

## HOMEWORK 3

**Q1:**

**QUERY VECTOR REPRESENTATION FOR W1:**

1 aeroelastic:0.2512 aircraft:0.2512 constructing:0.3268 heated:0.3268 high:0.2512 law:0.3268 model:0.3268 must:0.3268 obeyed:0.3268 similarity:0.3268 speed:0.2512  
2 aeroelastic:0.3255 aircraft:0.3255 associated:0.4234 flight:0.4234 high:0.3255 problem:0.1959 speed:0.3255 structural:0.4234  
3 composite:0.3833 conduction:0.3833 far:0.3833 heat:0.2946 problem:0.1774 slab:0.3833 so:0.3833 solved:0.3833  
4 assumption:0.2452 based:0.2452 can:0.1317 chemical:0.1885 chemically:0.2452 criterion:0.2452 developed:0.2452 empirically:0.2452 equilibrium:0.2452 flow:0.0859 gas:0.1885 instantaneous:0.2452 local:0.2452 mixture:0.2452 on:0.1317 reacting:0.2452 show:0.2452 simplifying:0.2452 solution:0.1885 to:0.0985 validity:0.2452  
5 aerodynamic:0.326 applicable:0.4242 chemical:0.326 hypersonic:0.4242 kinetic:0.4242 problem:0.1963 system:0.4242 to:0.1705  
6 behaviour:0.3563 couette:0.3563 do:0.3563 experimental:0.2739 flow:0.1249 guide:0.3563 theoretical:0.3563 to:0.1432 turbulent:0.3563 we:0.3563  
7 angle:0.242 at:0.242 attack:0.242 available:0.1417 distribution:0.2027 equivalent:0.2637 forebody:0.3821 lower:0.2637 ogive:0.3821 possible:0.2027 pressure:0.242 relate:0.2637 surface:0.2637 to:0.1536 zero:0.2637  
8 angle:0.2079 approximate:0.3283 at:0.2079 attack:0.2079 available:0.1764 body:0.1764 dash:0.4756 exact:0.3283 method:0.3283 predicting:0.3283 presently:0.3283 pressure:0.2079  
9 flow:0.1482 heat:0.325 internal:0.4228 on:0.2272 paper:0.325 slip:0.4228 study:0.4228 transfer:0.4228  
10 air:0.325 available:0.1746 density:0.325 enthalpy:0.325 gas:0.2498 over:0.325 property:0.2498 range:0.325 real:0.325 transport:0.325 wide:0.325  
11 analytical:0.3306 approximation:0.3306 blast:0.3306 find:0.3306 newtonian:0.3306 possible:0.2541 problem:0.153 similar:0.3306 solution:0.2541 strong:0.3306 to:0.1329 wave:0.2541  
12 aerodynamic:0.3091 calculated:0.3091 can:0.216 channel:0.4021 effect:0.3091 flow:0.1409 ground:0.4021 machine:0.4021 performance:0.4021  
13 aileron:0.4667 basic:0.3587 buzz:0.4667 mechanism:0.4667 transonic:0.4667  
14 interaction:0.473 on:0.2541 paper:0.3635 shock:0.473 sound:0.473 wave:0.3635  
15 material:0.7543 photoelastic:0.5206 property:0.4001  
16 about:0.2894 body:0.2023 by:0.2894 calculated:0.2894 can:0.2023 computer:0.3765 efficiently:0.3765 electronic:0.3765 flow:0.1319 potential:0.2894 revolution:0.2384 transverse:0.2894  
17 about:0.2623 body:0.1834 can:0.1834 dimensional:0.4945 flow:0.1196 potential:0.2623 problem:0.2289 reduced:0.3413 revolution:0.2161 three:0.3413 to:0.1372 transverse:0.2623 two:0.3413  
18 angle:0.3146 at:0.3146 attack:0.3146 available:0.2669 body:0.2669

distribution:0.3818 experimental:0.3818 on:0.2669 pressure:0.3146 revolution:0.3146  
**19** basic:0.2041 combining:0.2655 consideration:0.2655 doe:0.2655 dynamic:0.2655  
effect:0.2041 entry:0.2655 exist:0.2655 good:0.2655 re:0.2655 realistic:0.2655  
relative:0.2655 result:0.2655 simplicity:0.2655 treatment:0.2655  
**20** anyone:0.2446 by:0.188 condition:0.2446 convection:0.2446 current:0.2446  
determined:0.2446 flow:0.0857 formally:0.2446 free:0.2446 general:0.2446 ha:0.2446  
heating:0.2446 induced:0.2446 influence:0.2446 joule:0.2446  
magnetohydrodynamic:0.2446 produced:0.2446 under:0.2446

## **QUERY VECTOR REPRESENTATION FOR W2:**

**1** aeroelastic:0.2512 aircraft:0.2512 constructing:0.3268 heated:0.3268 high:0.2512  
law:0.3268 model:0.3268 must:0.3268 obeyed:0.3268 similarity:0.3268 speed:0.2512  
**2** aeroelastic:0.3255 aircraft:0.3255 associated:0.4234 flight:0.4234 high:0.3255  
problem:0.1959 speed:0.3255 structural:0.4234  
**3** composite:0.3833 conduction:0.3833 far:0.3833 heat:0.2946  
problem:0.1774 slab:0.3833 so:0.3833 solved:0.3833  
**4** assumption:0.2452 based:0.2452 can:0.1317 chemical:0.1885  
chemically:0.2452 criterion:0.2452 developed:0.2452 empirically:0.2452  
equilibrium:0.2452 flow:0.0859 gas:0.1885 instantaneous:0.2452 local:0.2452  
mixture:0.2452 on:0.1317 reacting:0.2452 show:0.2452 simplifying:0.2452  
solution:0.1885 to:0.0985 validity:0.2452  
**5** aerodynamic:0.326 applicable:0.4242 chemical:0.326 hypersonic:0.4242  
kinetic:0.4242 problem:0.1963 system:0.4242 to:0.1705  
**6** behaviour:0.3563 couette:0.3563 do:0.3563 experimental:0.2739 flow:0.1249  
guide:0.3563 theoretical:0.3563 to:0.1432 turbulent:0.3563 we:0.3563  
**7** angle:0.2162 at:0.2162 attack:0.2162 available:0.1582 distribution:0.2264  
equivalent:0.2945 forebody:0.3414 lower:0.2945 ogive:0.3414 possible:0.2264  
pressure:0.2162 relate:0.2945 surface:0.2945 to:0.1372 zero:0.2945  
**8** angle:0.2167 approximate:0.3421 at:0.2167 attack:0.2167 available:0.1838  
body:0.1838 dash:0.3993 exact:0.3421 method:0.3421 predicting:0.3421  
presently:0.3421 pressure:0.2167  
**9** flow:0.1482 heat:0.325 internal:0.4228 on:0.2272 paper:0.325 slip:0.4228 study:0.4228  
transfer:0.4228  
**10** air:0.325 available:0.1746 density:0.325 enthalpy:0.325 gas:0.2498 over:0.325  
property:0.2498 range:0.325 real:0.325 transport:0.325 wide:0.325  
**11** analytical:0.3306 approximation:0.3306 blast:0.3306 find:0.3306  
newtonian:0.3306 possible:0.2541 problem:0.153 similar:0.3306  
solution:0.2541 strong:0.3306 to:0.1329 wave:0.2541  
**12** aerodynamic:0.3091 calculated:0.3091 can:0.216 channel:0.4021 effect:0.3091  
flow:0.1409 ground:0.4021 machine:0.4021 performance:0.4021  
**13** aileron:0.4667 basic:0.3587 buzz:0.4667 mechanism:0.4667 transonic:0.4667  
**14** interaction:0.473 on:0.2541 paper:0.3635 shock:0.473  
sound:0.473 wave:0.3635  
**15** material:0.6722 photoelastic:0.587 property:0.4512  
**16** about:0.2894 body:0.2023 by:0.2894 calculated:0.2894 can:0.2023

computer:0.3765 efficiently:0.3765electronic:0.3765 flow:0.1319  
potential:0.2894 revolution:0.2384 transverse:0.2894  
**17** about:0.2772 body:0.1938 can:0.1938 dimensional:0.4207 flow:0.1264  
potential:0.2772 problem:0.1947reduced:0.3607 revolution:0.2284 three:0.3607  
to:0.145 transverse:0.2772 two:0.3607  
**18** angle:0.3146 at:0.3146 attack:0.3146 available:0.2669 body:0.2669  
distribution:0.3818 experimental:0.3818on:0.2669 pressure:0.3146 revolution:0.3146  
**19** basic:0.2041 combining:0.2655 consideration:0.2655 doe:0.2655 dynamic:0.2655  
effect:0.2041 entry:0.2655exist:0.2655 good:0.2655 re:0.2655 realistic:0.2655  
relative:0.2655 result:0.2655 simplicity:0.2655 treatment:0.2655  
**20** anyone:0.2446 by:0.188 condition:0.2446 convection:0.2446 current:0.2446  
determined:0.2446 flow:0.0857formally:0.2446 free:0.2446 general:0.2446 ha:0.2446  
heating:0.2446 induced:0.2446 influence:0.2446 joule:0.2446  
magnetohydrodynamic:0.2446 produced:0.2446 under:0.2446

**Q2: Answered in retrieval\_summary\_w1.txt & retrieval\_summary\_w2.txt**

**Q3: Relevant and non-relevant documents for each query**

Query #	Relevant doc #		Non-relevant doc #	
	W1	W2	W1	W2
1	13, 486, 184, 359	13, 486, 665	665	573,184
2	12, 51, 884, 746	12, 51, 746	1170	1170,875
3	485, 5, 399, 144, 181	485, 5, 399, 144, 181		
4	166, 488, 1189	166, 488	1061,236	1189, 1255, 1061
5	103, 943	943, 103	1032, 1272, 625	1032, 1272, 625
6	491, 385, 257, 121, 386	491, 121, 257, 385		148
7	492, 434, 56	492, 124, 122, 56	124,973	232
8	122, 556, 232, 492	232,556,122	19	19,492
9	21, 45, 306, 550, 102	45, 306, 21, 550, 102		
10	493, 302, 949,1143,332	302, 493, 332		1010, 1264
11	495, 654, 472, 72	495, 472, 72, 654	305	570
12	624, 650, 941	624, 650, 506, 966	506,966	1232
13	496, 903, 38	496, 903, 520	520, 313	38, 313
14	64, 291, 256, 170, 568	64, 291, 256, 568, 170		
15	462,463	462	1099, 761, 1043	1099, 1043, 817,981
16	498, 106, 1006	498, 1006, 106	93, 869	93, 869
17	700, 445, 1281, 916, 1108	372, 916, 700, 1108, 445		
18	248, 498, 197, 492	498, 248, 1006, 197	1006	56

19	706	706, 1279	82, 1346, 716, 554	716, 82, 1346
20	500, 268, 88, 270, 87	500, 268, 88, 270		1008

**Q4:**

The top ranked documents got a higher score because they might contain some high-weighted terms of low importance from the query, which may be responsible for the overall document getting a higher rank. Also, 'keywords' are considered for calculating relevance and not the context, which may further contribute to few documents getting a higher rank.

**Q5:**

1. W1 scheme uses max-tf weighting and W2 scheme considers document length in their approaches.
2. W1 is based on term frequency and hence weight and score depend on the term frequency. Therefore, if a query term occurs frequently in a non-relevant document, then the document gets a higher score and rank.
3. W2 considers length of document. Therefore, if a document has large number of non-relevant terms, then doclen is also larger, resulting in a higher score and rank for the document.
4. W2 seems to be a worse scheme as compared to W1 as it resulted in more non-relevant terms than W1.

**Q6:**

**Program Description:**

1. Indexing is performed on both Cranfield collection and hw3.queries using the approach from Homework 2.
2. Document and query weight vectors are created of the format term:normalized\_weight using both weighting schemes W1&W2, where  

$$W1 = [0.4 + 0.6 * \log (tf + 0.5) / \log (maxtf + 1.0)] * [\log (collectionsize / df) / \log (collectionsize)]$$

$$W2 = [0.4 + 0.6 * (tf / (tf + 0.5 + 1.5 * (doclen / avgdoclen) ) ) ] * [\log (collectionsize / df) / \log (collectionsize)]$$
3. A query and document dictionary is created each containing their terms and corresponding normalized weights. While retrieving documents, cosine products are computed for weights if terms found if query term found in document. A cosine product dictionary containing cosine product and document vector representation is also stored.
4. Top five documents are displayed according to descending order of cosine product score for each query for both weighting schemes.

