**Simple Search Engine in Spark**



**Made by:**

**Sobirdzhon Bobiev (s.bobiev@innopolis.university)**

**Beni Zakaria Ar Ridlo (b.ridho@innopolis.university)**

**Utih Amartiwi (u.amartiwi@innopolis.university)**

**Big Data Technology and Analysis Class 2019**

**CHAPTER I**

**INTRODUCTION**

1. **Background**

Since it was built in 2009, Apache Spark has emerged as the next generation big data processing engine [1].It is faster, highly accessible, and can integrate with other Big Data tools. In this project, we want to build a simple search engine using Spark. A fundamental goal of search engine is to identify some relevant documents. As number of files grows significantly, potentially relevant document identified by search engine also increase. Search engine can store some documents even it only contains 1-3 words that similar with user’s keywords. Thus, the users have to check again whether the information is really they need or not. It is difficult and takes much time to check it one by one.

In order to improve the relevancy of search result, we need to make an index of the similarity of the documents and rank them. The most common index used is TF/IDF weight that using frequency of term in indexing. The indexing result will be used to calculate the rank of similarity. There are some of rankers, but 2 most popular rankers are Inner Product and Okapi BM25. Inner product ranker comes from vector space theory that identifies the distance between vectors as the similarity of files. On the other hand, Okapi BM25 computes the similarity using weight between document and query based on term probability [2]. In some cases, inner product shows better result than BM25 [3] but in other cases are vice versa [2]. That motivates us to learn more the difference result of both ranker. Therefore, we will implement both rankers and find the best ranker for our search engine.

1. **Objective**

The aim of this project is to learn how to build a good search engine. Here we will implement big data theory from lecture and the programming skill using Spark. Furthermore, to provide the user the right information they need, we have to use the best ranker. Therefore, we will compare the result of inner product and BM25 using Mean Average Precision (MAP) as the metrics. We will analyze and determine the best one of them in our case. All in all, we hope we can get new insight from problem and result during the process and share it to the reader.

**CHAPTER II**

**PROCESS**

1. **Methods and Steps**
2. Indexing Engine

First, all the words from text corpus will be enumerated in Vocabulary. For each term, we give a unique ID that represents it. In each document *d*, we calculate the frequency of occurrence of term *t* in document *d* that called **Term-Frequency (TF)**. We also calculate the number of documentsthat contain term *t*. It is called **Inverse Document Frequency (IDF).** IDF(t) also can be calculated:

Where N is number of documents and n(t) the number of documents containing the term *t*. We calculate document weight (**Wd**) by:

Then we transform all terms in each document as format: (,).

1. Ranking Engine

After we do indexing the documents, we process the query. Here we calculate the weight of query (**Wq**) using formula above. Then we create a vectorized representation for the query with format: (,). Suppose is a term in query and is a term in document *d,* if we calculate the relevance analysis:

1. Inner Product

Suppose the document and query to be vectors **d** and **q** respectively. The inner product computes the distance between vectors:

is the relevance of document **d** with respect to query **q**.

1. Okapi BM25

Okapi BM25 computes a weight between document and query based on term probability.

and are assumed as same as TF and IDF in indexing engine. **Avgdl** is the average document length in the corpus, **|d|** is the size of the current document, **b** and **k1** are free parameters that will be set to 0.75 and 2.0 respectively.

1. Query Response

If , it means the document doesn’t contains the same term with query so it will not appear in Query Response. If it means the document has the relevant term with query. Query response will list some relevant documents and appear in order based on their rank.

1. Testing and Evaluating

Precision is the proportion of search result is relevant to the user. Mean Average Precision (MAP) for a set of queries is the mean of the average precision scores for each query.

Where **Nrel** is the number of relevant document returned by the ranker, **P(k)** is the precision calculated for the positions from 1 to k, **rel(k)** is an indicator function that is equal 1 when document at position k is a relevant document, and 0 otherwise. Here we calculate MAP of the result from both inner product ranker and BM25 ranker. We compare it and analyze whether their difference is significant or not and which one better than other using *t-test* [4].

1. **Chart of Steps**

|  |
| --- |
| **Indexing Engine**  Query  Text Corpus  Word Enumeration  ID(t)  Vocabulary  IDF(t)  Document Count  Indexer  ID(t)  Wd(t)  **Ranker Engine**  Relevance Analizator  Query Vectorizer  Wq(t)  If ID(td) = ID(tq)  doc ids  R(q,d)  Content Extractor  doc content  Query Response |

Chart 1. Vector Space Model with Map Reduce

1. **Dataset Description**

The dataset is from English Wikipedia Dump. We got it by accessing Hadoop cluster with HDFS that contain folder EnWikiMedium. In EnWikiMedium, there are many files that contain multiple JSON strings, each string corresponding to one Wikipedia article.

1. **Tasks Distribution**

|  |  |
| --- | --- |
| **Member’s Name** | **Tasks** |
| Sobirdzhon Bobiev | Implement indexer for naive search engine  Implement two rankers: inner product and BM25 |
| Beni Zakaria Ar Ridlo | Testing and comparing two rankers using MAP |
| Utih Amartiwi | Writing the report |

Repository: https://github.com/sobir-git/Search-Engine

**CHAPTER III**

**DISCUSSION OF RESULT**

In order to evaluate the result, here we use 10 same queries to input in search engine and compare between Inner product and BM25 ranker. The words are : ‘big data’, ‘game of thrones’, ‘alice in wonderland’, ‘shakespeare’, ‘data science’, ‘university’, ‘bali’, ‘black hole’, ‘westlife’, and ‘shah rukh khan’.

To compare the result statistically, we use t-test with initial hypothesis that the result of both ranker is almost same.

H0 : The Precision of Inner product = The Precision of BM25

H1 : The Precision of Inner product The Precision of BM25

|  |  |  |
| --- | --- | --- |
| Ranker | MAP | T-test |
| Inner Product | 0.80727 | T-statistics = -0.79161 P-Value = 0.43889 |
| BM25 | 0.874469 |

P-Value > 0.05, it means we accept H0. So, statistically, there is no significant different result of Inner Product ranker and BM25 ranker. Both ranker provide the high number of MAP that more than 80%. It means more than 80% of our prediction is correct. However, if we look at the MAP result, BM25 is a bit bigger than the Inner product. If we compare the search result, we think that BM25 result is more relevant than inner product. The full search result is provided in Appendix. All in all, in this case, the search result of BM25 is better than Inner product.

**CHAPTER IV**

**CONCLUSION**

1. Spark makes data processing easy using its data structure, like RDD, dataframes, and datasets.
2. Both Inner product and BM25 are suitable for this search engine. Statistically, there is no significant different from both result. But, if we compare the MAP, BM25 is better than the inner product.
3. Spark can be a tool for us to learn the big idea of search engine.

**APPENDIX**

**Inner Product**

Query results for:  Big Data

     DocID Title                                                        Relevance

     78358 Big Ten Conference                                           0.000010

     78024 Big East Conference (1979–2013)                              0.000007

      7990 Data warehouse                                               0.000005

      4391 Big Brother (franchise)                                      0.000005

      4116 Big Bang                                                     0.000005

     42253 Data mining                                                  0.000005

      8377 Database                                                     0.000004

     12398 Geographic information system                                0.000004

     78362 Big 12 Conference                                            0.000004

     75691 Betting in poker                                             0.000004

      4472 The Big O                                                    0.000003

     32073 USB                                                          0.000003

     47676 Data (Star Trek)                                             0.000003

     16245 Janis Joplin                                                 0.000003

     30538 Transmission Control Protocol                                0.000003

      8013 Data compression                                             0.000003

     18167 Linked list                                                  0.000003

     61419 Tokenization (data security)                                 0.000003

     44578 Big O notation                                               0.000003

      6799 COBOL                                                        0.000003

Query results for: Game of thrones

     DocID Title                                                        Relevance

     12301 A Song of Ice and Fire                                       0.001283

     12302 World of A Song of Ice and Fire                              0.000670

     33604 War of the Spanish Succession                                0.000502

     12300 George R. R. Martin                                          0.000335

     57550 Charles Edward Stuart                                        0.000279

     74667 Crown jewels                                                 0.000168

     44705 House of Bourbon                                             0.000167

     46852 George I of Great Britain                                    0.000167

     15925 Jacobitism                                                   0.000167

     77740 Salic law                                                    0.000167

     61987 James Francis Edward Stuart                                  0.000167

     21148 Netherlands                                                  0.000112

     13212 History of Europe                                            0.000112

     62583 History of Italy                                             0.000112

     52266 War of the Austrian Succession                               0.000112

     71920 House of Orange-Nassau                                       0.000112

     55116 Dubrovnik                                                    0.000112

     76496 Max von Sydow                                                0.000112

     50213 Morganatic marriage                                          0.000112

     70760 Union Jack                                                   0.000112

Query results for: Alice in Wonderland

     DocID Title                                                        Relevance

     62446 Alice (Alice's Adventures in Wonderland)                     0.002230

     52853 Alice's Adventures in Wonderland                             0.001371

     31623 Through the Looking-Glass                                    0.001287

     15612 John Tenniel                                                 0.000473

     15907 Jabberwocky                                                  0.000347

     31501 Tim Burton                                                   0.000336

     50670 Marcy Playground                                             0.000320

     62180 Alice in Wonderland (1933 film)                              0.000278

     71870 Johnny Depp                                                  0.000274

     80335 Helena Bonham Carter                                         0.000270

     32917 Walt Disney                                                  0.000267

      1500 Dodo (Alice's Adventures in Wonderland)                      0.000223

     37398 The Walt Disney Company                                      0.000201

     52856 Alice in Wonderland (disambiguation)                         0.000199

     54436 The Wonderful Wizard of Oz                                   0.000197

     61141 Song of the South                                            0.000197

     52847 Children's literature                                        0.000195

     50214 Alice in Chains                                              0.000193

     63234 Grampians National Park                                      0.000192

     74534 Alice Comedies                                               0.000162

Query results for: Shakespeare

     DocID Title                                                        Relevance

      7712 Continuation War                                             0.000454

     11772 Finnish Civil War                                            0.000407

     33924 Winter War                                                   0.000404

     10577 Finland                                                      0.000388

     14378 History of Finland                                           0.000306

     50434 Finnish grammar                                              0.000202

     16777 Kalevala                                                     0.000139

     13696 Helsinki                                                     0.000136

     50432 Colloquial Finnish                                           0.000136

     10715 Finnish Defence Forces                                       0.000126

     50433 Finnish phonology                                            0.000120

     27122 Sami people                                                  0.000104

     25071 Paavo Nurmi                                                  0.000091

     10710 Demographics of Finland                                      0.000088

     49644 Jean Sibelius                                                0.000085

     75649 Sauna                                                        0.000085

     48672 Turku                                                        0.000079

     22618 Operation Barbarossa                                         0.000076

     10712 Economy of Finland                                           0.000060

     32048 Urho Kekkonen                                                0.000057

Query results for: Data science

     DocID Title                                                        Relevance

     26787 Science fiction                                              0.000009

     29266 Relationship between religion and science                    0.000007

      7990 Data warehouse                                               0.000005

     42253 Data mining                                                  0.000005

     26700 Science                                                      0.000005

     12398 Geographic information system                                0.000004

      8377 Database                                                     0.000004

     37010 Philosophy of science                                        0.000004

      7683 Creation science                                             0.000004

     32073 USB                                                          0.000003

     14400 History of science                                           0.000003

     15295 Intelligent design                                           0.000003

     47676 Data (Star Trek)                                             0.000003

     11257 Feminist science fiction                                     0.000003

     14573 Isaac Asimov                                                 0.000003

      5323 Computer science                                             0.000003

     30538 Transmission Control Protocol                                0.000003

      8013 Data compression                                             0.000003

     18167 Linked list                                                  0.000003

     26685 Statistics                                                   0.000003

Query results for: University

     DocID Title                                                        Relevance

     32127 University of Chicago                                        0.000003

     61114 Boston University                                            0.000003

     60611 University of Florida                                        0.000003

     60919 University of London                                         0.000002

     73276 University of Waterloo                                       0.000002

     32031 University of Texas at Austin                                0.000002

     39569 University of Glasgow                                        0.000002

     59801 University of Virginia                                       0.000002

      6310 Columbia University                                          0.000002

     31797 University of Oxford                                         0.000002

     32098 University of Victoria                                       0.000002

     42730 Emory University                                             0.000002

      5689 College                                                      0.000002

     66772 Rector (academia)                                            0.000002

     52029 University College London                                    0.000002

     32015 University of Arizona                                        0.000002

     31740 University of Michigan                                       0.000002

     19735 Macquarie University                                         0.000002

     58920 Washington University in St. Louis                           0.000002

     31801 Uppsala University                                           0.000002

Query results for: Bali

     DocID Title                                                        Relevance

      4147 Bali                                                         0.012449

     14579 Indonesia                                                    0.002716

     18360 Lombok                                                       0.002037

     14643 History of Indonesia                                         0.000905

     14644 Geography of Indonesia                                       0.000679

     30075 Tiger                                                        0.000679

     69336 Java                                                         0.000453

     54920 Cockfight                                                    0.000340

     14649 Transport in Indonesia                                       0.000340

     44245 Ramayana                                                     0.000340

     28741 Southeast Asia                                               0.000340

     31898 United Nations Framework Convention on Climate Change        0.000340

     33873 World Trade Organization                                     0.000340

     53831 Missionary                                                   0.000340

     76398 Trance and Dance in Bali                                     0.000340

     52828 Lindsay Davenport                                            0.000340

     13543 Hinduism                                                     0.000340

     20201 Musical notation                                             0.000226

     32454 Geography of Venezuela                                       0.000226

     14604 Foreign relations of India                                   0.000226

Query results for: Black Hole

     DocID Title                                                        Relevance

     64065 National Register of Historic Places                         0.000005

     23053 Periodic table                                               0.000004

     19692 Monopoly (game)                                              0.000003

      3870 Bundle theory                                                0.000003

      5320 Carbon nanotube                                              0.000003

     19622 Materials science                                            0.000003

       666 Alkali metal                                                 0.000003

     23479 Physicalism                                                  0.000003

     37436 Emergence                                                    0.000003

      5387 Condensed matter physics                                     0.000002

      5370 Category of being                                            0.000002

     37738 Soil                                                         0.000002

     23001 Polymer                                                      0.000002

      9476 Electron                                                     0.000002

     27568 Substance theory                                             0.000002

     62903 Complex systems                                              0.000001

     28305 String theory                                                0.000001

     19053 Mineral                                                      0.000001

     21058 Motel                                                        0.000001

     25453 Rheology                                                     0.000001

Query results for: WESTLIFE

     DocID Title                                                        Relevance

     80356 History of Vietnam                                           1.000000

     29544 Scientific revolution                                        0.000000

     29549 Self-replication                                             0.000000

     29545 Salian dynasty                                               0.000000

     29556 List of people from Sardinia                                 0.000000

     29550 Shmuel Yosef Agnon                                           0.000000

     29551 Steve Ditko                                                  0.000000

     29555 List of tourist attractions in Sardinia                      0.000000

     29579 Miller test                                                  0.000000

     29558 Gavinus                                                      0.000000

     29559 Sienna                                                       0.000000

     29564 Super Bowl XXXVI                                             0.000000

     29570 Scansano                                                     0.000000

     29574 List of maritime explorers                                   0.000000

     29577 Wednesday Morning, 3 A.M.                                    0.000000

     29578 Sheldon Rampton                                              0.000000

     29580 Set-top box                                                  0.000000

     29582 Scatology                                                    0.000000

     29586 Sigma-algebra                                                0.000000

     29587 Second Battle of El Alamein                                  0.000000

Query results for: Shah Rukh Khan

     DocID Title                                                        Relevance

     17253 Kubla Khan                                                   0.000372

     27098 Star Trek II: The Wrath of Khan                              0.000206

      8133 Dalai Lama                                                   0.000190

     58481 Badakhshan                                                   0.000190

     24260 Pervez Musharraf                                             0.000183

     36973 Pashtuns                                                     0.000182

     10312 East Pakistan                                                0.000178

     59597 Disc brake                                                   0.000178

      9058 European influence in Afghanistan                            0.000138

     18984 Mongols                                                      0.000134

     67397 Ulaanbaatar                                                  0.000130

     15245 Ismail Khan                                                  0.000126

      2695 Abdur Rahman Khan                                            0.000123

     40096 History of Bangladesh                                        0.000107

     34039 West Pakistan                                                0.000107

       737 Afghanistan                                                  0.000103

      2425 Aurangzeb                                                    0.000099

     34360 Yahya Khan                                                   0.000091

     21950 Khyber Pakhtunkhwa                                           0.000091

     11617 Feynman diagram                                              0.000089

BM 25

Query results for: Big Data

     DocID Title                                                        Relevance

     21125 Usenet newsgroup                                             9.363536

      7376 Cosmic microwave background                                  9.101180

      4261 Big Brother (Nineteen Eighty-Four)                           8.764335

      8495 Data set                                                     8.391962

      2732 Au file format                                               8.291928

     43026 Endianness                                                   8.220434

     31849 Telecommunications in Uruguay                                8.202649

     39137 Quintessence (physics)                                       8.174278

      2230 Analysis of algorithms                                       8.131582

     64489 Purchasing power parity                                      8.123138

     38798 Big Four accounting firms                                    7.751299

      5382 Inflation (cosmology)                                        7.671265

     47226 4179 Toutatis                                                7.453011

     20266 Mainframe computer                                           7.448758

      5378 Physical cosmology                                           7.439649

     20911 Multiverse                                                   7.391802

     26262 Redshift                                                     7.377892

     20894 Maximum transmission unit                                    7.339667

     22649 Observation                                                  7.294026

      2593 Accounting                                                   7.275168

Query results for: Game of thrones

     DocID Title                                                        Relevance

     12301 A Song of Ice and Fire                                       16.580270

     12300 George R. R. Martin                                          14.373137

     57493 Turbine (company)                                            13.712001

     76496 Max von Sydow                                                13.705445

      6607 Chaosium                                                     13.475070

     70653 The Lost Vikings (series)                                    12.836729

     12302 World of A Song of Ice and Fire                              12.314104

     21489 NetHack                                                      11.486217

     57550 Charles Edward Stuart                                        11.111935

     69567 Potiphar and his wife                                        10.913735

     39946 Charles XV of Sweden                                         10.211129

     61987 James Francis Edward Stuart                                  10.122846

     19537 MUD                                                          9.839998

     53014 List of Spanish monarchs                                     9.827676

     77825 TNT (TV channel)                                             9.691570

     69516 Philibert II, Duke of Savoy                                  9.134235

     24942 Patrilineality                                               9.013012

     79258 Iphigenia                                                    8.833192

     77446 Leucothea                                                    8.677797

     55116 Dubrovnik                                                    8.604608

Query results for: Alice in Wonderland

     DocID Title                                                        Relevance

     62446 Alice (Alice's Adventures in Wonderland)                     23.681057

     31623 Through the Looking-Glass                                    23.162918

     52853 Alice's Adventures in Wonderland                             22.996797

     62180 Alice in Wonderland (1933 film)                              22.050693

     52856 Alice in Wonderland (disambiguation)                         21.976944

      1500 Dodo (Alice's Adventures in Wonderland)                      21.148621

     74534 Alice Comedies                                               20.792648

     15612 John Tenniel                                                 20.398554

     15907 Jabberwocky                                                  19.122218

     45250 The Annotated Alice                                          18.727864

     78497 Jeff Noon                                                    18.659037

     80335 Helena Bonham Carter                                         18.093735

     45227 The Hunting of the Snark                                     15.692171

     31501 Tim Burton                                                   15.691451

     71870 Johnny Depp                                                  15.108505

     28217 Serial Experiments Lain                                      13.667296

     19500 Mervyn Peake                                                 13.493841

     32917 Walt Disney                                                  12.986983

     22725 On Fairy-Stories                                             12.806804

     45323 British Shorthair                                            12.775506

Query results for: Shakespeare

     DocID Title                                                        Relevance

      7712 Continuation War                                             10.228918

     50432 Colloquial Finnish                                           10.171282

     33924 Winter War                                                   10.128289

     50433 Finnish phonology                                            10.061059

     11772 Finnish Civil War                                            10.046662

     10715 Finnish Defence Forces                                       10.039559

     67408 Maamme                                                       10.014021

     50434 Finnish grammar                                              9.977256

     47661 Perkele                                                      9.951597

     10577 Finland                                                      9.939562

     14378 History of Finland                                           9.929193

     51083 Mikael Agricola                                              9.911247

     78302 Flag of Finland                                              9.873517

     66001 Ukko                                                         9.844840

     13696 Helsinki                                                     9.836820

     80522 Kokkola                                                      9.821437

     34446 Zachris Topelius                                             9.718290

     10710 Demographics of Finland                                      9.714814

     16777 Kalevala                                                     9.684821

     48672 Turku                                                        9.657883

Query results for: Data science

     DocID Title                                                        Relevance

      5323 Computer science                                             9.164645

      7671 Committee on Data for Science and Technology                 8.970093

     46127 Robert Tarjan                                                8.774521

     22649 Observation                                                  8.752901

     78204 International Polar Year                                     8.726871

     47795 Voyager program                                              8.683044

      5213 Computing                                                    8.588465

     45344 Collision (computer science)                                 8.558447

     45868 National Center for Supercomputing Applications              8.483373

     15627 Junk science                                                 8.438771

     58731 Peter Naur                                                   8.396507

     42253 Data mining                                                  8.359674

     11734 Fred Singer                                                  8.343525

     78172 International Geophysical Year                               8.299881

     26833 Scientific method                                            8.255052

     31492 The Skeptical Environmentalist                               8.236704

     16253 J. Philippe Rushton                                          8.223476

     26685 Statistics                                                   8.216694

     37876 National Oceanic and Atmospheric Administration              8.203268

     26781 Social science                                               8.197975

Query results for: University

     DocID Title                                                        Relevance

     72296 Peking University                                            3.893440

     60919 University of London                                         3.879486

     19735 Macquarie University                                         3.872889

     77940 University of North Carolina                                 3.872454

     32099 University of Manitoba                                       3.868881

     72599 Fudan University                                             3.864025

     43861 University of Southern Denmark                               3.863639

     66772 Rector (academia)                                            3.862307

     38091 Charles University                                           3.861856

     73276 University of Waterloo                                       3.859860

     27157 Stockholm University                                         3.856910

     53049 University of Vienna                                         3.855923

     77666 University of Stirling                                       3.855867

     38308 Manchester Metropolitan University                           3.855548

     39569 University of Glasgow                                        3.853994

     31928 University of Sydney                                         3.852159

     31798 Umeå University                                              3.851842

     32127 University of Chicago                                        3.850559

     32098 University of Victoria                                       3.850553

     66167 University constituency                                      3.850278

Query results for: Bali

     DocID Title                                                        Relevance

      4147 Bali                                                         14.515627

     76398 Trance and Dance in Bali                                     13.247874

     18360 Lombok                                                       13.189092

     14644 Geography of Indonesia                                       12.117559

     14579 Indonesia                                                    11.569830

     14649 Transport in Indonesia                                       11.336032

     32454 Geography of Venezuela                                       10.235141

     28696 Sumbawa                                                      10.027963

     44243 Nataraja                                                     9.604719

     69336 Java                                                         9.114847

     14643 History of Indonesia                                         8.935354

     33873 World Trade Organization                                     8.888242

     31898 United Nations Framework Convention on Climate Change        8.555007

     30075 Tiger                                                        8.310941

     52828 Lindsay Davenport                                            7.682327

     18183 Relative direction                                           7.638301

     22738 Operation Enduring Freedom                                   7.446773

     54920 Cockfight                                                    7.398711

     11100 Kite                                                         7.348049

     14651 Foreign relations of Indonesia                               7.266028

Query results for: Black Hole

     DocID Title                                                        Relevance

      7794 Crystallography                                              8.964742

      9730 Electron microscope                                          8.742224

     18151 Laser construction                                           8.462071

     72489 Electron energy loss spectroscopy                            8.234701

     11758 FASA                                                         8.193289

      8449 Developmental biology                                        8.069641

     61275 Cathodoluminescence                                          7.813923

     48366 Polyurethane                                                 7.701243

     25927 Rutherfordium                                                7.689396

     48579 Atomic, molecular, and optical physics                       7.536919

      7450 Context menu                                                 7.499346

     67958 Copernicium                                                  7.357776

     29107 Semantics                                                    7.353465

     16920 Knowledge representation and reasoning                       7.281770

     62903 Complex systems                                              7.263059

     12241 Gallium                                                      7.216225

     18120 Lysosome                                                     7.209814

     22208 Organic chemistry                                            7.192812

     24354 Pharmacology                                                 7.159937

     11429 False document                                               7.126060

Query results for: WESTLIFE

     DocID Title                                                        Relevance

     80356 History of Vietnam                                           3.459545

     29544 Scientific revolution                                        0.000000

     29549 Self-replication                                             0.000000

     29545 Salian dynasty                                               0.000000

     29556 List of people from Sardinia                                 0.000000

     29550 Shmuel Yosef Agnon                                           0.000000

     29551 Steve Ditko                                                  0.000000

     29555 List of tourist attractions in Sardinia                      0.000000

     29579 Miller test                                                  0.000000

     29558 Gavinus                                                      0.000000

     29559 Sienna                                                       0.000000

     29564 Super Bowl XXXVI                                             0.000000

     29570 Scansano                                                     0.000000

     29574 List of maritime explorers                                   0.000000

     29577 Wednesday Morning, 3 A.M.                                    0.000000

     29578 Sheldon Rampton                                              0.000000

     29580 Set-top box                                                  0.000000

     29582 Scatology                                                    0.000000

     29586 Sigma-algebra                                                0.000000

     29587 Second Battle of El Alamein                                  0.000000

Query results for: Shah Rukh Khan

     DocID Title                                                        Relevance

     24260 Pervez Musharraf                                             11.587601

     15245 Ismail Khan                                                  10.509455

      2695 Abdur Rahman Khan                                            10.339457

     58481 Badakhshan                                                   10.260425

      1546 Aga Khan III                                                 10.251657

     10312 East Pakistan                                                10.156445

     53459 Inayat Khan                                                  10.148452

     20886 Mohammed Zahir Shah                                          10.096899

     17253 Kubla Khan                                                   10.079720

     36973 Pashtuns                                                     10.020464

     27098 Star Trek II: The Wrath of Khan                              10.014914

      9058 European influence in Afghanistan                            9.995112

     34039 West Pakistan                                                9.973024

     53462 Vilayat Inayat Khan                                          9.944898

     34360 Yahya Khan                                                   9.889075

     17342 Kalat, Pakistan                                              9.879450

     51456 Mohammed Daoud Khan                                          9.849158

     45028 Mongol invasion of Rus'                                      9.805428

     67397 Ulaanbaatar                                                  9.755096

     60759 Pashtunistan                                                 9.709606

**REFERENCES**

|  |  |
| --- | --- |
| [1] | M. Zahara, H. Karau, A. Konwinski and P. Wendell, Learning Spark, Sebastopol: O’Reilly Media, Inc., 2015. |
| [2] | A. Trotman, "Learning to Rank," *Information Retrieval,* pp. 359-381, 2005. |
| [3] | A. Trotman, "Choosing Document Structure Weight," *Information processing & management,* pp. 243-264, 2005. |
| [4] | R. E. Walpole, R. H. Myers, S. L. Myers and K. Ye, Probability and Statistics for Engineers and Scientist, 9 ed., Prentice Hall, 2012. |