INFO300.LabExercise 5--Similarity

Date: November 06, 2024

Student Name: *Qirui Zhang* Class: *B* Email:*320220941080@lzu。edu.cn*

Goals: Practice with ElasticSearch Similarity

Notes:

- Use your name to replace "suwei".
- Run the following command and write the response of each command.
- Answer the questions.
- Please submit both MD and PDF files named "Week5Lab.yourName".

1). BM25 and boolean Similarity

Run the Command:

```
PUT /zhangqirui_simitest
{
  "mappings":{
    "properties": {
      "f1":{
        "type": "text"
      },
      "f2":{
        "type":"text",
        "similarity": "boolean"
      }
    }
  }
}
```

Your Response:

```
{
  "acknowledged": true,
  "shards_acknowledged": true,
  "index": "zhangqirui_simitest"
}
```

- Q1: What's the similarity function of field "f1" and "f2"?
- Your Answer:

```
The similarity function of f1 is the default similarity function BM25, based on
TF-IDF. While the similiarity function of f2 is boolean.
```

Run the Command:

```
POST /zhangqirui_simitest/_doc/1
{
  "f1":"Beijing Beijing Shanghai",
  "f2":"Lanzhou Lanzhou"
}
```

Your Response:

```
{
  "_index": "zhangqirui_simitest",
  "_id": "1",
  "_version": 1,
  "result": "created",
  "_shards": {
    "total": 2,
    "successful": 1,
    "failed": 0
  },
  "_seq_no": 0,
  "_primary_term": 1
}
```

Run the Command:

```
POST /zhangqirui_simitest/_doc/2
{
   "f1":"Beijing Tianjin",
   "f2":"Lanzhou Lanzhou Tianjin"
}
```

Your Response:

```
{
  "_index": "zhangqirui_simitest",
  "_id": "2",
  "_version": 1,
  "result": "created",
  "_shards": {
    "total": 2,
    "successful": 1,
    "failed": 0
  },
  "_seq_no": 1,
  "_primary_term": 1
}
```

Run the Command:

```
POST /zhangqirui_simitest/_doc/3
{
   "f1":"Beijing Beijing Lanzhou Tianjin",
   "f2":"Lanzhou Shanghai Tianjin"
}
```

Your Response:

```
{
  "_index": "zhangqirui_simitest",
  "_id": "3",
  "_version": 1,
  "result": "created",
  "_shards": {
    "total": 2,
    "successful": 1,
    "failed": 0
  },
  "_seq_no": 2,
  "_primary_term": 1
}
```

Run the Command:

```
GET /zhangqirui_simitest/_search
{
  "query": {
    "multi_match" : {
      "query": "Beijing",
      "fields": ["f1"]
    }
  }
}
```

Your Response:

```
{
  "took": 1,
  "timed_out": false,
  "_shards": {
    "total": 1,
    "successful": 1,
    "skipped": 0,
    "failed": 0
  },
  "hits": {
    "total": {
      "value": 3,
      "relation": "eq"
    },
    "max_score": 0.18360566,
    "hits": [
      {
        "_index": "zhangqirui_simitest",
        "_id": "1",
        "_score": 0.18360566,
        "_source": {
          "f1": "Beijing Beijing Shanghai",
          "f2": "Lanzhou Lanzhou"
        }
      },
      {
        "_index": "zhangqirui_simitest",
        "_id": "3",
        "_score": 0.16786805,
        "_source": {
          "f1": "Beijing Beijing Lanzhou Tianjin",
          "f2": "Lanzhou Shanghai Tianjin"
        }
      },
      {
        "_index": "zhangqirui_simitest",
        "_id": "2",
        "_score": 0.1546153,
        "_source": {
          "f1": "Beijing Tianjin",
          "f2": "Lanzhou Lanzhou Tianjin"
        }
      }
    ]
  }
}
```

- Q2: What's the sequence of retrieved documents? What's the score of each retrieved document?
- Your Answer:

```
```

```
Document1 with the score of 0.18360566, document3 with the score of 0.16786805 and
document 2 with the score of 0.1546153
```

Run the Command:

```
GET /zhangqirui_simitest/_search
{
  "query": {
    "multi_match" : {
      "query":    "Lanzhou",
      "fields": ["f2"]
    }
  }
}
```

Your Response:

```
{
  "took": 1,
  "timed_out": false,
  "_shards": {
    "total": 1,
    "successful": 1,
    "skipped": 0,
    "failed": 0
  },
  "hits": {
    "total": {
      "value": 3,
      "relation": "eq"
    },
    "max_score": 1,
    "hits": [
      {
        "_index": "zhangqirui_simitest",
        "_id": "1",
        "_score": 1,
        "_source": {
          "f1": "Beijing Beijing Shanghai",
          "f2": "Lanzhou Lanzhou"
        }
      },
      {
        "_index": "zhangqirui_simitest",
        "_id": "2",
        "_score": 1,
        "_source": {
          "f1": "Beijing Tianjin",
          "f2": "Lanzhou Lanzhou Tianjin"
        }
      },
      {
        "_index": "zhangqirui_simitest",
        "_id": "3",
        "_score": 1,
        "_source": {
          "f1": "Beijing Beijing Lanzhou Tianjin",
          "f2": "Lanzhou Shanghai Tianjin"
        }
      }
    ]
  }
}
```

- Q3: What's the sequence of retrieved documents? What's the score of each retrieved document?
- Your Answer:

```
Document1, document2 and document3 got the same score, because they all contain
the required word and were given the highest score when being retrieved with
boolean similarity function.
```

Run the Command:

```
GET /suwei_simitest/_search?explain=true
{
  "query": {
    "multi_match" : {
      "query": "Beijing",
      "fields": ["f1"]
    }
  }
}
```

Your Response:

```
{
  "took": 2,
  "timed_out": false,
  "_shards": {
    "total": 1,
    "successful": 1,
    "skipped": 0,
    "failed": 0
  },
  "hits": {
    "total": {
      "value": 3,
      "relation": "eq"
    },
    "max_score": 0.18360566,
    "hits": [
      {
        "_shard": "[zhangqirui_simitest][0]",
        "_node": "XlrvdNasSlSKrcRF6sFS9Q",
        "_index": "zhangqirui_simitest",
        "_id": "1",
        "_score": 0.18360566,
        "_source": {
          "f1": "Beijing Beijing Shanghai",
          "f2": "Lanzhou Lanzhou"
        },
        "_explanation": {
          "value": 0.18360566,
          "description": "weight(f1:beijing in 0) [PerFieldSimilarity], result of:",
          "details": [
            {
              "value": 0.18360566,
              "description": "score(freq=2.0), computed as boost * idf * tf from:",
              "details": [
                {
                  "value": 2.2,
                  "description": "boost",
                  "details": []
                },
                {
                  "value": 0.13353139,
                  "description": "idf, computed as log(1 + (N - n + 0.5) / (n + 0.5)) from:",
                  "details": [
                    {
                      "value": 3,
                      "description": "n, number of documents containing term",
                      "details": []
                    },
                    {
                      "value": 3,
                      "description": "N, total number of documents with field",
                      "details": []
                    }
                  ]
                },
                {
                  "value": 0.625,
                  "description": "tf, computed as freq / (freq + k1 * (1 - b + b * dl / avgdl)) from:",
                  "details": [
                    {
                      "value": 2,
                      "description": "freq, occurrences of term within document",
                      "details": []
                    },
                    {
                      "value": 1.2,
                      "description": "k1, term saturation parameter",
                      "details": []
                    },
                    {
                      "value": 0.75,
                      "description": "b, length normalization parameter",
                      "details": []
                    },
                    {
                      "value": 3,
                      "description": "dl, length of field",
                      "details": []
                    },
                    {
                      "value": 3,
                      "description": "avgdl, average length of field",
                      "details": []
                    }
                  ]
                }
              ]
            }
          ]
        }
      },
      {
        "_shard": "[zhangqirui_simitest][0]",
        "_node": "XlrvdNasSlSKrcRF6sFS9Q",
        "_index": "zhangqirui_simitest",
        "_id": "3",
        "_score": 0.16786805,
        "_source": {
          "f1": "Beijing Beijing Lanzhou Tianjin",
          "f2": "Lanzhou Shanghai Tianjin"
        },
        "_explanation": {
          "value": 0.16786805,
          "description": "weight(f1:beijing in 0) [PerFieldSimilarity], result
of:",
          "details": [
            {
              "value": 0.16786805,
              "description": "score(freq=2.0), computed as boost * idf * tf
from:",
              "details": [
                {
                  "value": 2.2,
                  "description": "boost",
                  "details": []
                },
                {
                  "value": 0.13353139,
                  "description": "idf, computed as log(1 + (N - n + 0.5) / (n + 0.5)) from:",
                  "details": [
                    {
                      "value": 3,
                      "description": "n, number of documents containing term",
                      "details": []
                    },
                    {
                      "value": 3,
                      "description": "N, total number of documents with field",
                      "details": []
                    }
                  ]
                },
                {
                  "value": 0.5714286,
                  "description": "tf, computed as freq / (freq + k1 * (1 - b + b * dl / avgdl)) from:",
                  "details": [
                    {
                      "value": 2,
                      "description": "freq, occurrences of term within document",
                      "details": []
                    },
                    {
                      "value": 1.2,
                      "description": "k1, term saturation parameter",
                      "details": []
                    },
                    {
                      "value": 0.75,
                      "description": "b, length normalization parameter",
                      "details": []
                    },
                    {
                      "value": 4,
                      "description": "dl, length of field",
                      "details": []
                    },
                    {
                      "value": 3,
                      "description": "avgdl, average length of field",
                      "details": []
                    }
                  ]
                }
              ]
            }
          ]
        }
```

```
      },
      {
        "_shard": "[zhangqirui_simitest][0]",
        "_node": "XlrvdNasSlSKrcRF6sFS9Q",
        "_index": "zhangqirui_simitest",
        "_id": "2",
        "_score": 0.1546153,
        "_source": {
          "f1": "Beijing Tianjin",
          "f2": "Lanzhou Lanzhou Tianjin"
        },
        "_explanation": {
          "value": 0.1546153,
          "description": "weight(f1:beijing in 0) [PerFieldSimilarity], result of:",
          "details": [
            {
              "value": 0.1546153,
              "description": "score(freq=1.0), computed as boost * idf * tf from:",
              "details": [
                {
                  "value": 2.2,
                  "description": "boost",
                  "details": []
                },
                {
                  "value": 0.13353139,
                  "description": "idf, computed as log(1 + (N - n + 0.5) / (n + 0.5)) from:",
                  "details": [
                    {
                      "value": 3,
                      "description": "n, number of documents containing term",
                      "details": []
                    },
                    {
                      "value": 3,
                      "description": "N, total number of documents with field",
                      "details": []
                    }
                  ]
                },
                {
                  "value": 0.5263158,
                  "description": "tf, computed as freq / (freq + k1 * (1 - b + b * dl / avgdl)) from:",
                  "details": [
                    {
                      "value": 1,
                      "description": "freq, occurrences of term within document",
                      "details": []
                    },
                    {
                      "value": 1.2,
                      "description": "k1, term saturation parameter",
                      "details": []
                    },
                    {
                      "value": 0.75,
                      "description": "b, length normalization parameter",
                      "details": []
                    },
                    {
                      "value": 2,
                      "description": "dl, length of field",
                      "details": []
                    },
                    {
                      "value": 3,
                      "description": "avgdl, average length of field",
                      "details": []
                    }
                  ]
                }
              ]
            }
          ]
        }
      }
    ]
  }
}
```

- Q4: Please give the formula for computing the similarity score. Please write down the values of each parameter and variable for Document “3”.
- Your Answer:

$score(d, q)=\sum_{t\in q}idf(t)\frac{(k_1+1)tf(t,d)}{k_1(1-b+b\frac{dl(d)}{aygdl})+tf(t,d)}$

```
d:document
q:query
t:term
tf(t,d):the frequence of term t in document d
k: parameter with the default value of 1.2
b: parameter with the default value of 0.75
dl(d): length of document
avgdl: average length of document.

For document3, tf(t,d)=2, idf(t)=0.13353, dl(d)=4, avgdl=3

score(d,q)=0.16786805
```

Run the Command:

```
GET /zhangqirui_simitest/_search?explain=true
{
  "query": {
    "multi_match" : {
      "query":    "Lanzhou",
      "fields": ["f2"]
    }
  }
}
```

Your Response:

```
{
  "took": 2,
  "timed_out": false,
  "_shards": {
    "total": 1,
    "successful": 1,
    "skipped": 0,
    "failed": 0
  },
  "hits": {
    "total": {
      "value": 3,
      "relation": "eq"
    },
    "max_score": 1,
    "hits": [
      {
        "_shard": "[zhangqirui_simitest][0]",
        "_node": "XlrvdNasSlSKrcRF6sFS9Q",
        "_index": "zhangqirui_simitest",
        "_id": "1",
        "_score": 1,
        "_source": {
          "f1": "Beijing Beijing Shanghai",
          "f2": "Lanzhou Lanzhou"
        },
        "_explanation": {
          "value": 1,
          "description": "weight(f2:lanzhou in 0) [PerFieldSimilarity], result
of:",
          "details": [
            {
              "value": 1,
              "description": "score(BooleanWeight), computed from:",
              "details": [
                {
                  "value": 1,
                  "description": "boost, query boost",
                  "details": []
                }
              ]
            }
          ]
        }
      },
      {
        "_shard": "[zhangqirui_simitest][0]",
        "_node": "XlrvdNasSlSKrcRF6sFS9Q",
        "_index": "zhangqirui_simitest",
        "_id": "2",
        "_score": 1,
        "_source": {
          "f1": "Beijing Tianjin",
          "f2": "Lanzhou Lanzhou Tianjin"
        },
        "_explanation": {
          "value": 1,
          "description": "weight(f2:lanzhou in 0) [PerFieldSimilarity], result of:",
          "details": [
            {
              "value": 1,
              "description": "score(BooleanWeight), computed from:",
              "details": [
                {
                  "value": 1,
                  "description": "boost, query boost",
                  "details": []
                }
              ]
            }
          ]
        }
      },
      {
        "_shard": "[zhangqirui_simitest][0]",
        "_node": "XlrvdNasSlSKrcRF6sFS9Q",
        "_index": "zhangqirui_simitest",
        "_id": "3",
        "_score": 1,
        "_source": {
          "f1": "Beijing Beijing Lanzhou Tianjin",
          "f2": "Lanzhou Shanghai Tianjin"
        },
        "_explanation": {
          "value": 1,
          "description": "weight(f2:lanzhou in 0) [PerFieldSimilarity], result of:",
          "details": [
            {
              "value": 1,
              "description": "score(BooleanWeight), computed from:",
              "details": [
                {
                  "value": 1,
                  "description": "boost, query boost",
                  "details": []
                }
              ]
            }
          ]
        }
      }
    ]
  }
}
```

2). DFR as Default Similarity

Run the Command:

```
GET /zhangqirui_simitest/_mapping
```

Your Response:

```
{
  "zhangqirui_simitest": {
    "mappings": {
      "properties": {
        "f1": {
          "type": "text"
        },
        "f2": {
          "type": "text",
          "similarity": "boolean"
        }
      }
    }
  }
}
```

Run the Command:

```
POST /zhangqirui_simitest/_close?wait_for_active_shards=0
```

Your Response:

```
{
  "acknowledged": true,
  "shards_acknowledged": true,
  "indices": {
    "zhangqirui_simitest": {
      "closed": true
    }
  }
}
```

Run the Command:

```
PUT /zhangqirui_simitest/_settings
{
  "index": {
    "similarity": {
      "default": {
        "type": "DFR",
        "basic_model": "g",
        "after_effect": "l",
        "normalization": "h2",
        "normalization.h2.c": "3.0"
      }
    }
  }
}
```

Your Response:

```
{
  "acknowledged": true
}
```

Run the Command:

```
POST /zhangqirui_simitest/_open
```

Your Response:

```
{
  "acknowledged": true,
  "shards_acknowledged": true
}
```

Run the Command:

```
GET /zhangqirui_simitest/_search?explain=true
{
  "query": {
    "multi_match" : {
      "query": "Beijing",
      "fields": ["f1"]
    }
  }
}
```

Your Response:

```
{
  "took": 2,
  "timed_out": false,
  "_shards": {
    "total": 1,
    "successful": 1,
    "skipped": 0,
    "failed": 0
  },
  "hits": {
    "total": {
      "value": 3,
      "relation": "eq"
    },
    "max_score": 1.2049356,
    "hits": [
      {
        "_shard": "[zhangqirui_simitest][0]",
        "_node": "XlrvdNasSlSKrcRF6sFS9Q",
        "_index": "zhangqirui_simitest",
        "_id": "1",
        "_score": 1.2049356,
        "_source": {
          "f1": "Beijing Beijing Shanghai",
          "f2": "Lanzhou Lanzhou"
        },
        "_explanation": {
          "value": 1.2049356,
          "description": "weight(f1:beijing in 0) [PerFieldSimilarity], result of:",
          "details": [
            {
              "value": 1.2049356,
              "description": "score(DFRSimilarity, freq=2.0), computed as boost * basicModel.score(stats, tfn) * afterEffect.score(stats, tfn) from:",
              "details": [
                {
                  "value": 4,
                  "description": "NormalizationH2, computed as tf * log2(1 + c * avgfl / fl) from:",
                  "details": [
                    {
                      "value": 2,
                      "description": "tf, number of occurrences of term in the document",
                      "details": []
                    },
                    {
                      "value": 3,
                      "description": "c, hyper-parameter",
                      "details": []
                    },
                    {
                      "value": 3,
                      "description": "avgfl, average length of field across all documents",
                      "details": []
                    },
                    {
                      "value": 3,
                      "description": "fl, field length of the document",
                      "details": []
                    }
                  ]
                },
                {
                  "value": 6.0246778,
                  "description": "BasicModelG, computed as log2(lambda + 1) + tfn * log2((1 + lambda) / lambda) from:",
                  "details": [
                    {
                      "value": 4,
                      "description": "tfn, normalized term frequency",
                      "details": []
                    },
                    {
                      "value": 0.6666667,
                      "description": "lambda, computed as F / (N + F) from:",
                      "details": [
                        {
                          "value": 6,
                          "description": "F, total number of occurrences of term across all docs + 1",
                          "details": []
                        },
                        {
                          "value": 3,
                          "description": "N, total number of documents with field",
                          "details": []
                        }
                      ]
                    }
                  ]
                },
                {
                  "value": 0.2,
                  "description": "AfterEffectL, computed as 1 / (tfn + 1) from:",
                  "details": [
                    {
                      "value": 4,
                      "description": "tfn, normalized term frequency",
                      "details": []
                    }
                  ]
                }
              ]
            }
          ]
        }
      },
      {
        "_shard": "[zhangqirui_simitest][0]",
        "_node": "XlrvdNasSlSKrcRF6sFS9Q",
        "_index": "zhangqirui_simitest",
        "_id": "3",
        "_score": 1.1890086,
        "_source": {
          "f1": "Beijing Beijing Lanzhou Tianjin",
          "f2": "Lanzhou Shanghai Tianjin"
        },
        "_explanation": {
          "value": 1.1890086,
          "description": "weight(f1:beijing in 0) [PerFieldSimilarity], result
of:",
          "details": [
            {
              "value": 1.1890086,
              "description": "score(DFRSimilarity, freq=2.0), computed as boost *
basicModel.score(stats, tfn) * afterEffect.score(stats, tfn) from:",
              "details": [
                {
                  "value": 3.4008794,
                  "description": "NormalizationH2, computed as tf * log2(1 + c *
avgfl / fl) from:",
                  "details": [
                    {
                      "value": 2,
                      "description": "tf, number of occurrences of term in the
document",
                      "details": []
                    },
                    {
                      "value": 3,
                      "description": "c, hyper-parameter",
                      "details": []
                    },
                    {
                      "value": 3,
                      "description": "avgfl, average length of field across all documents",
                      "details": []
                    },
                    {
                      "value": 4,
                      "description": "fl, field length of the document",
                      "details": []
                    }
                  ]
                },
                {
                  "value": 5.2326837,
                  "description": "BasicModelG, computed as log2(lambda + 1) + tfn * log2((1 + lambda) / lambda) from:",
                  "details": [
                    {
                      "value": 3.4008794,
                      "description": "tfn, normalized term frequency",
                      "details": []
                    },
                    {
                      "value": 0.6666667,
                      "description": "lambda, computed as F / (N + F) from:",
                      "details": [
                        {
                          "value": 6,
                          "description": "F, total number of occurrences of term across all docs + 1",
                          "details": []
                        },
                        {
                          "value": 3,
                          "description": "N, total number of documents with field",
                          "details": []
                        }
                      ]
                    }
                  ]
                },
                {
                  "value": 0.22722732,
                  "description": "AfterEffectL, computed as 1 / (tfn + 1) from:",
                  "details": [
                    {
                      "value": 3.4008794,
                      "description": "tfn, normalized term frequency",
                      "details": []
                    }
                  ]
                }
              ]
            }
          ]
        }
      },
      {
        "_shard": "[zhangqirui_simitest][0]",
        "_node": "XlrvdNasSlSKrcRF6sFS9Q",
        "_index": "zhangqirui_simitest",
        "_id": "2",
        "_score": 1.152836,
        "_source": {
          "f1": "Beijing Tianjin",
          "f2": "Lanzhou Lanzhou Tianjin"
        },
        "_explanation": {
          "value": 1.152836,
          "description": "weight(f1:beijing in 0) [PerFieldSimilarity], result of:",
          "details": [
            {
              "value": 1.152836,
              "description": "score(DFRSimilarity, freq=1.0), computed as boost * basicModel.score(stats, tfn) * afterEffect.score(stats, tfn) from:",
              "details": [
                {
                  "value": 2.4594316,
                  "description": "NormalizationH2, computed as tf * log2(1 + c * avgfl / fl) from:",
                  "details": [
                    {
                      "value": 1,
                      "description": "tf, number of occurrences of term in the document",
                      "details": []
                    },
                    {
                      "value": 3,
                      "description": "c, hyper-parameter",
                      "details": []
                    },
                    {
                      "value": 3,
                      "description": "avgfl, average length of field across all documents",
                      "details": []
                    },
                    {
                      "value": 2,
                      "description": "fl, field length of the document",
                      "details": []
                    }
                  ]
                },
                {
                  "value": 3.9881573,
                  "description": "BasicModelG, computed as log2(lambda + 1) + tfn * log2((1 + lambda) / lambda) from:",
                  "details": [
                    {
                      "value": 2.4594316,
                      "description": "tfn, normalized term frequency",
                      "details": []
                    },
                    {
                      "value": 0.6666667,
                      "description": "lambda, computed as F / (N + F) from:",
                      "details": [
                        {
                          "value": 6,
                          "description": "F, total number of occurrences of term across all docs + 1",
                          "details": []
                        },
                        {
                          "value": 3,
                          "description": "N, total number of documents with field",
                          "details": []
                        }
                      ]
                    }
                  ]
                },
                {
                  "value": 0.28906482,
                  "description": "AfterEffectL, computed as 1 / (tfn + 1) from:",
                  "details": [
                    {
                      "value": 2.4594316,
                      "description": "tfn, normalized term frequency",
                      "details": []
                    }
                  ]
                }
              ]
            }
          ]
        }
      }
    ]
  }
}
```

- Q5: Look into the document of Elastic Search, please write out all the possible values of each parameter for DFR similarity.
- Your Answer:

```
 "basic_model": 6.0246778, 5.2326837, 3.9881573
 "after_effect": 0.2, 0.22722732, 0.28906482
 "normalization": 4, 3.4008794, 2.4594316
```

3). DFI and IB Similarity

Run the Command:

```
PUT /suwei_simitest1
{
  "settings":{
    "index": {
      "similarity":{
        "my_dfi":{
          "type":"DFI",
          "independence_measure":"standardized"
        },
        "my_ib":{
          "type":"IB",
          "distribution":"ll",
          "lambda":"df",
          "normalization":"h1"
        }
      }
    }
  }
}
```

Your Response:

```
{
  "acknowledged": true,
  "shards_acknowledged": true,
  "index": "zhangqirui_simitest1"
}
```

Run the Command:

```
PUT /zhangqirui_simitest1/_mapping
{
  "properties":{
      "f1":{
        "type": "text",
        "similarity":"my_dfi"
      },
      "f2":{
        "type":"text",
        "similarity": "my_ib"
      }
    }
}
```

Your Response:

```
{
  "acknowledged": true
}
```

Run the Command:

```
POST /suwei_simitest1/_doc/1
{
  "f1":"Beijing Lanzhou",
  "f2":"Lanzhou Lanzhou"
}

POST /suwei_simitest1/_doc/2
{
   "f1":"Beijing Beijing Tianjin",
   "f2":"Beijing Tianjin Tianjin"
}

POST /suwei_simitest1/_doc/3
{
   "f1":"Lanzhou Lanzhou Tianjin",
   "f2":"Lanzhou Tianjin Lanzhou"
}
```

Run the Command:

```
GET /suwei_simitest1/_search?explain=true
{
  "query": {
    "multi_match" : {
      "query": "Beijing",
      "fields": ["f1"]
    }
    }
  }
```

Your Response:

```
{
  "took": 14,
  "timed_out": false,
  "_shards": {
    "total": 1,
    "successful": 1,
    "skipped": 0,
    "failed": 0
  },
  "hits": {
    "total": {
      "value": 2,
      "relation": "eq"
    },
    "max_score": 0.65750307,
    "hits": [
      {
        "_shard": "[zhangqirui_simitest1][0]",
        "_node": "XlrvdNasSlSKrcRF6sFS9Q",
        "_index": "zhangqirui_simitest1",
        "_id": "2",
        "_score": 0.65750307,
        "_source": {
          "f1": "Beijing Beijing Tianjin",
          "f2": "Beijing Tianjin Tianjin"
        },
        "_explanation": {
          "value": 0.65750307,
          "description": "weight(f1:beijing in 0) [PerFieldSimilarity], result of:",
          "details": [
            {
              "value": 0.65750307,
              "description": "score(DFISimilarity, freq=2.0), computed as boost * log2(measure + 1) from:",
              "details": [
                {
                  "value": 1,
                  "description": "boost, query boost",
                  "details": []
                },
                {
                  "value": 0.57735026,
                  "description": "measure, computed as independence.score(freq, expected) from:",
                  "details": [
                    {
                      "value": 2,
                      "description": "freq, occurrences of term within document",
                      "details": []
                    },
                    {
                      "value": 1.3333334,
                      "description": "expected, computed as (F + 1) * dl / (T + 1) from:",
                      "details": [
                        {
                          "value": 3,
                          "description": "F, total number of occurrences of term across all docs",
                          "details": []
                        },
                        {
                          "value": 3,
                          "description": "dl, length of field",
                          "details": []
                        },
                        {
                          "value": 8,
                          "description": "T, total number of tokens in the field",
                          "details": []
                        }
                      ]
                    }
                  ]
                }
              ]
            }
          ]
        }
      },
      {
        "_shard": "[zhangqirui_simitest1][0]",
        "_node": "XlrvdNasSlSKrcRF6sFS9Q",
        "_index": "zhangqirui_simitest1",
        "_id": "1",
        "_score": 0.16072807,
        "_source": {
          "f1": "Beijing Lanzhou",
          "f2": "Lanzhou Lanzhou"
        },
        "_explanation": {
          "value": 0.16072807,
          "description": "weight(f1:beijing in 0) [PerFieldSimilarity], result of:",
          "details": [
            {
              "value": 0.16072807,
              "description": "score(DFISimilarity, freq=1.0), computed as boost * log2(measure + 1) from:",
              "details": [
                {
                  "value": 1,
                  "description": "boost, query boost",
                  "details": []
                },
                {
                  "value": 0.11785113,
                  "description": "measure, computed as independence.score(freq, expected) from:",
                  "details": [
                    {
                      "value": 1,
                      "description": "freq, occurrences of term within document",
                      "details": []
                    },
                    {
                      "value": 0.8888889,
                      "description": "expected, computed as (F + 1) * dl / (T + 1) from:",
                      "details": [
                        {
                          "value": 3,
                          "description": "F, total number of occurrences of term across all docs",
                          "details": []
                        },
                        {
                          "value": 2,
                          "description": "dl, length of field",
                          "details": []
                        },
                        {
                          "value": 8,
                          "description": "T, total number of tokens in the field",
                          "details": []
                        }
                      ]
                    }
                  ]
                }
              ]
            }
          ]
        }
      }
    ]
  }
}
```

Run the Command:

```
GET /suwei_simitest1/_search?explain=true
{
  "query": {
    "multi_match" : {
      "query": "Lanzhou",
      "fields": ["f2"]
    }
  }
}
```

Your Response:

```

{
  "took": 2,
  "timed_out": false,
  "_shards": {
    "total": 1,
    "successful": 1,
    "skipped": 0,
    "failed": 0
  },
  "hits": {
    "total": {
      "value": 2,
      "relation": "eq"
    },
    "max_score": 1.5163475,
    "hits": [
      {
        "_shard": "[zhangqirui_simitest1][0]",
        "_node": "XlrvdNasSlSKrcRF6sFS9Q",
        "_index": "zhangqirui_simitest1",
        "_id": "1",
        "_score": 1.5163475,
        "_source": {
          "f1": "Beijing Lanzhou",
          "f2": "Lanzhou Lanzhou"
        },
        "_explanation": {
          "value": 1.5163475,
          "description": "weight(f2:lanzhou in 0) [PerFieldSimilarity], result
of:",
          "details": [
            {
              "value": 1.5163475,
              "description": "score(IBSimilarity, freq=2.0), computed as boost *
distribution.score(stats, normalization.tfn(stats, freq, docLen),
lambda.lambda(stats)) from:",
              "details": [
                {
                  "value": 2.6666667,
         "description": "NormalizationH1, computed as tf * c * (avgfl / fl) from:",
         "details": [
          {
           "value": 2,
           "description": "tf, number of occurrences of term in the document",
           "details": []
          },
          {
           "value": 1,
           "description": "c, hyper-parameter",
           "details": []
          },
          {
           "value": 2.6666667,
           "description": "avgfl, average length of field across all documents",
           "details": []
          },
          {
           "value": 2,
           "description": "fl, field length of the document",
           "details": []
          }
         ]
        },
        {
         "value": 0.75,
         "description": "LambdaDF, computed as (n + 1) / (N + 1) from:",
         "details": [
          {
           "value": 2,
           "description": "n, number of documents containing term",
           "details": []
          },
          {
           "value": 3,
           "description": "N, total number of documents with field",
           "details": []
          }
         ]
        },
        {
         "value": 1.5163475,
         "description": "DistributionLL",
         "details": []
        }
       ]
      }
     ]
    }
   },
   {
        "_shard": "[zhangqirui_simitest1][0]",
        "_node": "XlrvdNasSlSKrcRF6sFS9Q",
        "_index": "zhangqirui_simitest1",
        "_id": "3",
        "_score": 1.2150227,
        "_source": {
          "f1": "Lanzhou Lanzhou Tianjin",
          "f2": "Lanzhou Tianjin Lanzhou"
        },
        "_explanation": {
          "value": 1.2150227,
          "description": "weight(f2:lanzhou in 1) [PerFieldSimilarity], result of:",
          "details": [
            {
              "value": 1.2150227,
              "description": "score(IBSimilarity, freq=2.0), computed as boost * distribution.score(stats, normalization.tfn(stats, freq, docLen), lambda.lambda(stats)) from:",
              "details": [
                {
                  "value": 1.7777778,
                  "description": "NormalizationH1, computed as tf * c * (avgfl / fl) from:",
                  "details": [
                    {
                      "value": 2,
                      "description": "tf, number of occurrences of term in the document",
                      "details": []
                    },
                    {
                      "value": 1,
                      "description": "c, hyper-parameter",
                      "details": []
                    },
                    {
                      "value": 2.6666667,
                      "description": "avgfl, average length of field across all documents",
                      "details": []
                    },
                    {
                      "value": 3,
                      "description": "fl, field length of the document",
                      "details": []
                    }
                  ]
                },
                {
                  "value": 0.75,
                  "description": "LambdaDF, computed as (n + 1) / (N + 1) from:",
                  "details": [
                    {
                      "value": 2,
                      "description": "n, number of documents containing term",
                      "details": []
                    },
                    {
                      "value": 3,
                      "description": "N, total number of documents with field",
                      "details": []
                    }
                  ]
                },
                {
                  "value": 1.2150227,
                  "description": "DistributionLL",
                  "details": []
                }
              ]
            }
          ]
        }
      }
    ]
  }
}
```

4). Scripted Similarity

Run the Command:

```
PUT /suwei_simitest_mytfidf/
{
  "settings": {
    "number_of_shards": 1,
    "similarity":{
      "su_tfidf":{
        "type":"scripted",
        "script":{
          "source":"double tf=Math.sqrt(doc.freq);double idf=Math.log((field.docCount+3.0)/(term.docFreq+2.0))+3.0;double norm=1/Math.sqrt(doc.length);return query.boost * tf * idf * norm;"
        }
      }
    }
  },
  "mappings": {
    "properties": {
      "books":{
        "type":"text",
        "similarity": "su_tfidf"
      }
    }
  }
}
```

Your Response:

```
{
  "acknowledged": true,
  "shards_acknowledged": true,
  "index": "zhangqirui_simitest_mytfidf"
}
```

Run the Command:

```
PUT /suwei_simitest_mytfidf/_doc/1
{
  "books":"Lanzhou University, Information Retrieval, I love Lanzhou, Lanzhou Beef
Noodle, Information Theory"
}

Get /suwei_simitest_mytfidf/_search?explain=true
{
  "query":{
    "query_string": {
      "query": "Information^1.8 Lanzhou",
      "default_field": "books"
    }
  }
}
```

Response:

```
{
  "took": 5,
  "timed_out": false,
  "_shards": {
    "total": 1,
    "successful": 1,
    "skipped": 0,
    "failed": 0
  },
  "hits": {
    "total": {
      "value": 1,
      "relation": "eq"
    },
    "max_score": 4.059784,
    "hits": [
      {
        "_shard": "[zhangqirui_simitest_mytfidf][0]",
        "_node": "XlrvdNasSlSKrcRF6sFS9Q",
        "_index": "zhangqirui_simitest_mytfidf",
        "_id": "1",
        "_score": 4.059784,
        "_source": {
          "books": "Lanzhou University, Information Retrieval, I love Lanzhou, Lanzhou Beef Noodle, Information Theory"
        },
        "_explanation": {
          "value": 4.059784,
          "description": "sum of:",
          "details": [
            {
              "value": 2.415943,
              "description": "weight(books:information in 0) [PerFieldSimilarity], result of:",
              "details": [
                {
                  "value": 2.415943,
                  "description": "score from ScriptedSimilarity(weightScript=[null], script=[Script{type=inline, lang='painless', idOrCode='double tf=Math.sqrt(doc.freq);double idf=Math.log((field.docCount+3.0)/(term.docFreq+2.0))+3.0;double norm=1/Math.sqrt(doc.length);return query.boost * tf * idf * norm;', options={}, params={}}]) computed from:",
                  "details": [
                    {
                      "value": 1,
                      "description": "weight",
                      "details": []
                    },
                    {
                      "value": 1.8,
                      "description": "query.boost",
                      "details": []
                    },
                    {
                      "value": 1,
                      "description": "field.docCount",
                      "details": []
                    },
                    {
                      "value": 9,
                      "description": "field.sumDocFreq",
                      "details": []
                    },
                    {
                      "value": 12,
                      "description": "field.sumTotalTermFreq",
                      "details": []
                    },
                    {
                      "value": 1,
                      "description": "term.docFreq",
                      "details": []
                    },
                    {
                      "value": 2,
                      "description": "term.totalTermFreq",
                      "details": []
                    },
                    {
                      "value": 2,
                      "description": "doc.freq",
                      "details": []
                    },
                    {
                      "value": 12,
                      "description": "doc.length",
                      "details": []
                    }
                  ]
                }
              ]
            },
            {
              "value": 1.643841,
              "description": "weight(books:lanzhou in 0) [PerFieldSimilarity],
result of:",
              "details": [
                {
                  "value": 1.643841,
                  "description": "score from ScriptedSimilarity(weightScript=
[null], script=[Script{type=inline, lang='painless', idOrCode='double
tf=Math.sqrt(doc.freq);double
idf=Math.log((field.docCount+3.0)/(term.docFreq+2.0))+3.0;double
norm=1/Math.sqrt(doc.length);return query.boost * tf * idf * norm;', options={},
params={}}]) computed from:",
                  "details": [
                    {
                      "value": 1,
                      "description": "weight",
                      "details": []
                    },
                    {
                      "value": 1,
                      "description": "query.boost",
                      "details": []
                    },
                    {
                      "value": 1,
                      "description": "field.docCount",
                      "details": []
                    },
                    {
                      "value": 9,
                      "description": "field.sumDocFreq",
                      "details": []
                    },
                    {
                      "value": 12,
                      "description": "field.sumTotalTermFreq",
                      "details": []
                    },
                    {
                      "value": 1,
                      "description": "term.docFreq",
                      "details": []
                    },
                    {
                      "value": 3,
                      "description": "term.totalTermFreq",
                      "details": []
                    },
                    {
                      "value": 3,
                      "description": "doc.freq",
                      "details": []
                    },
                    {
                      "value": 12,
                      "description": "doc.length",
                      "details": []
                    }
                  ]
                }
              ]
            }
          ]
        }
      }
    ]
  }
}
```

- Q6: Please give the formula for computing the similarity score. Please write down the values of each parameter for the query "Information^1.8 Lanzhou".
- Your Answer:

```
tf=Math.sqrt(doc.freq);
idf=Math.log((field.docCount+3.0)/(term.docFreq+2.0))+3.0;
norm=1/Math.sqrt(doc.length);
score=query.boost * tf * idf * norm;

part1:
weight:1
doc.freq:2
```

```
field.docCount:1
term.docFreq:1
doc.length:12
query.boost:1.8
tf:1.414
idf:3.125
norm:0.289
score: 2.415943

part2:
weight:1
doc.freq:1
field.docCount:1
term.docFreq:1
doc.length:12
query.boost:1
tf:1
idf:3.125
norm:0.289
score:1.643841
```