CAMPUS TOUR			
Rank	Authority	Hub	PageRank
1.	924	24105	24117
2.	24024	24138	24190
3.	23671	24092	24198
4.	24082	24186	24053
5.	24061	24113	24061
6.	24192	24166	24082
7.	24191	24100	24052
8.	2283	24065	24191
9.	24052	24086	24192
10.	24105	24144	24112

TRANSCRIPTS			
Rank	Authority	Hub	PageRank
1.	3900	3854	22957
2.	3984	3855	1055
3.	4353	3850	3850
4.	3910	3858	21983
5.	4251	3852	21982
6.	3957	3859	1021
7.	4105	4353	999
8.	3875	3857	1051
9.	3830	3853	852
10.	4103	3856	3853

ADMISSIONS			
Rank	Authority	Hub	PageRank
1.	1082	1081	938
2.	1086	1082	935
3.	1084	1086	1075
4.	1080	1084	992
5.	1085	1080	963
6.	1089	1085	941
7.	1095	1089	936
8.	1090	1095	937
9.	1083	1090	939
10.	1088	1083	1081

EMPLOYEE BENEFITS			
Rank	Authority	Hub	PageRank
1.	4580	4570	4599
2.	4664	4552	4543
3.	4543	4543	4591
4.	4604	4580	223
5.	4582	4501	4592
6.	4434	4573	4590
7.	751	4564	4438
8.	4466	4529	787
9.	781	4510	4627
10.	4587	4555	4464

PARKING DECAL			
Rank	Authority	Hub	PageRank
1.	2283	2285	2406
2.	2243	2286	649
3.	2260	2283	4595
4.	2272	2280	648
5.	2247	2279	644
6.	2237	2284	643
7.	2246	2281	645
8.	2234	2282	2282
9.	2273	2287	2287
10.	2280	2243	642

SRC			
Rank	Authority	Hub	PageRank
1.	24420	24342	14460
2.	24342	24356	18370
3.	24356	24343	14556
4.	24343	24348	20289
5.	24361	24355	20286
6.	24355	24346	20211
7.	24346	24349	20215
8.	24349	24362	20204
9.	24362	24360	20288
10.	24360	24357	21278

Table 1.

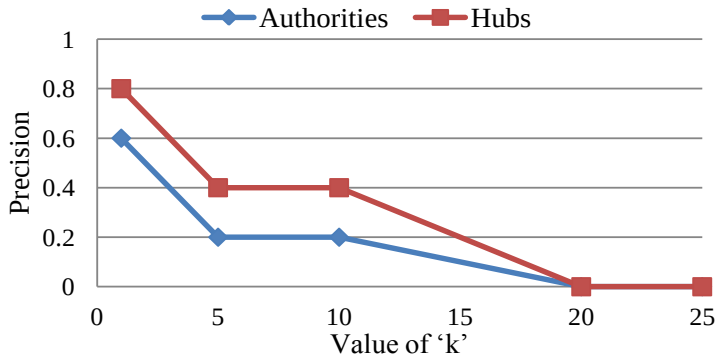
Q: Top 10 authorities and hubs for K = 10 for the sample queries.

A: Please refer table 1.

Q: What effect does changing the root set size have on quality (relevance) of results? Why?

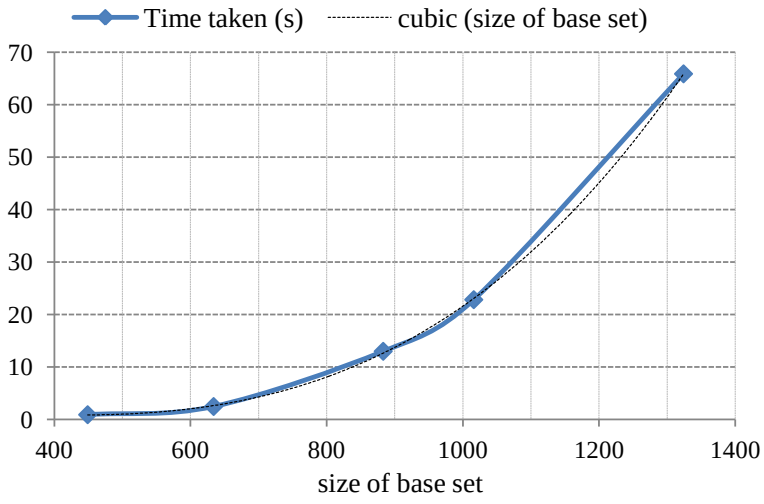
A: The relevance decreases when the root size is increased. The rate of decrease will depend on the change in base set size. For query "*Computing Commons*" the precision for Authorities and for Hubs was 1 till the value of $k < 33$. As the value of root set is increased the relevance decreases because pages with no similarity to the query will also be accounted for ranking.

For query "*admission*" plot1 shows the precision vs size of root set curve.



Plot1: Precision vs Size of root set. Query "*admission*"

Q: What effect does changing the root set size have on time taken? Why?
A: As the root set is increased, the base set might increase according to the query result size.
From the example query "html" (from table 2) we can observe the following:
(Base set size for root set 20)/ (Base set size for root set 10) = R_b = 634/449 = 1.41
(Total time for root set 20)/ (Total time set size for root set 10) = R_t = 2441/880 = 2.77
$(R_b)^3 = (1.41)^3 = 2.80$
Which is approximately a constant integer multiple (1) of the time ratio 'R_t' which implies that the time taken will be cubic on size of base set.
Time taken $\simeq$ c.(size of base set)^3 , where 'c' is some constant.
Which is true theoretically also because matrix multiplication is $O(n^3)$.
The same observation can be made for other values.
It can also be observed from plot2 (below).



Plot2: for varying value of root set for query "html"

Q: Compare the time taken by various phases of the algorithm.
A: Refer table 2 for time taken in various phases.
The dominant phase is the matrix multiplication which would be cubic (as can also be seen in the table)

	HTML				
Root set size	10	20	30	40	50
Base set size	449	634	883	1016	1324
Base set created in	13	31	7	11	9
Adjacency matrix computed in	5	8	17	14	47
$A^T.A$ and $A.A^T$ computed in	**826**	**2360**	**12881**	**22717**	**65690**
Aut and Hub Vector computed in	33	40	59	66	124
Total Time	880	2441	12967	22811	65872

Table 2 *(*time in milliseconds)*

Q: Pagerank / similarity output (sample queries below. use w = 0.4)?
A: Please refer to table1 for the output.

Q: Which results do you think are more relevant - TF/IDF, Authorities, Hubs or Pagerank? Why? Support your answer with examples or statistics?
A: Assumption – size of root set used for Authorities and Hubs is 10. Value of 'c' and 'w' for PageRank is 0.8 and 0.4 respectively.

- **Campus Tour**:
  - The top results of ***TF/IDF*** and ***pageRank*** are same. The first result is the main page for online campus tour for Tempe, which according to the query is very relevant. The second, third and fourth results are also main pages for online campus tour for Polytechnic campus, West campus and Downtown Phoenix campus respectively. And they are *very relevant* to the query.
  - ***Authorities***: The first result is not opening. The second result is related to ASU scholar student profile which is irrelevant to the query. Third result is about welcome week for new students and families, which again is *irrelevant* to the query. The fourth result is "ASU campus tour stops by building code" which is a related to the query and is still relevant but not compared to results given by TF/IDF and pageRank.
  - ***Hubs***: The top 4 results are tour links of H.B. Farmer Building, Moeur Building, Danforth Chapel and Wilson Hall which according to me are irrelevant as compared to results given by TF/IDF and pageRank.
- **Transcripts**:
  - Again top results of ***TF/IDF*** and ***pageRank*** are same. The first result is about "how to order official and unofficial transcripts" which is according to me is quite relevant. The second and third results are related to transcript requirement for undergraduate and graduate admission procedure. These are again valid results as the query is not specific. The fourth result is regarding "unofficial transcript evaluation for ATP candidates only" and is again relevant as the query is not very specific.

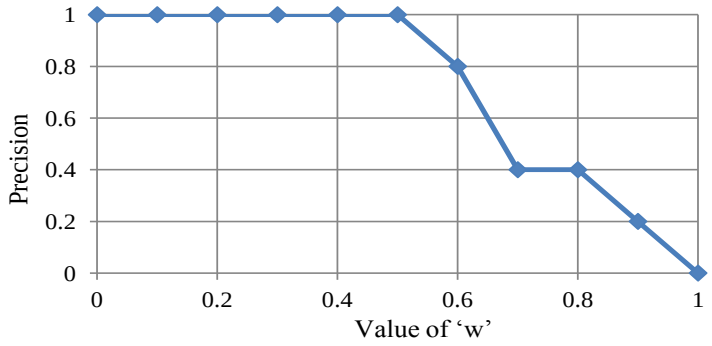
- **Authorities**: The first and second results are about "Current graduate student resources at ASU" and "Financial Support for graduate students", and they don't talk about transcripts at all. They are irrelevant. Third and fourth results give "division of graduate studies website index" and "diversity" which is totally irrelevant to the query.
- **Hubs**: The top 4 results are regarding "graduate admissions" and talks about "Requirements for F1/J1 applicants", "FAQ", "checklist" and "Issuance of I-20" respectively. The third result talks about transcript requirement and other details for graduate admission. This result is still relevant to the query and is also same as 3rd result of pageRank but the other results are irrelevant.

- **Admissions**:
  - Top results for ***TF/IDF*** and ***pageRank*** are more or less same: Top 4 results are under category of undergraduate admissions. They are "Contact details", "Counselor and advisors", "Admissions appointment" and "non-degree checklist" respectively. The first and third results are a little related to the query as per my expectation but second and fourth are irrelevant.
  - Top results for ***Authorities*** and ***Hubs*** are quite same: 4 top results are regarding "why choose ASU" fall under "undergraduate admissions" category. Each link talks about "location", "student story", "another student story" and "traditions and culture". My expectation was to get admission requirement pages for different programs. And according to me they are irrelevant.
- **SRC**: (According to current ASU website and Google results, SRC stands for Student Recreational Complex and I was directed to Sun Devil Fitness Complex)
  - ***TF/IDF***: The top results are for ASU photo gallery and are "Sun Devil 101: Welcome Week", "2004 in Review", "Futbol vs Football" and "Bird's-Eye view for ASU". As per the expectation mentioned earlier these links are totally irrelevant.
  - ***PageRank***: The first three results are regarding libraries at ASU and are "Contact", "Help Desk" and "Employment". These are again irrelevant according to the expectation. Fourth result is about "Publications" and mentions about different magazine at ASU. Again not relevant.
  - ***Authorities and Hubs***: Top results of both authorities and hubs are similar to results by TF/IDF and are pointing to photo gallery. Again irrelevant.
  - ***Why***?? SRC is an abbreviation but the search engine is checking the occurrence of html tag "src" in the page. That is why the top results for TF/IDF, Authorities, and Hubs were related to photo gallery. Note that it is a very frequent term, so it will have very low IDF value therefore the similarity value will be small. For pageRank the importance measure dominates and that is why the results also change but still they are irrelevant according to my expectations.

My Conclusion: According to me, in most case pageRank and IDF results are similar and are the relatively more relevant to the query as compared to Authorities and Hubs.

Q: What effect does varying w have on the relevance of the results? Why?

A: The effect of relevance can be seen from the plot3 below. As the value of 'w' increases the precision monotonically decreases. This is because as the value of 'w' increases the importance measure will dominate over similarity values, so high page ranked pages will become top results and similarity to the query will become a negligible quantity and hence relevance to the query is lost. This plot is for query "memorial union".



Plot3: Precision vs 'w'. Query "memorial union"

Value of 'w'	# of relevant results (among top 5)
0.0	5
0.1	5
0.2	5
0.3	5
0.4	5
0.5	5
0.6	4
0.7	2
0.8	2
0.9	1
1.0	0

Q: What effect does varying c have on the relevance of the results? Time taken? Why?

A: For c = 0, the page rank value for all the documents will be same. For different values of 'c' the relevance doesn't change. The results for the sample queries, turns out to be similar for all values of 'c'. The number of iterations and thus the time taken to converge changes and is mentioned in the table below. The relevance doesn't change because relative page rank values don't change much. For smaller values of 'c' the reset matrix dominates in M* matrix, so the deviation (maximum value in matrix – average value) will be low and hence convergence will be achieved in lesser iterations. Whereas for higher values of 'c', the M+Z matrix dominates and the deviation will be more. Thus, require more iterations or time to converge.

'c'	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Iteration count	1	2	2	3	3	3	3	3	3	3	16
Time (*seconds*)	3	6	6	9	9	9	9	9	9	10	49

**delta for convergence is 0.001*

Q: Did your PageRank computation converge? How many iterations did it take? How much memory did it take?

A: Taking the convergence condition to be:

for all abs($PR_i - PR_{i-1}$) ≤ delta, where PR_i is the PageRank vector computed in i^{th} iteration.

For different values of delta the iterations and time taken is shown in table below.

	Delta				
	0.001	0.0001	0.00001	0.000005	0.000001
Converged	Yes	Yes	Yes	Yes	No
Number of iterations	3	12	17	18	207 (terminated)
Time Taken (seconds)	10	40	52	63	----

The major memory is used by two vectors of type double used to store the previous and current values of pagerank vector. A boolean vector for the links in a row which is being computed. The length of all three are corpus size.

So, total memory = (2*8(byte) + 1*1(byte))*25054 = 425,918 bytes ≃ 0.43 MB

Q: Which document had the highest PageRank? Does that make sense?
A: The highest pageRank is for document **9048**. This makes sense because it has the largest number of citations which is equal to 7251. The document with second largest citations with value 3047 is 14460 which have the second highest pageRank value. Also, the number of links from 9048 is zero.

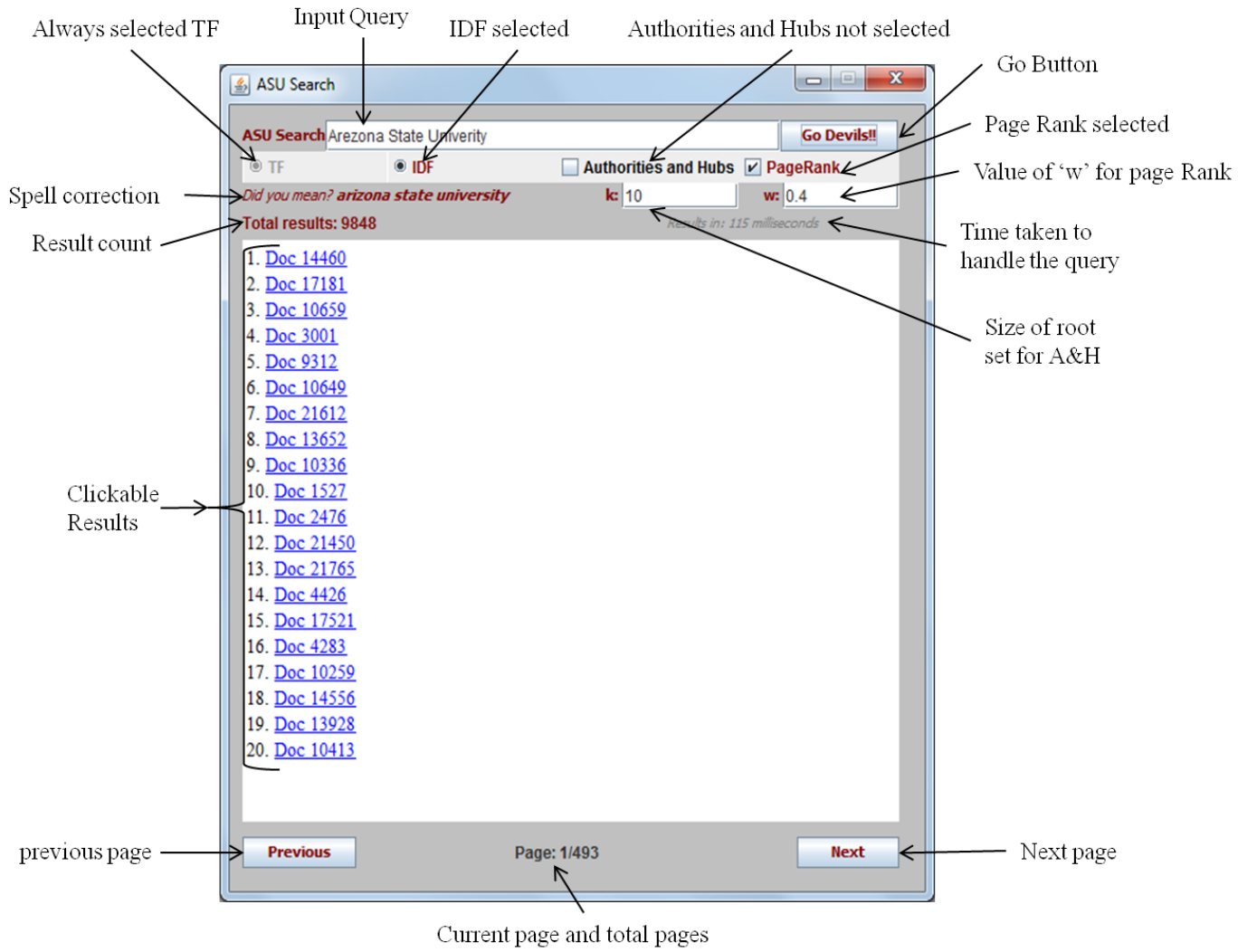
Q: Is there any correlation between high authority values and high PageRank values? What about hub values and PageRank values? Does this make sense?
A: Observations from the table below: The page rank values for the authorities are in descending order which makes sense because if a document has more citations then the page rank value will be higher whereas for hubs there is no correlation between page rank values and the ranking.

Campus Tour			
Top 5 Authorities	Page Rank value	Top 5 Hubs	Page Rank value
924	0.0148	24105	1.85×10^{-4}
24024	0.0155	24138	1.85×10^{-4}
23671	0.0208	24092	1.64×10^{-4}
24082	0.0140	24186	1.25×10^{-4}
24061	0.0140	24113	1.91×10^{-4}
Bottom 5 Authorities	Page Rank value	Bottom 5 Hubs	Page Rank value
24189	2.4×10^{-4}	24190	0
24117	0	924	0.0148
24053	0	24024	0.0155
24198	0	24066	3.54×10^{-4}
24190	0	24073	6×10^{-5}

Extra Implementation and Analysis:

- **Tolerant dictionary/Spell Corrector:**
  Tolerant dictionary using k-gram Jaccard similarity and IDF is implemented. '*SpellCorrector.java*' implements this functionality. It is an executable application and based on value of "k", it creates a '*spellIndex*' and '*indexMap*' and stores them in '*spellIndex.txt*' and '*termIndex.txt*' respectively. The '*spellIndex*' contains all the k-gram strings and set of index of the terms which contain this k-gram string. The '*indexMap*' gives the term corresponding to an index in '*spellIndex*'. A begin character '*#*' and end character '*$*' is also used for each term.
  When the search engine application is launched these files are read and respective indexes are populated. This takes around 500 milliseconds on my machine. When a query is entered, the 'begin' and 'end' characters are added to the query term. For each k-gram string of this query term the set of possible corpus terms is populated and then similarity for all these terms is computed as follows:
  $\text{Similarty}(q, t_i) = (\text{Jaccard similarity})/(1 + \text{IDF}(t_i))$, where Jaccard similarity = (union/intersection)
  The term with the highest similarity is returned as the corrected word.
  Interesting queries which can be tried: 'arezona', 'arzona', 'univerity'.

- **GUI**

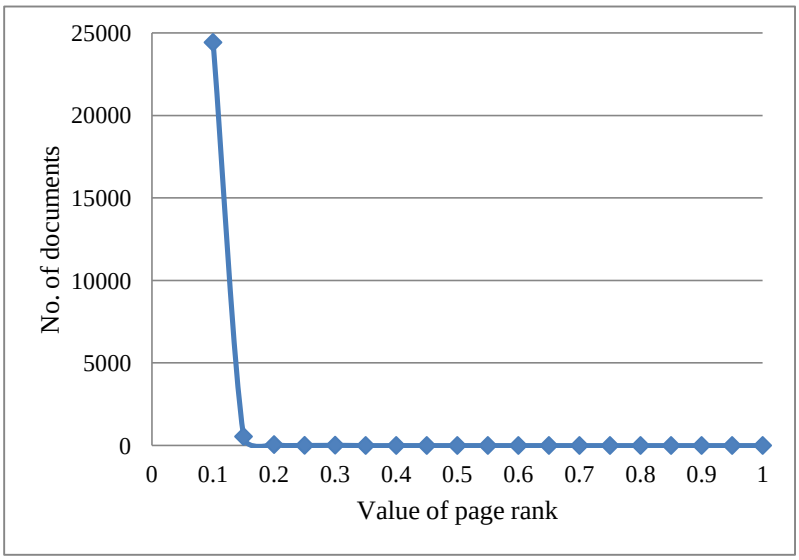


- **Prioritized Page Rank:** It is implemented but currently not in use. The following observations were made.

	Delta				
	0.001	0.0001	0.00001	0.000001	0.0000001
Converged	Yes	Yes	Yes	Yes	Yes
Number of iterations	3	6	13	24	34
Time Taken (seconds)	5	6	9	15	35

It converges for smaller values of delta also and time taken is remarkably less as compared to normal page rank computation.

- **Power law nature of the corpus:** The computed pageRank values were rounded off such that they lie in set of smaller ranges and the number of documents in these smaller ranges is computed. The plot below is between these smaller ranges and the count of documents. The curve shows power law behavior of the web pages.



NOTE: the '*spellIndex.txt*' and '*termIndex.txt*' should be present in under irs13 folder for the code to work.
The result3 folder which contains all the html files should also be present under the irs13 folder for the clickable links to work.