# Assignment 2

Juan Florez

Eduardo Jaime

## Setup

We indexed Lucene version 5.2.1 (23568 documents) and Rhino 1.7R5 (3718 documents). A custom script was used to preprocess the documents, and Lucene was used to index and perform retrieval.

## Results

### Lucene

Title only:

| Query | Title | | | | | | |
|---|---|---|---|---|---|---|---|
| | @5 | | @10 | | @20 | | Effectiveness |
| | Precision | Recall | Precision | Recall | Precision | Recall | |
| 1 | 0% | 0% | 0% | 0% | 5% | 100% | 13 |
| 2 | 0% | 0% | 0% | 0% | 0% | 0% | None |
| 3 | 0% | 0% | 0% | 0% | 0% | 0% | None |
| 4 | 0% | 0% | 0% | 0% | 0% | 0% | None |
| 5 | 0% | 0% | 0% | 0% | 0% | 0% | None |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Average | 0.00% | 0.00% | 0.00% | 0.00% | 1.00% | 20.00% | 13.0 |
| Median | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 13.0 |

Description only:

| Query | Description | | | | | | |
|---|---|---|---|---|---|---|---|
| | @5 | | @10 | | @20 | | Effectiveness |
| | Precision | Recall | Precision | Recall | Precision | Recall | |
| 1 | 0% | 0% | 0% | 0% | 5% | 100% | 15 |
| 2 | 0% | 0% | 0% | 0% | 0% | 0% | None |
| 3 | 0% | 0% | 0% | 0% | 0% | 0% | None |
| 4 | 0% | 0% | 0% | 0% | 5% | 100% | 14 |
| 5 | 0% | 0% | 0% | 0% | 0% | 0% | None |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Average | 0.00% | 0.00% | 0.00% | 0.00% | 2.00% | 40.00% | 14.5 |
| Median | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 14.5 |

Title and description:

| Query | Title and Description | | | | | | |
|---|---|---|---|---|---|---|---|
| | @5 | | @10 | | @20 | | Effectiveness |
| | Precision | Recall | Precision | Recall | Precision | Recall | |
| 1 | 0% | 0% | 10% | 100% | 5% | 100% | 7 |
| 2 | 0% | 0% | 0% | 0% | 0% | 0% | None |
| 3 | 0% | 0% | 0% | 0% | 0% | 0% | None |
| 4 | 0% | 0% | 0% | 0% | 5% | 100% | 14 |
| 5 | 0% | 0% | 0% | 0% | 0% | 0% | None |
| | | | | | | | |
| Average | 0.00% | 0.00% | 2.00% | 20.00% | 2.00% | 40.00% | 10.5 |
| Median | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 10.5 |

We believe the poor performance of the method on this corpus may be due to the corpus size. With over 20 thousand methods it is increasingly likely that many documents other than the relevant ones had a high textual similarity with each query. Additionally, most queries had a single document as a correct result, making it even harder to retrieve it among the first 20 results.

## Rhino

Title only:

| Query | Title | | | | | | |
|---|---|---|---|---|---|---|---|
| | @5 | | @10 | | @20 | | Effectiveness |
| | Precision | Recall | Precision | Recall | Precision | Recall | |
| 1 | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | 0 | None |
| 2 | 60.0% | 100.0% | 30.0% | 100.0% | 15.0% | 1 | 1 |
| 3 | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | 0 | None |
| 4 | 0.0% | 0.0% | 10.0% | 50.0% | 5.0% | 0.5 | 6 |
| 5 | 40.0% | 66.7% | 30.0% | 100.0% | 15.0% | 1 | 2 |
| | | | | | | | |
| Average | 20.00% | 33.33% | 14.00% | 50.00% | 7.00% | 50.00% | 3.0 |
| Median | 0.00% | 0.00% | 10.00% | 50.00% | 5.00% | 50.00% | 2.0 |

Description only:

| Query | Description | | | | | | |
|---|---|---|---|---|---|---|---|
| | @5 | | @10 | | @20 | | Effectiveness |
| | Precision | Recall | Precision | Recall | Precision | Recall | |
| 1 | 0 | 0 | 0.0% | 0.0% | 0.0% | 0.0% | None |
| 2 | 0.6 | 1 | 30.0% | 100.0% | 15.0% | 100.0% | 1 |
| 3 | 0 | 0 | 0.0% | 0.0% | 5.0% | 33.3% | 17 |
| 4 | 0 | 0 | 0.0% | 0.0% | 0.0% | 0.0% | None |
| 5 | 0 | 0 | 0.0% | 0.0% | 0.0% | 0.0% | None |
| | | | | | | | |
| Average | 12.00% | 20.00% | 6.00% | 20.00% | 4.00% | 26.67% | 9.0 |
| Median | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 9.0 |

Title and description:

| Query | Title and Description | | | | | | |
|---|---|---|---|---|---|---|---|
| | @5 | | @10 | | @20 | | Effectiveness |
| | Precision | Recall | Precision | Recall | Precision | Recall | |
| 1 | 0.0% | 0.0% | 0.0% | 0.0% | 5.0% | 9.1% | 18 |
| 2 | 60.0% | 100.0% | 30.0% | 100.0% | 15.0% | 100.0% | 1 |
| 3 | 0.0% | 0.0% | 10.0% | 33.3% | 5.0% | 33.3% | 9 |
| 4 | 20.0% | 50.0% | 10.0% | 50.0% | 5.0% | 50.0% | 4 |
| 5 | 0.0% | 0.0% | 0.0% | 0.0% | 5.0% | 33.3% | 18 |
| | | | | | | | |
| Average | 16.00% | 30.00% | 10.00% | 36.67% | 7.00% | 45.15% | 10.0 |
| Median | 0.00% | 0.00% | 10.00% | 33.33% | 5.00% | 33.33% | 9.0 |

# Total for both systems

Title only:

| | Title | | | | | | |
|---|---|---|---|---|---|---|---|
| | @5 | | @10 | | @20 | | Effectiveness |
| | Precision | Recall | Precision | Recall | Precision | Recall | |
| Average | 10.00% | 16.67% | 7.00% | 25.00% | 4.00% | 35.00% | 5.5 |
| Median | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 4.0 |

Description only:

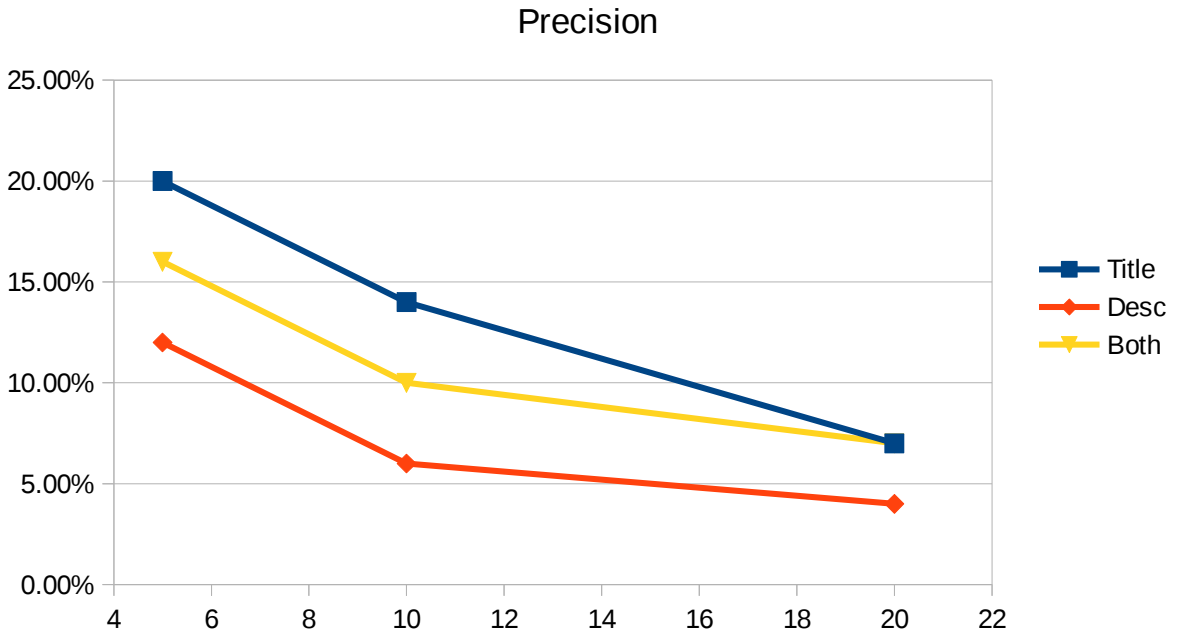
| | Description | | | | | | |
|---|---|---|---|---|---|---|---|
| | @5 | | @10 | | @20 | | Effectiveness |
| | Precision | Recall | Precision | Recall | Precision | Recall | |
| Average | 6.00% | 10.00% | 3.00% | 10.00% | 3.00% | 33.33% | 11.8 |
| Median | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 14.5 |

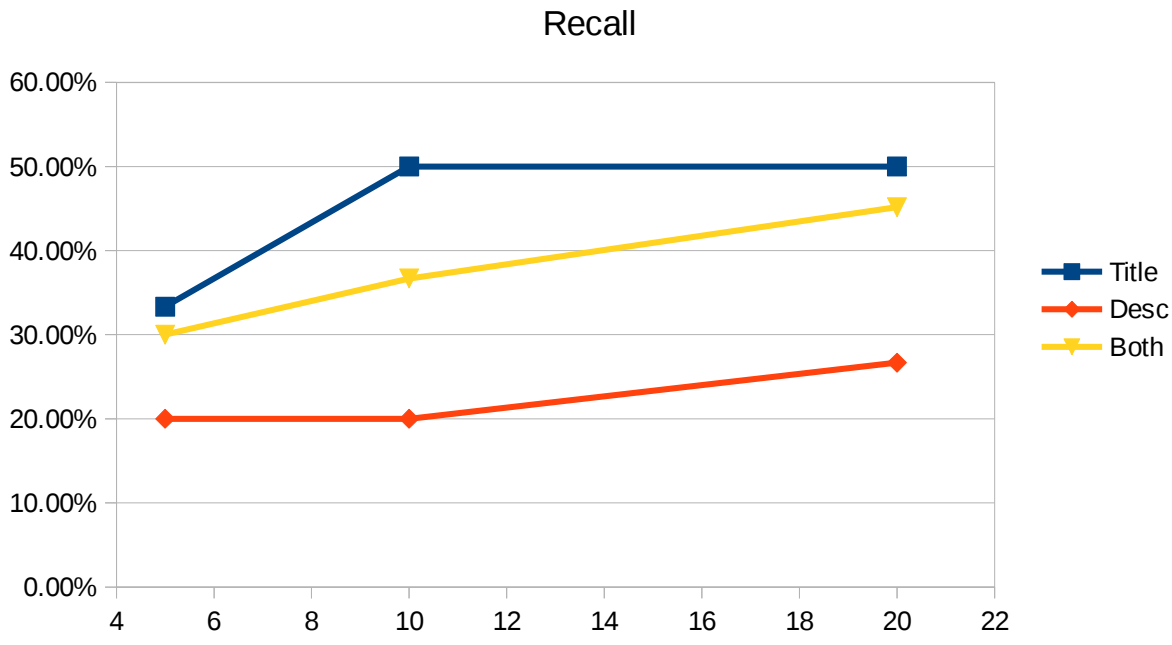
Title and description:

| | Title and Description | | | | | | |
|---|---|---|---|---|---|---|---|
| | @5 | | @10 | | @20 | | Effectiveness |
| | Precision | Recall | Precision | Recall | Precision | Recall | |
| Average | 8.00% | 15.00% | 6.00% | 28.33% | 4.50% | 42.58% | 10.1 |
| Median | 0.00% | 0.00% | 0.00% | 0.00% | 5.00% | 33.33% | 9.0 |

# Comparison of queries

Next, we present some statistics only about the Rhino results, since the ones from Lucene don't allow for a comprehensive analysis. The charts plot the average precision and recall over the 5 queries using each different type of query.




| Type of information | Average effectiveness over all queries |
|---|---|
| Title | 3 |
| Description | 9 |
| Both | 10 |

The results tell us using only titles works best in this case. This points to the title being a good summary of the bug report in most cases, and also suggests that more words are not always better. Bearing this in mind, it could be expected that applying some kind of summarization to the bug report could yield better results than simply using all words in it.

It can also be observed that the overall precision and recall of the approach are not very high, and this could be attributed to a diversity of factors. However, manually analyzing the results of the queries we realized that the approach consistently came close to the methods in the gold set, many times pointing to other methods in the same class. This clearly suggests that using simple TR can point the search in the right direction, but other more sophisticated approaches could be used then to improve the results.