MIDTERM REPORT: Search Module

CENG3530, INFORMATION RETRIEVAL SYSTEMS

Beyza Kurt
beyzakurt@posta.mu.edu.tr

1 Introduction

In this report, I am going to give informations about the search module I created and compare it and Apache Solr.

2 About Search Module

In my own search module, I have a dictionary representing the database which contains all the words in which documents. First, using this database, I find out which of the words in the searched word set are in or not in the database. After printing a text for words that are not in the database, I throw the words that in database to a list and I use this list in the rest of my code. Then, I use the algorithm that I wrote myself without using any sources to find the documents where the words are common. If I need to explain this algorithm with an example; Let's say the searched word set contains 5 words and all of them are available in the database. I can find all subsets of these 5 words using the algorithm. It can be seen more easily in the examples I made for comparison. After finding all the common documents in which the words are mentioned, I collect the specific values of the words according to the words in the document and I get numbers that show how much the document is related to the searched words. After sorting the results according to their relevance, I order the ones with the most words among themselves in a more relevant way.

3 Results

I searched the same sets of words both in my own search module and in Apache Solr. I compared if the document numbers in my first 10 results are within the top 10 results of Apache.

A. "I want to go to a vacation."

My Search Module

```
Enter a query: I want to go to a vacation.

'i - want - to - go - a' -> not in DB

10 search results were found.
+-------+----------+----------------------+
|  Doc  |  Terms   |      Relevance       |
+-------+----------+----------------------+
| 25.20 | vacation |  0.1417003966188046  |
| 24.38 | vacation | 0.05282135320388474  |
| 15.49 | vacation | 0.04225708256310779  |
|  23.8 | vacation | 0.03997291593807494  |
|  30.4 | vacation | 0.031978332750459956 |
| 24.28 | vacation | 0.028442267109784092 |
| 15.45 | vacation | 0.028306179707344933 |
| 25.17 | vacation | 0.02506776084252157  |
|  5.3  | vacation | 0.024146904321775884 |
| 16.19 | vacation | 0.023759002244317633 |
+-------+----------+----------------------+
```

Apache Solr

25.20 23.8 15.49 24.38 24.28 5.3 30.4 15.45 16.19 25.17

Compare

```
25.20 24.38 15.49  23.8  30.4 24.28 15.45 25.17  5.3 16.19
25.20 24.38 15.49  23.8  30.4 24.28 15.45 25.17  5.3 16.19
```

# of same results → 10/10

Accuracy = 1.0

B. Great news! The war is over, we are free.

My Search Module

```
Enter a query: Great news! The war is over, we are free.
```

```
'the - is - over - we - are' -> not in DB

192 search results were found.
+-------+----------------------------------+----------------------+
|  Doc  |              Terms               |      Relevance       |
+-------+----------------------------------+----------------------+
|  1.14 | ('free', 'great', 'news', 'war') | 0.07619160686766954  |
|  15.1 | ('free', 'great', 'news', 'war') | 0.043980000499888025 |
|  19.6 | ('free', 'great', 'news', 'war') | 0.01972030479267399  |
|  7.37 |     ('free', 'news', 'war')      | 0.06524335598135395  |
|  1.43 |     ('free', 'great', 'war')     | 0.06134714413591455  |
| 30.17 |     ('great', 'news', 'war')     | 0.06103635043684839  |
| 20.19 |     ('free', 'great', 'war')     | 0.059751638199581486 |
|  3.25 |     ('free', 'news', 'war')      | 0.058865269956655225 |
|  6.27 |     ('great', 'news', 'war')     | 0.05772655445280575  |
| 23.17 |    ('free', 'great', 'news')     | 0.056837387204287756 |
|  5.49 |     ('free', 'great', 'war')     | 0.05623427875663336  |
|  13.1 |    ('free', 'great', 'news')     | 0.055536781379439897 |
|  1.6  |    ('free', 'great', 'news')     | 0.05115592341872342  |
|  13.2 |    ('free', 'great', 'news')     | 0.046042086596456824 |
|  1.34 |    ('free', 'great', 'news')     | 0.040926637072762534 |
| 20.20 |    ('free', 'great', 'news')     | 0.04082956915845322  |
| 14.36 |    ('free', 'great', 'news')     | 0.040776314697842334 |
|  5.11 |    ('free', 'great', 'news')     | 0.03938092750285913  |
| 20.42 |    ('free', 'great', 'news')     | 0.03914721806959894  |
|  1.38 |    ('free', 'great', 'news')     | 0.03872628024089357  |
|  15.8 |    ('free', 'great', 'news')     | 0.03659782308569594  |
|  7.46 |     ('great', 'news', 'war')     | 0.03571228133249095  |
|  1.21 |    ('free', 'great', 'news')     | 0.031399772791600396 |
|  4.5  |     ('great', 'news', 'war')     | 0.028584853330328244 |
| 10.26 |     ('great', 'news', 'war')     | 0.028245049953114408 |
|  3.46 |    ('free', 'great', 'news')     | 0.02769285517036979  |
| 10.17 |     ('free', 'great', 'war')     | 0.026776805305634574 |
|  10.9 |    ('free', 'great', 'news')     | 0.026562742267186627 |
| 10.33 |     ('free', 'great', 'war')     | 0.026045889845091603 |
|  20.9 |    ('free', 'great', 'news')     | 0.02523905787679272  |
| 20.35 |    ('free', 'great', 'news')     | 0.024464853647443253 |
| 10.18 |     ('free', 'great', 'war')     | 0.01804828268778634  |
| 27.27 |        ('great', 'news')         | 0.13836364885612806  |
|  4.35 |        ('great', 'free')         | 0.11968817466535078  |
|  27.8 |        ('great', 'news')         | 0.11323692169677557  |

           ---results are hidden because of the length---

| 10.16 |         ('great', 'war')         | 0.011100365361885123 |
|  8.31 |        ('great', 'news')         | 0.010612487866248996 |
|  10.1 |        ('great', 'free')         | 0.00972864346110339  |
+-------+----------------------------------+----------------------+
```

Apache Solr

1.14 3.25 7.37 6.27 30.17 15.1 30.12 1.43 13.1 5.49

Compare

```
1.14 3.25 7.37 6.27 30.17  15.1 30.12 1.43 13.1  5.49
1.14 15.1 19.6 7.37  1.43 30.17 20.19 3.25 6.27 23.17

# of same results → 7/10
Accuracy = 0.7
```

C. rich gold man a party last night

My Search Module

```
Enter a query: rich gold man a party last night

'a - last' -> not in DB

223 search results were found.
+-------+-----------------------------------+----------------------+
|  Doc  |               Terms               |      Relevance       |
+-------+-----------------------------------+----------------------+
|  7.14 | ('man', 'night', 'party', 'rich') | 0.09790495493794443  |
|  4.13 |     ('man', 'night', 'party')     | 0.10092598690853145  |
|  7.13 |     ('man', 'night', 'party')     | 0.09038599460250166  |
|  7.35 |     ('man', 'night', 'party')     | 0.08078095941072953  |
| 15.50 |      ('gold', 'man', 'rich')      | 0.07474911889239255  |
|  2.37 |     ('man', 'night', 'party')     | 0.07244551340027769  |
| 14.22 |     ('man', 'night', 'party')     | 0.06790364827345975  |
| 10.35 |     ('man', 'night', 'party')     | 0.06676941325498395  |
| 10.27 |     ('man', 'night', 'party')     | 0.06676120100100276  |
|  4.14 |     ('night', 'party', 'rich')    | 0.06668501864682684  |
|  1.26 |     ('man', 'night', 'party')     | 0.05578043556975844  |
| 10.31 |     ('man', 'night', 'party')     | 0.04602241931339364  |
|  13.2 |     ('man', 'night', 'party')     | 0.045552048447721094 |

           ---results are hidden because of the length---

|  10.5 |      ('man', 'party', 'rich')     | 0.019440381091840597 |
|  2.34 |         ('party', 'night')        | 0.16937896573930306  |
| 23.32 |          ('man', 'party')         | 0.16244190971185057  |

           ---results are hidden because of the length---

|  7.46 |         ('party', 'night')        | 0.012330335733965872 |
|  10.7 |          ('man', 'party')         | 0.011799703361399017 |
|  10.6 |          ('man', 'party')         | 0.010016027271885213 |
|  10.1 |          ('man', 'party')         | 0.009769886715109206 |
+-------+-----------------------------------+----------------------+
```

Apache Solr

7.14 15.50 10.19 4.14 7.13 26.40 26.12 14.13 4.36 7.35

Compare

```
7.14 15.50 10.19 4.14  7.13 26.40 26.12 14.13  4.36 7.35
7.14  4.13  7.13 7.35 15.50  2.37 14.22 10.35 10.27 4.14

# of same results → 5/10
Accuracy = 0.5
```

4 Conclusion

The main reason why the results are not compatible is the Tokenizer methods. In my method, only the words are converted to lowercase and processing is done. Apache Solr uses a much more advanced method and can search for more detailed information by finding the origins, types and suffixes of the words. This is the reason why the searches made using non-gravity words are higher in accuracy.