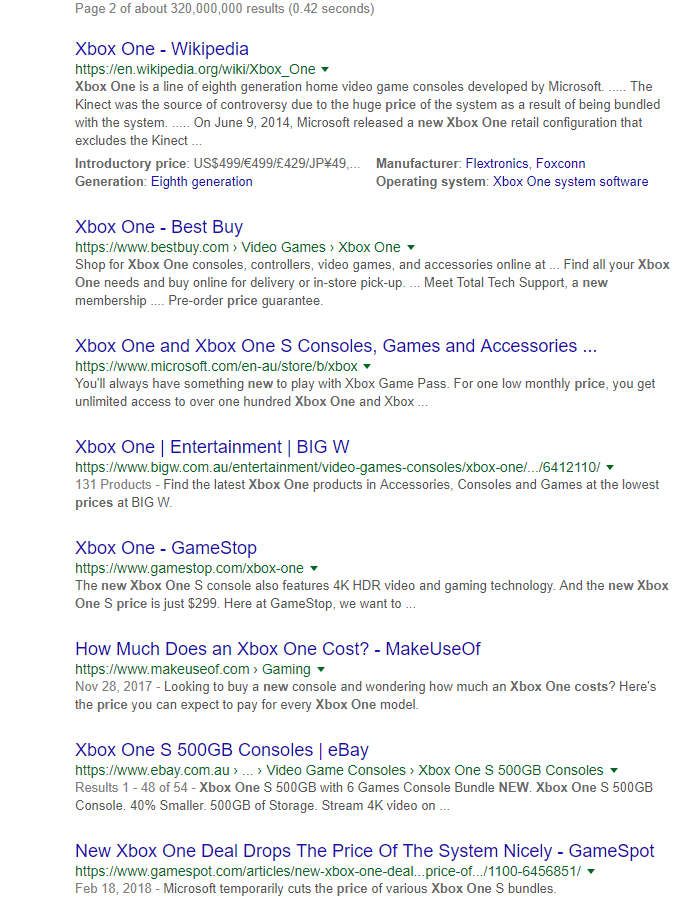
a)

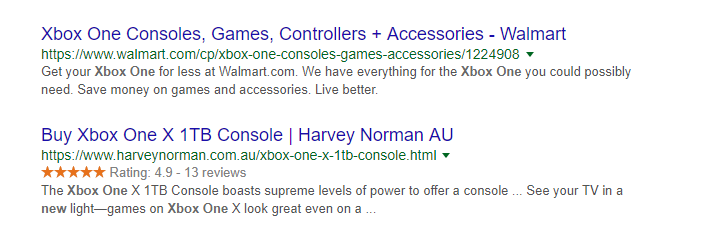
**Query1:** Price of new Xbox one

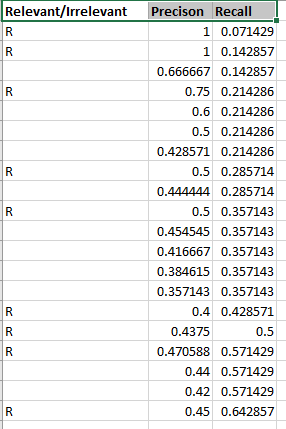
****



****

****

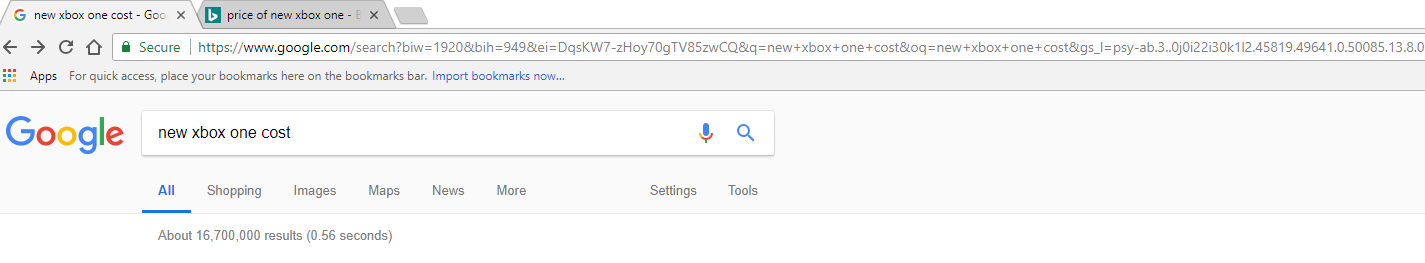
****

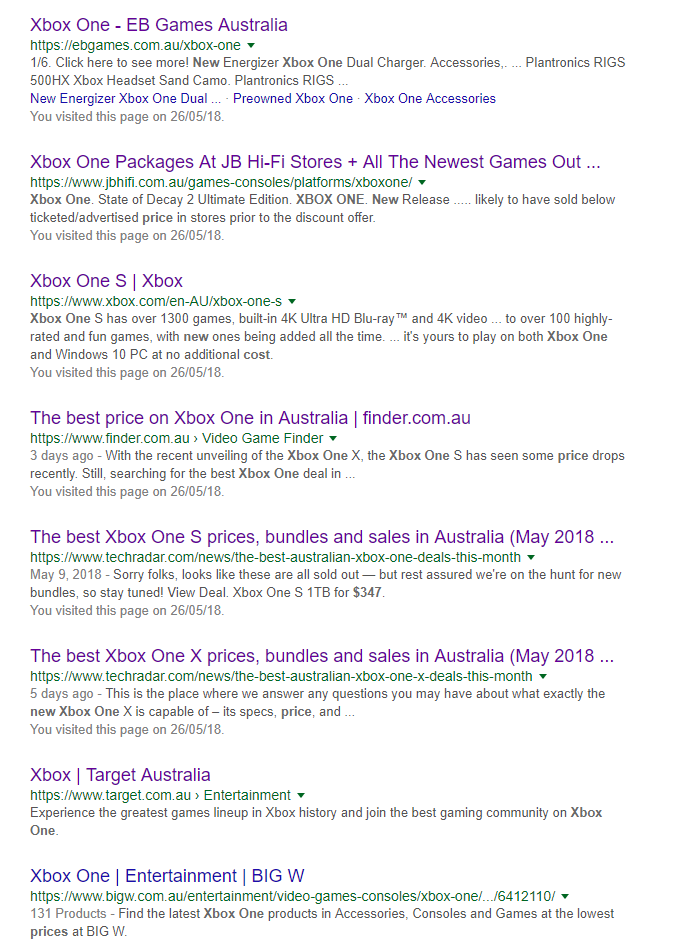


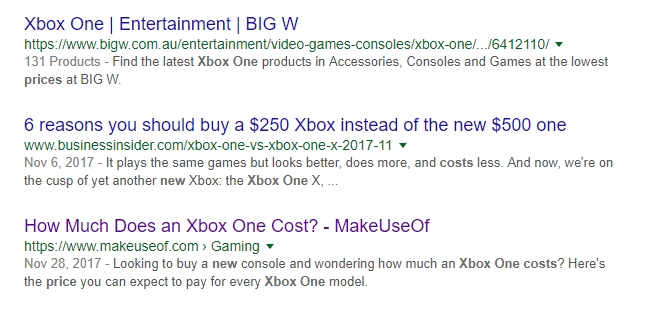
**Recall vs. Precision curve of Query1**

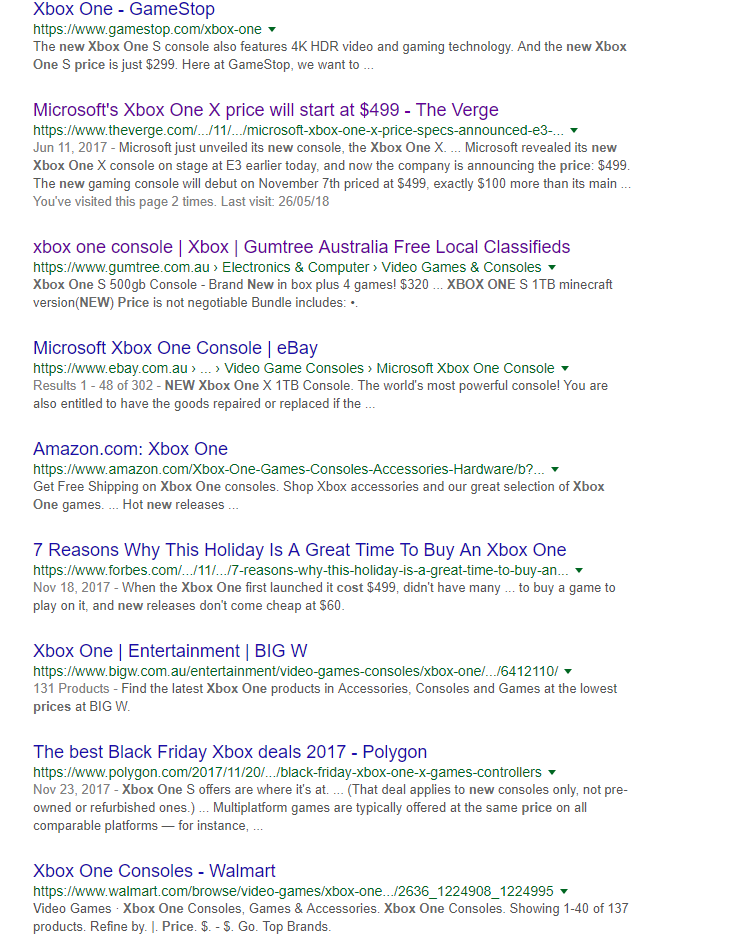
**Relevant Documents: 1, 2, 4, 8, 10, 15, 16, 17, 20**

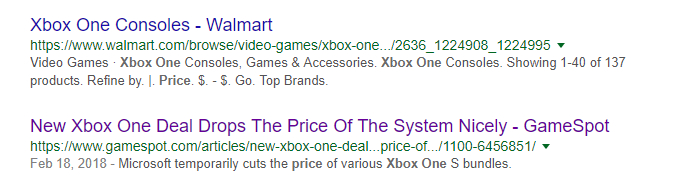
**Query2:** new Xbox one cost

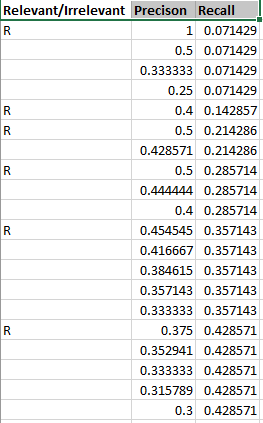


****

****

****

****



**Recall vs. Precision curve of Query2**

**Relevant Documents: 1, 5, 6, 8, 11, 16**

**Mean Average Precision for Query1:** 0.393435

**Mean Average Precision for Query2:** 0.230682

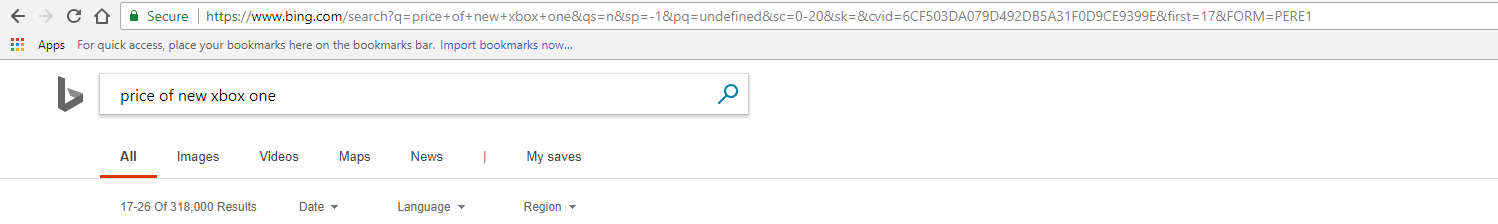
**Overall Mean Average Precision:** 0.312058

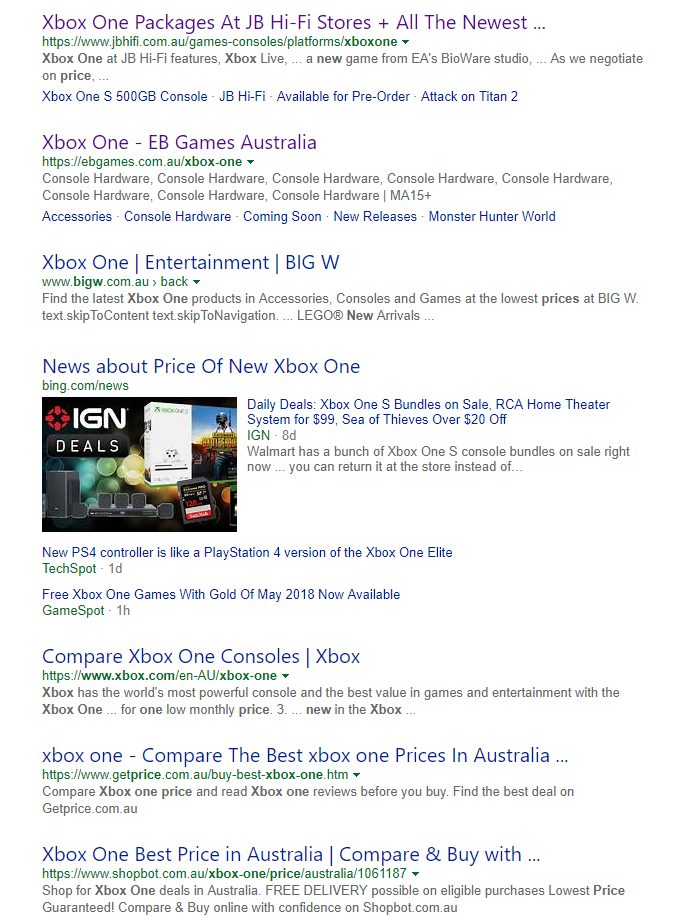
**Search Engine:** Bing

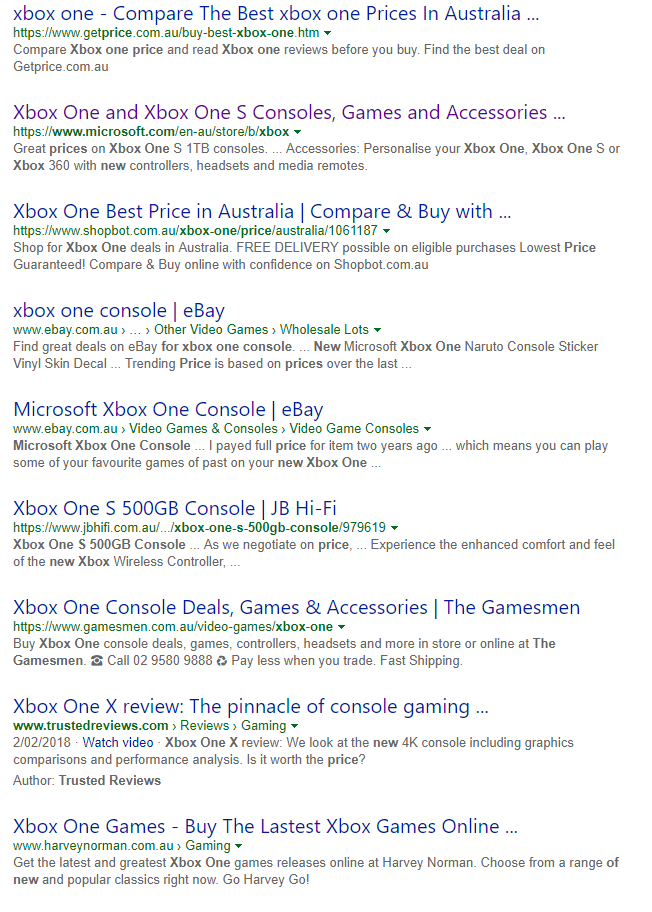
**Target 5:** obtain the price of new Xbox one.

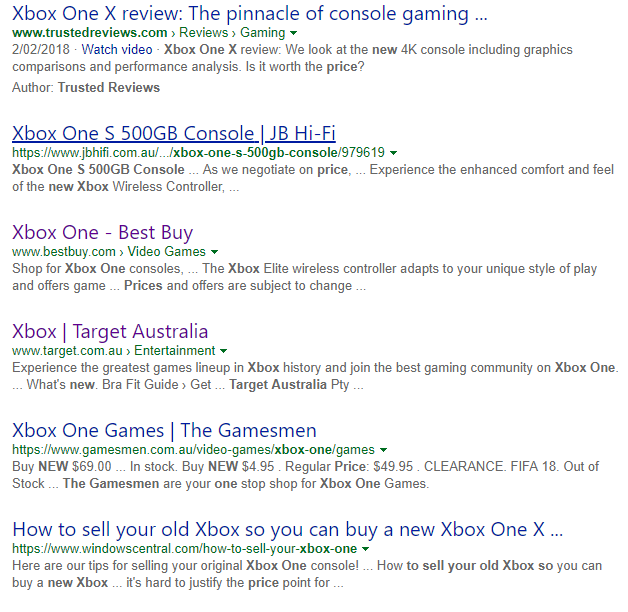
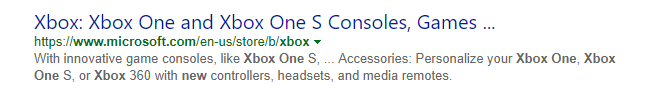
b)

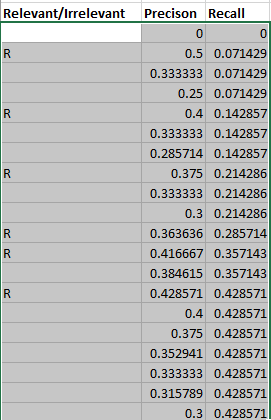
**Query1:** Price of new Xbox one

****

****

****

****

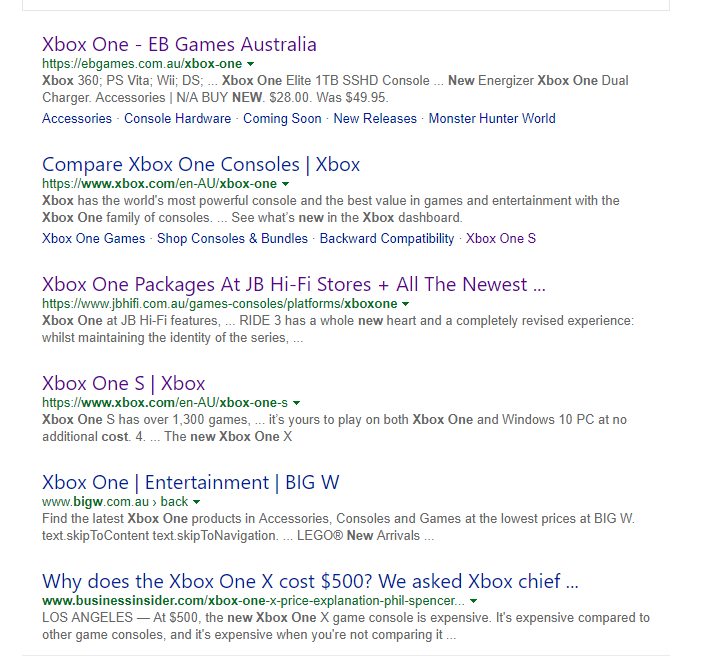


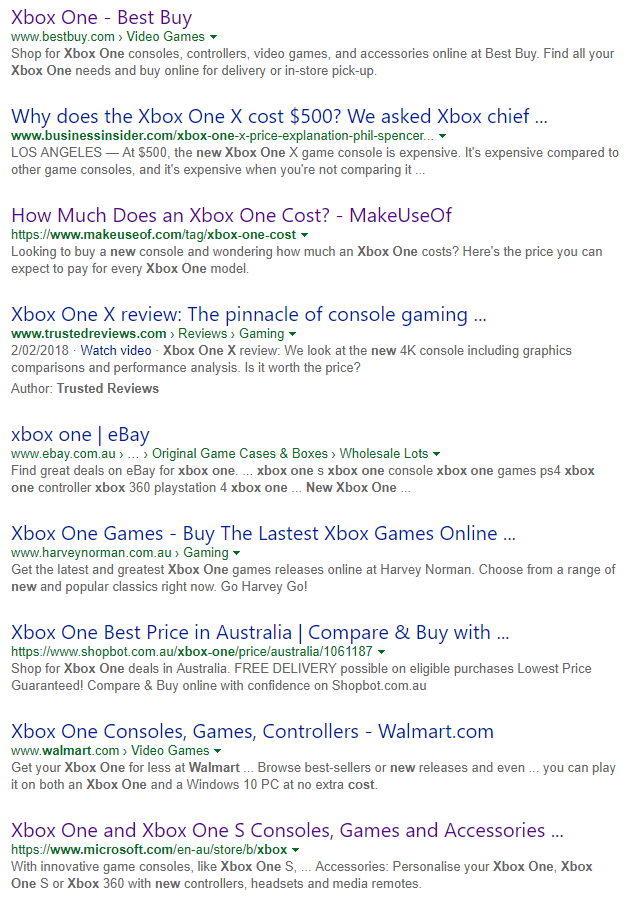
**Recall vs. Precision curve of Query1**

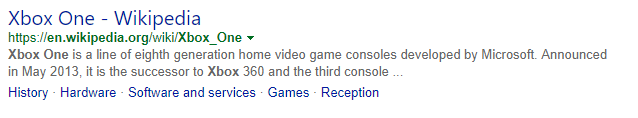
**Relevant Documents: 2, 5, 8, 11, 12, 14**

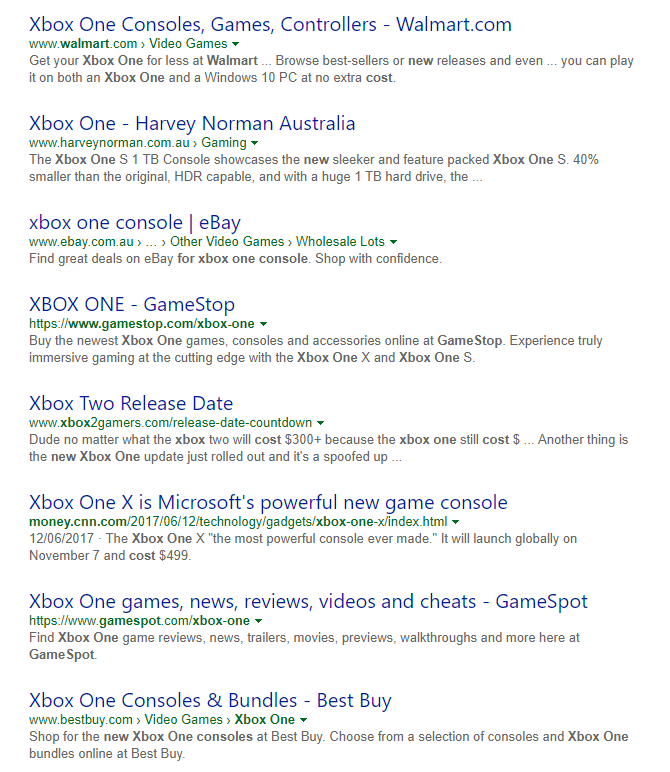
**Query2:** new Xbox one cost

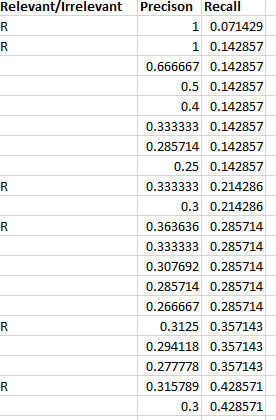


****

****

****

****



**Recall vs. Precision curve of Query2**

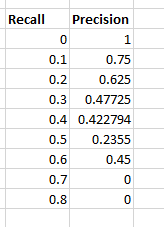
**Relevant Documents: 1, 2, 9,11,16,19**

**Mean Average Precision for Query1:** 0.17742

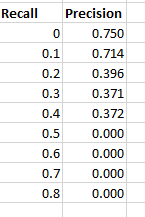
**Mean Average Precision for Query2:** 0.237518

**Overall Mean Average Precision:** 0.207469

c)



**Average for search engine- Google**



**Average for search engine- Bing**

When you see the graph you will find that blue line always has greater precision than orange line over all the values of the recall. Even at the 10% of the recall value both of them have similar value but Google is slightly higher precision. Same is the case over 40% recall value. Except this, all other recall values you will find Google is a winner.

Google search engine is better/superior than Bing in terms of getting relevant documents to the users.

**References:**

Future Learn Limited 2018, SIT772 Database & Information retrieval, Future Learn Limited, retrieved 15 April 2018, < [https://www.futurelearn.com/your-programs/database-and-information-retrieval/4](https://www.futurelearn.com/your-programs/database-and-information-retrieval/4%20) >