

1. The number of word tokens in the database (before ):  
240535
2. The number of unique words in the database:7377  
The number of word tokens in the database (after): 114826
3. The number of words that occur only once in the  
database: 2850
4. The average number of word tokens per document :  
317.06188118811883
4. For 30 most frequent words in the database, provide:  
TF, scaled TF (1+log(tf)), IDF, TF\*IDF and probabilities

For 30 most frequent words in the database, provide:

Term	Tf	N	Tf(weight)	df	IDF	TF.IdF	probabilities
like	3119	404	9.0453	363	0.1070	333.7706	0.0272
m	2418	404	8.7907	404	0.0000	0.0000	0.0211
um	2377	404	8.7736	338	0.1784	423.9831	0.0207
know	2240	404	8.7142	358	0.1209	270.7754	0.0195
go	1689	404	8.4319	392	0.0302	50.9285	0.0147
realli	1426	404	8.2626	308	0.2713	386.8953	0.0124
one	1153	404	8.0501	332	0.1963	226.3107	0.0100
get	1049	404	7.9556	304	0.2844	298.3221	0.0091
video	1014	404	7.9217	278	0.3738	379.0269	0.0088
re	979	404	7.8865	400	0.0100	9.7414	0.0085
uh	962	404	7.8690	233	0.5504	529.4621	0.0084
yeah	886	404	7.7867	255	0.4602	407.6941	0.0077
think	834	404	7.7262	264	0.4255	354.8385	0.0073
want	827	404	7.7178	270	0.4030	333.2751	0.0072
peopl	796	404	7.6796	236	0.5376	427.9161	0.0069
thing	770	404	7.6464	270	0.4030	310.3045	0.0067
ve	742	404	7.6093	394	0.0251	18.5975	0.0065
make	712	404	7.5681	279	0.3702	263.5846	0.0062
say	671	404	7.5088	267	0.4142	277.9055	0.0058
ll	662	404	7.4953	396	0.0200	13.2404	0.0058
see	655	404	7.4846	286	0.3454	226.2521	0.0057
guy	609	404	7.4118	215	0.6308	384.1431	0.0053
time	603	404	7.4019	252	0.4720	284.6074	0.0053
got	603	404	7.4019	258	0.4485	270.4185	0.0053
right	558	404	7.3244	242	0.5125	285.9623	0.0049
xxxx	548	404	7.3063	175	0.8366	458.4726	0.0048
good	542	404	7.2953	237	0.5334	289.0783	0.0047
would	512	404	7.2383	215	0.6308	322.9577	0.0045
well	509	404	7.2324	221	0.6033	307.0554	0.0044
lot	507	404	7.2285	228	0.5721	290.0391	0.0044

Done