# IR Assignment 2 - Report

**GROUP MEMBERS -**

**2017A7PS0184P** **TANMAY MOGHE**
**2017A7PS0224P** **AMRATANSHU SHRIVASTAVA**
**2016B5A70560P** **PARTH MISRA**

## PART-1 :

**File used -** AG/ wiki_40

**Query 1 -** "**As in the Five hoosier Painters**"

**No of Total Checked Documents -** 3410

| Top K docs | Score | Relevant? |
|---|---|---|
| 453 | 0.14509490047932058 | YES |
| 1266 | 0.09792048145742804 | NO |
| 22 | 0.09318015511770117 | NO |
| 1853 | 0.0920679046801944 | NO |
| 3421 | 0.08471177053171496 | NO |
| 1688 | 0.08185997385526618 | NO |
| 1761 | 0.0812744625671755 | NO |
| 1665 | 0.0812744625671755 | NO |
| 228 | 0.07924630290589128 | NO |
| 1480 | 0.0750831362372001 | NO |

**Query 2 -** “the modern re-interpretations of classical music”

**No of Total Checked Documents -** 3400

| **Top K docs** | **Score** | **Relevant?** |
|---|---|---|
| 1055 | 0.18097363400918084 | NO |
| 158 | 0.13431847193268856 | NO |
| 1395 | 0.11982177602218191 | YES |
| 2227 | 0.11065940839749591 | NO |
| 2103 | 0.10605685601861932 | NO |
| 1536 | 0.10435725864276822 | NO |
| 2902 | 0.10381718811397504 | NO |
| 2070 | 0.10320373409540515 | NO |
| 2443 | 0.10268699910226566 | NO |
| 249 | 0.10215178638225375 | NO |

**Query 3 -** “Highway construction”

**No of Total Checked Documents -** 229

| **Top K docs** | **Score** | **Relevant?** |
|---|---|---|
| 3498 | 0.38906153277780803 | NO |
| 3288 | 0.37961648079931404 | NO |
| 3298 | 0.3756852943664315 | NO |
| 3295 | 0.3756852943664315 | NO |
| 3514 | 0.35121071186458924 | NO |
| 3456 | 0.3423608527772028 | NO |

| | | |
|---|---|---|
| 3244 | 0.341687558776387 | NO |
| 3458 | 0.3363934577723619 | NO |
| 3097 | 0.3294548249093206 | NO |
| 3476 | 0.32490620424096894 | NO |

**Query 4 - "School of Drama"**

**No of Total Checked Documents -** 3517

| **Top K docs** | **Score** | **Relevant?** |
|---|---|---|
| 422 | 0.2618943100453689 | NO |
| 77 | 0.22936999933375904 | NO |
| 76 | 0.2251385005267035 | YES |
| 1495 | 0.22058261953074598 | NO |
| 2953 | 0.21644699212633328 | NO |
| 865 | 0.20695732288529847 | NO |
| 2585 | 0.20399844405617149 | NO |
| 1942 | 0.1940949799712458 | NO |
| 2058 | 0.1914977043892086 | YES |
| 290 | 0.190345853832252 | NO |

**Query 5 -** "**heart attack**"

**No of Total Checked Documents -** 160

| **Top K docs** | **Score** | **Relevant?** |
|---|---|---|
| 3337 | 0.23840159528767724 | NO |

| | | |
|---|---|---|
| 1044 | 0.21690756568193661 | NO |
| 1554 | 0.13240562329627978 | YES |
| 112 | 0.10524567694838673 | NO |
| 1950 | 0.10337374203569925 | YES |
| 182 | 0.10328641418985274 | YES |
| 2301 | 0.09754699285122778 | YES |
| 2806 | 0.09080811034596947 | YES |
| 630 | 0.08593961067042936 | NO |
| 245 | 0.08189026430953389 | NO |

**Query 6 -** “**InterContinental Restaurant**”

**No of Total Checked Documents -** 50

| **Top K docs** | **Score** | **Relevant?** |
|---|---|---|
| 1677 | 0.1602044684166896 | YES |
| 1180 | 0.07663245906874039 | NO |
| 95 | 0.06516313100294978 | NO |
| 577 | 0.05056367293781623 | NO |
| 2339 | 0.04567578146656186 | NO |
| 3053 | 0.045638525623719756 | NO |
| 3221 | 0.041757660343651214 | NO |
| 1883 | 0.04004165974982279 | NO |
| 2910 | 0.03986451485698695 | NO |
| 727 | 0.038744828106700206 | NO |

**Query 7 - “Chief Executive Officer”**

**No of Total Checked Documents -** 249

| Top K docs | Score | Relevant? |
|---|---|---|
| 3210 | 0.16320076477650775 | YES |
| 867 | 0.14701142726044325 | YES |
| 336 | 0.11667975837604694 | YES |
| 2966 | 0.10992287103284161 | NO |
| 2159 | 0.10114244805758311 | NO |
| 2964 | 0.09832439929041993 | NO |
| 3441 | 0.09806551943433839 | NO |
| 2856 | 0.09741271755188063 | NO |
| 3185 | 0.0948701737713282 | NO |
| 2038 | 0.08958588562121536 | NO |

**Query 8 - “medieval architecture”**

**No of Total Checked Documents -** 82

| Top K docs | Score | Relevant? |
|---|---|---|
| 2336 | 0.1266504172164324 | YES |
| 332 | 0.12322922025939545 | YES |
| 955 | 0.10157258586961192 | NO |
| 2723 | 0.09138799111638844 | YES |
| 802 | 0.08578670103564713 | NO |
| 2802 | 0.08171621563069045 | YES |
| 1204 | 0.07847868499617959 | NO |

| | | |
|---|---|---|
| 2811 | 0.0779897736720232 | YES |
| 1024 | 0.07202072146679961 | NO |
| 1747 | 0.07153003509534009 | NO |

**Query 9 -** “**university teacher**”

**No of Total Checked Documents -** 467

| **Top K docs** | **Score** | **Relevant?** |
|---|---|---|
| 1267 | 0.18657176786512833 | NO |
| 127 | 0.16537620463732122 | YES |
| 2481 | 0.13828635683728155 | YES |
| 252 | 0.13398558219852177 | YES |
| 2876 | 0.1277448944479284 | YES |
| 1754 | 0.12387898133751052 | NO |
| 2427 | 0.11916001745360645 | YES |
| 3057 | 0.1191180834677854 | YES |
| 2245 | 0.11682066205288566 | YES |
| 3511 | 0.1160820720798712 | YES |

**Query 10 -** “**seven quality coach**” (ambiguous)

**No of Total Checked Documents -** 329

| **Top K docs** | **Score** | **Relevant?** |
|---|---|---|
| 68 | 0.207805907294367 | SOMEWHAT |
| 1076 | 0.13113833524331556 | NO |
| 1754 | 0.1234175640890745 | NO |

| | | |
|---|---|---|
| 1442 | 0.11578844409676164 | NO |
| 1619 | 0.11161599663132674 | NO |
| 28 | 0.10809777561352757 | NO |
| 3180 | 0.10741571599651598 | NO |
| 3445 | 0.10659096869161729 | NO |
| 180 | 0.1060372470842922 | NO |
| 1314 | 0.09998977908425799 | NO |

## PART-2 :

1. **Problem -** In our system, every document is a candidate (to be checked for cosine similarity) if it contains at least one of the query terms

2. **Improvement 1** - **ONLY HIGH IDF TERMS**
   (Terms with idf lower than 2.00 are dropped from the query)

   **Improvement 2 - DOCS CONTAINING MANY QUERY TERMS**
   (Documents which contain at least 70% of the query terms are taken to check)

3. **Improvement 1** - Terms like 'the' and 'as' don't alter the ranking of the system much. Rather, they add many documents to the no of documents checked for similarity as they occur too frequently. Here, we consider only terms with high idf. Low idf terms from the query are dropped, thus decreasing the no of total documents to be checked.

   **Improvement 2** - Many docs containing just one term are not relevant. With phrases like "New York" which are bound to occur together, our previous system doesn't work as expected. Taking more than 70% of these 2 words, will mean that only docs containing BOTH these words will be taken into account and hence, more relevant documents will be checked.

4. **Proximity Search** is still a problem after this improvement. This improvement does not work for the following cases and the like -

   Query - "**Outspoken Minds**" will retrieve the documents which contain the maximum number of times the words "outspoken" and "minds" whereas we wanted the document where both of these words occur together, even once or twice.

5. Actual impact on the number of documents checked for cosine similarity -

   (3 test queries - demonstration)

**Improvement query 1-**

"**As in the Five hoosier Painters"**
No of docs checked without improvements - 3410
No of docs checked with improvement 1- 359
No of docs checked with improvement 2- 4
No of docs checked with both improvements - 1

(Best case scenario result !)

(only one doc checked and hence, only one returned)

| **Top K docs** | **Score** | **Relevant?** |
|---|---|---|
| 453 | 0.11684739176220117 | YES |

**Improvement query 2-**

"**the modern re-interpretations of classical music"**
No of docs checked without improvements - 3400
No of docs checked with improvement 1- 430
No of docs checked with improvement 2- 7
No of docs checked with both improvements- 7

| **Top K docs** | **Score** | **Relevant?** |
|---|---|---|
| 1395 | 0.09256531771135673 | YES |
| 249 | 0.07291422739098215 | NO |
| 446 | 0.04668260814517258 | YES |
| 1921 | 0.04454834546738897 | NO |
| 2394 | 0.03594974095550584 | YES |
| 895 | 0.03254713117437016 | NO |
| 3281 | 0.024273047721452723 | YES |

**Improvement query 3-**

"**Chief Executive Officer"**
No of docs checked without improvements - 249
No of docs checked with improvement 1- 249 (since no low idf term is present)
No of docs checked with improvement 2- 13
No of docs checked with both improvements - 13

| **Top K docs** | **Score** | **Relevant?** |
|---|---|---|
| 3210 | 0.16320076477650775 | YES |
| 867 | 0.1470114272604433 | YES |
| 336 | 0.11667975837604694 | YES |
| 2966 | 0.10992287103284161 | NO |
| 2038 | 0.0895858856212153 | NO |
| 167 | 0.08270129994191028 | YES |
| 2373 | 0.06992177758218406 | NO |
| 1401 | 0.06486127976799042 | YES |
| 2499 | 0.061944070510219995 | NO |
| 1418 | 0.06155479673173009 | YES |

**RESULT -** Not only has the improvements (both) reduced the number of documents to be checked for similarity but the top K retrieved documents now are much more Relevant (ratio of Relevant docs in the top K has increased) as you can see from the tables.