**Date:20/07/2021**

**Information Retrieval**

**Submitted by:**

Narmeen Ijaz

Roll No (20201009-056)

Department# MSCS(3rd-Semester)

**Submitted to:**

Dr. Shafiq Khan

**UNIVERSITY OF SIALKOT**

****

**Question:**

Consider 4 documents with one line of text.

1. Use an inverted index to index the documents
2. Construct a 4 words query
3. Apply VSM model, language model, and TF-IDF model for document retrieval

**Solution:**

**Consider 4 documents with one line of text.**

**Document 1:**

Computer system managed the computer resources.

**Document 2:**

Operating system is an interface between hardware and software.

**Document 3:**

Other application run in computer with the help of operating system.

**Document 4:**

Operating system manage computer memory and resources.

1. **Use an inverted index to index the documents**

**Remove stop words:**

An, and, between, help, is, in, of, other, the, with

**Distinct Terms:**

Application, computer, hardware, manage, memory, operating system, resources, run, software

**Stop words**

|  |  |  |
| --- | --- | --- |
| **ID** | **Term** | **Document** |
| **1** | Application | 3 |
| **2** | Computer | 1,3,4 |
| **3** | Hardware | 2 |
| **4** | Interface | 2 |
| **5** | Manage | 1,4 |
| **6** | Memory | 4 |
| **7** | Operating System | 1,2,3,4 |
| **8** | Resources | 1,4 |
| **9** | Run | 3 |
| **10** | Software | 2 |

**Inverted Index**

an

and

between

help

is

in

of

other

the

with

Operating system is an interface between hardware and software.

**Document 2**

Operating system i Other application run in computer with the help of operating system.

**Document 3**

Operating system manage computer memory and resources.

**Document 4**

Computer system managed the computer resources.

**Document 1**

1. **Construct a 4 words query**

**Query:**

**“**Operating System manage Computer Software”

1. **Apply VSM model, language model, and TF-IDF model for document retrieval**

**VSM Model**

**Query: “**Operating System manage Computer Software”

**Document 1:** Computer system managed the computer resources.

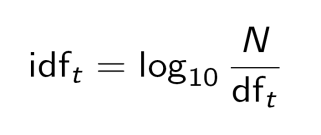
**Document 2:** Operating system is an interface between hardware and software.

**Document 3:** Other application run in computer with the help of operating system.

**Document 4:** Operating system manage computer memory and resources.

**Idf**

**Formula**



**IDF of Terms:**

|  |  |  |  |
| --- | --- | --- | --- |
| Idf(Operating System)=0 | Idf(is)=0.6020 | Idf(and)=0.3010 | Idf(in)=0.6020 |
| Idf(Manage)=0.3010 | Idf(an)=0.6020 | Idf(software)=0.6020 | Idf(with)=0.6020 |
| Idf(the)=0.3010 | Idf(interface)=0.6020 | Idf(other)=0.6020 | Idf(of)=0.6020 |
| Idf(computer)=0.1249 | Idf(between)=0.6020 | Idf(application)=0.6020 | Idf(memory)=0.6020 |
| Idf(resources)=0.3010 | Idf(hardware)=0.6020 | Idf(run)=0.6020 |

* Weight the items of vectors using tf-idf weighting scheme

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Doc ID** | **Operating System** | **Manage** | **the** | **Computer** | **Resources** | **Is** | **an** | **Interface** | **between** | **Hardware** |
| **D1** | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| **D2** | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 |
| **D3** | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| **D4** | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| **Q** | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Doc ID** | **and** | **Software** | **other** | **application** | **run** | **in** | **with** | **of** | **memory** |
| **D1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **D2** | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **D3** | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| **D4** | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| **Q** | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

**Weight of TF-IDF:**

**Formula: **

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Doc ID** | **Operating System** | **Manage** | **the** | **Computer** | **Resources** | **Is** | **an** | **in** | **Interface** |
| **D1** | 0 | 0.3010 | 0.3010 | 0.124 | 0.3010 | 0 | 0 | 0 | 0 |
| **D2** | 0 | 0 | 0 | 0 | 0 | 0.602 | 0.602 | 0 | 0.602 |
| **D3** | 0 | 0 | 0.3010 | 0.124 | 0 | 0 | 0 | 0.602 | 0 |
| **D4** | 0 | 0.3010 | 0 | 0.124 | 0.3010 | 0 | 0 | 0 | 0 |
| **Q** | 0 | 0.3010 | 0 | 0.124 | 0 | 0 | 0 | 0 | 0 |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Doc ID** | **between** | **Hardware** | **and** | **Software** | **other** | **application** | **run** | **with** | **of** | **memory** |
| **D1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **D2** | 0.602 | 0.602 | 0.3010 | 0.602 | 0 | 0 | 0 | 0 | 0 | 0 |
| **D3** | 0 | 0 | 0 | 0 | 0.602 | 0.602 | 0.602 | 0.602 | 0.602 | 0 |
| **D4** | 0 | 0 | 0.3010 | 0 | 0 | 0 | 0 | 0 | 0 | 0.602 |
| **Q** | 0 | 0 | 0 | 0.602 | 0 | 0 | 0 | 0 | 0 | 0 |

**Query:**

“Operating System manage Computer Software”

SC(q,D1) = Operating System + manage + computer + software

= (0)(0)+(0.3010)(0.3010)+(0.124)(0.124)+(0)(0.602)

= 0.1059

SC(q,D2) = Operating System + manage + computer + software

= (0)(0)+(0)(0.3010)+(0)(0.124)+(0.602)(0.602)

= 0.362

SC(q,D3) = Operating System + manage + computer + software

= (0)(0)+(0)(0.3010)+(0.124)(0.124)+(0)(0.602)

= 0.0153

SC(q,D4) = Operating System + manage + computer + software

= (0)(0)+(0.3010)(0.3010)+(0.124)(0.124)+(0)(0.602)

= 0.1059

**Ranking would be D2,D1,D4 and D3**

**Language Model**

**D1**: Computer system managed the computer resources.

**D2:** Operating system is an interface between hardware and software.

**D3:** Other application run in computer with the help of operating system.

**D4:** Operating system manage computer memory and resources.

**Query: “**Operating System manage Computer Software”

**Use mixture model with =1/2**

**P(D1\q) =** Operating System \* manage \* computer \* software

= [(1/5)0.5+(4/4)0.5] \* [(1/5)0.5+(2/4)0.5] \* [(1/5)0.5+(3/4)0.5] \* [(0/5)0.5+(1/4)0.5]

= (0.1+0.5)\*(0.1+0.25)\*(0.1+0.375)\*(0+0.125)

= 0.6\*0.26\*0.474\*0.125

**P(D1\q) =** 0.0092625

**P(D2\q) =** Operating System \* manage \* computer \* software

= [(1/8)0.5+(4/4)0.5] \* [(0/8)0.5+(2/4)0.5] \* [(0/8)0.5+(3/4)0.5] \* [(1/8)0.5+(1/4)0.5]

= (0.0625+0.5)\*(0+0.25)\*(0+0.375)\*(0.0625+0.125)

= 0.5625\*0.25\*0.375\*0.1875

**P(D2\q) =** 0.00988

**P(D3\q) =** Operating System \* manage \* computer \* software

= [(1/11)0.5+(4/4)0.5] \* [(0/11)0.5+(2/4)0.5] \* [(1/11)0.5+(3/4)0.5] \* [(0/11)0.5+(1/4)0.5]

= (0.04+0.5)\*(0+0.25)\*(0.04+0.375)\*(0+0.125)

= 0.54\*0.25\*0.415\*0.125

**P(D3\q) =** 0.0070

**P(D4\q) =** Operating System \* manage \* computer \* software

= [(1/6)0.5+(4/4)0.5] \* [(1/6)0.5+(2/4)0.5] \* [(1/6)0.5+(3/4)0.5] \* [(0/6)0.5+(1/4)0.5]

= (0.08+0.5)\*(0.08+0.25)\*(0.08+0.375)\*(0+0.125)

= 0.58\*0.33\*0.455\*0.125

**P(D4\q) =** 0.010

**Ranking= D4,D2,D1,D3**

**TF-IDF Model**

**D1**: Computer system managed the computer resources.

**D2:** Operating system is an interface between hardware and software.

**D3:** Other application run in computer with the help of operating system.

**D4:** Operating system manage computer memory and resources.

**Query: “**Operating System manage Computer Software”

**IDF of Terms:**

|  |  |  |  |
| --- | --- | --- | --- |
| Idf(Operating System)=0 | Idf(is)=0.6020 | Idf(and)=0.3010 | Idf(in)=0.6020 |
| Idf(Manage)=0.3010 | Idf(an)=0.6020 | Idf(software)=0.6020 | Idf(with)=0.6020 |
| Idf(the)=0.3010 | Idf(interface)=0.6020 | Idf(other)=0.6020 | Idf(of)=0.6020 |
| Idf(computer)=0.1249 | Idf(between)=0.6020 | Idf(application)=0.6020 | Idf(memory)=0.6020 |
| Idf(resources)=0.3010 | Idf(hardware)=0.6020 | Idf(run)=0.6020 |

**Term Frequency:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Doc ID** | **Operating System** | | **Manage** | | **the** | | **Computer** | | **Resources** | | **Is** | | **an** | **between** | | | **Hardware** | |
| **D1** | 1 | | 1 | | 1 | | 1 | | 1 | | 0 | | 0 | 0 | | | 0 | |
| **D2** | 1 | | 0 | | 0 | | 0 | | 0 | | 1 | | 1 | 1 | | | 1 | |
| **D3** | 1 | | 0 | | 1 | | 1 | | 0 | | 0 | | 0 | 0 | | | 0 | |
| **D4** | 1 | | 1 | | 0 | | 1 | | 1 | | 0 | | 0 | 0 | | | 0 | |
| **Q** | 1 | | 1 | | 0 | | 1 | | 0 | | 0 | | 0 | 0 | | | 0 | |
| **Doc ID** | **and** | **Software** | | **other** | | **application** | | **run** | | **in** | | **with** | | | **of** | **memory** | |
| **D1** | 0 | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | | 0 | 0 | |
| **D2** | 1 | 1 | | 0 | | 0 | | 0 | | 0 | | 0 | | | 0 | 0 | |
| **D3** | 0 | 0 | | 1 | | 1 | | 1 | | 1 | | 1 | | | 1 | 0 | |
| **D4** | 1 | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | | 0 | 1 | |
| **Q** | 0 | 1 | | 0 | | 0 | | 0 | | 0 | | 0 | | | 0 | 1 | |

**Weight of TF-IDF:**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Doc ID** | **Operating System** | **Manage** | **the** | **Computer** | **Resources** | **Is** | **an** | **in** | **Interface** |
| **D1** | 0 | 0.3010 | 0.3010 | 0.124 | 0.3010 | 0 | 0 | 0 | 0 |
| **D2** | 0 | 0 | 0 | 0 | 0 | 0.602 | 0.602 | 0 | 0.602 |
| **D3** | 0 | 0 | 0.3010 | 0.124 | 0 | 0 | 0 | 0.602 | 0 |
| **D4** | 0 | 0.3010 | 0 | 0.124 | 0.3010 | 0 | 0 | 0 | 0 |
| **Q** | 0 | 0.3010 | 0 | 0.124 | 0 | 0 | 0 | 0 | 0 |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Doc ID** | **between** | **Hardware** | **and** | **Software** | **other** | **application** | **run** | **with** | **of** | **memory** |
| **D1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **D2** | 0.602 | 0.602 | 0.3010 | 0.602 | 0 | 0 | 0 | 0 | 0 | 0 |
| **D3** | 0 | 0 | 0 | 0 | 0.602 | 0.602 | 0.602 | 0.602 | 0.602 | 0 |
| **D4** | 0 | 0 | 0.3010 | 0 | 0 | 0 | 0 | 0 | 0 | 0.602 |
| **Q** | 0 | 0 | 0 | 0.602 | 0 | 0 | 0 | 0 | 0 | 0 |

**Query:** Operating System + manage + Computer + Software

Cos(q,d1) =SIM(q,d1) = 0 + (0.3010/4\*0.3010/5) + (0.124/4\*0.124/5) + 0 = 0.00526

Cos (q,d2) =SIM(q,d2) = 0 + (0.3010/4\*0.3010/8) + 0 + (0.602/4\*0.602/8) = 0.0141

Cos (q,d3) =SIM(q,d3) = 0 + 0 + (0.124/4\*0.124/10) + 0 = 0.00038

Cos (q,d4) =SIM(q,d4) = 0 + (0.3010/4\*0.3010/6) + (0.124/4\*0.124/6) + 0 =0.00437

**Ranking would be D2,D1,D4 and D3**