Information Retrieval
TCSS554 A

Homework One

1.1 Tokens before processing : **240540**
1.2 Tokens after processing : **117881**
2. The number of unique words in the database : **7450**
3. The number of words that occur only once in the database : **2881**
4. The average number of word tokens per document : **291.784653**
5. For 30 most frequent words in the database :

Term	Tf	Tf(weight)	df	IDF	tf*idf	p(term)
like	3119	4.494015	363	0.046475	144.954714	0.026459
im	2400	4.380211	372	0.035838	86.012221	0.02036
um	2377	4.376029	294	0.138034	328.1069	0.020164
know	2240	4.350248	358	0.052498	117.596278	0.019002
go	1689	4.22763	331	0.086553	146.188644	0.014328
realli	1426	4.15412	308	0.117831	168.026505	0.012097
dont	1362	4.134177	331	0.086553	117.885692	0.011554
one	1153	4.061829	304	0.123508	142.404472	0.009781
get	1049	4.020775	294	0.138034	144.797702	0.008899
video	1014	4.006038	277	0.163902	166.196218	0.008602
uh	962	3.983175	221	0.261989	252.033506	0.008161
yeah	886	3.947434	255	0.199841	177.05929	0.007516
that	864	3.936514	289	0.145484	125.697763	0.007329
think	834	3.921166	264	0.184777	154.104383	0.007075
want	827	3.917506	270	0.175018	144.739556	0.007016
peopl	796	3.900913	236	0.233469	185.841612	0.006753
thing	770	3.886491	263	0.186426	143.547725	0.006532
make	712	3.85248	277	0.163902	116.697936	0.00604
say	671	3.826723	265	0.183135	122.883915	0.005692
see	655	3.816241	265	0.183135	119.953747	0.005556
well	613	3.78746	245	0.217215	133.152967	0.0052
guy	609	3.784617	215	0.273943	166.831229	0.005166
time	603	3.780317	252	0.204981	123.603437	0.005115
got	603	3.780317	237	0.231633	139.674711	0.005115
right	557	3.745855	229	0.246546	137.326057	0.004725

xxxx	547	3.737987	175	0.363343	198.748794	0.00464
good	542	3.733999	229	0.246546	133.627868	0.004598
ive	538	3.730782	212	0.280046	150.664481	0.004564
lot	507	3.705008	219	0.265937	134.830186	0.004301
gonna	501	3.699838	167	0.383665	192.216112	0.00425