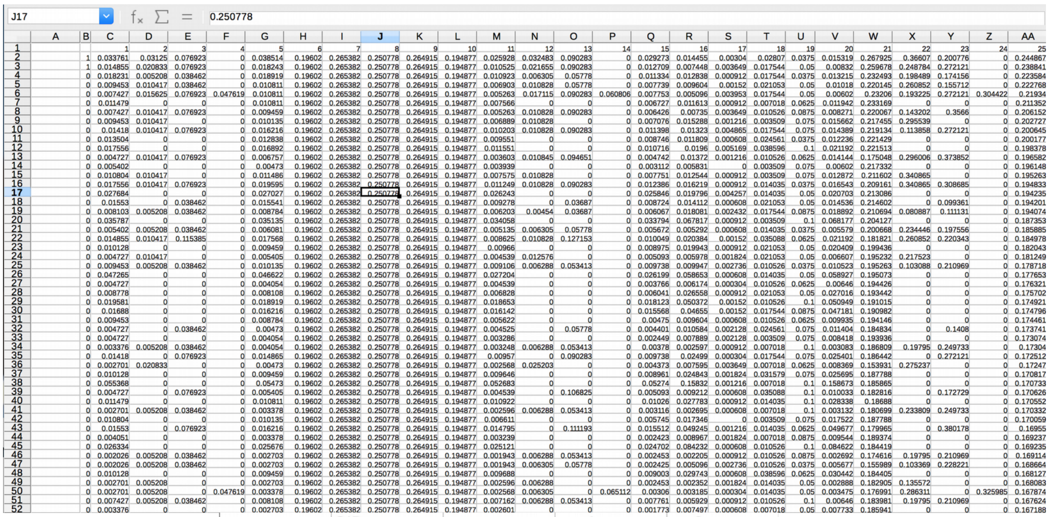
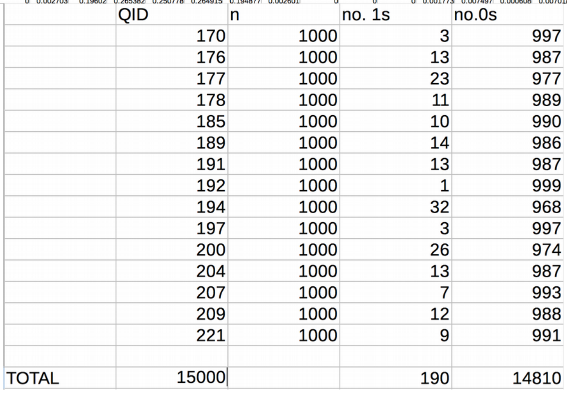
Analysis

motivation(1): make CSV files that can be read easily into python or excel spread sheet
method: write a string manipulation program called reformat_data.py, get rid of the redundant characters, such as ‘qid:WT04:’ , “1:” etc...



J17 0.250778

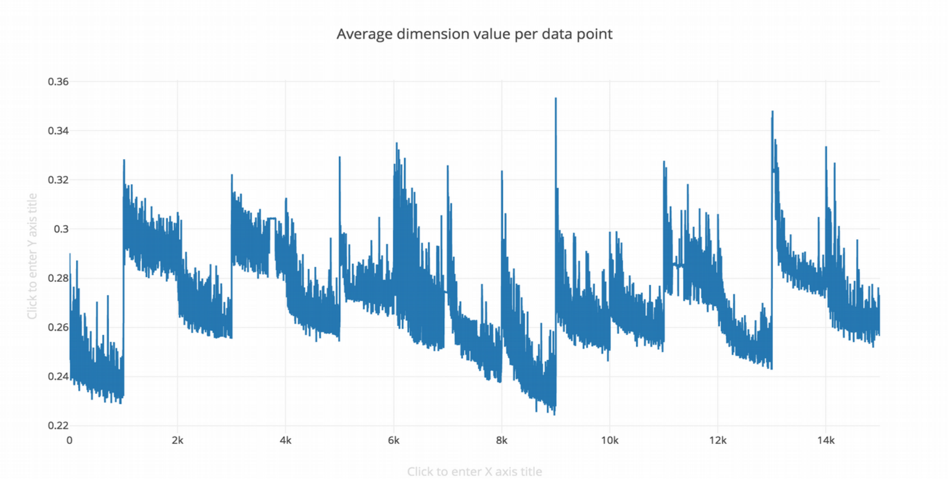
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
2		1	0.033761	0.03125	0.076923	0	0.038514	0.19602	0.265382	0.250778	0.264915	0.194877	0.025928	0.032483	0.090283	0	0.029273	0.014455	0.00304	0.02807	0.0375	0.015319	0.267925	0.36607	0.200776	0	0.244867
3		1	0.014855	0.020833	0.076923	0	0.018243	0.19602	0.265382	0.250778	0.264915	0.194877	0.010525	0.021655	0.090283	0	0.012709	0.007448	0.003649	0.017544	0.05	0.00832	0.259678	0.248784	0.272121	0	0.238841
4		0	0.018231	0.005208	0.038462	0	0.018919	0.19602	0.265382	0.250778	0.264915	0.194877	0.010923	0.006305	0.05778	0	0.011334	0.012838	0.000912	0.017544	0.0375	0.013215	0.232493	0.198489	0.174156	0	0.223584
5		0	0.009453	0.010417	0.038462	0	0.010811	0.19602	0.265382	0.250778	0.264915	0.194877	0.006903	0.010828	0.05778	0	0.007739	0.009604	0.00152	0.021053	0.05	0.01018	0.220145	0.260852	0.155712	0	0.222768
6		0	0.007427	0.015625	0.076923	0.047619	0.010811	0.19602	0.265382	0.250778	0.264915	0.194877	0.005263	0.017115	0.090283	0.060806	0.007753	0.005096	0.003953	0.017544	0.05	0.00602	0.23206	0.193225	0.272121	0.304422	0.21934
7		0	0.011479	0	0	0	0.010811	0.19602	0.265382	0.250778	0.264915	0.194877	0.007566	0	0	0	0.006727	0.011613	0.000912	0.007018	0.0625	0.011942	0.233169	0	0	0	0.211352
8		0	0.007427	0.010417	0.076923	0	0.009459	0.19602	0.265382	0.250778	0.264915	0.194877	0.005263	0.010828	0.090283	0	0.006426	0.00735	0.003649	0.010526	0.0875	0.008271	0.220067	0.143202	0.3566	0	0.206152
9		0	0.009453	0.010417	0	0	0.010135	0.19602	0.265382	0.250778	0.264915	0.194877	0.006889	0.010828	0	0	0.007076	0.015288	0.001216	0.003509	0.075	0.015662	0.217455	0.295539	0	0	0.202727
10		0	0.01418	0.010417	0.076923	0	0.016216	0.19602	0.265382	0.250778	0.264915	0.194877	0.010203	0.010828	0.090283	0	0.011398	0.01323	0.004865	0.017544	0.075	0.014389	0.219134	0.113858	0.272121	0	0.200645
11		0	0.013504	0	0	0	0.012838	0.19602	0.265382	0.250778	0.264915	0.194877	0.009551	0	0	0	0.008746	0.011809	0.000608	0.024561	0.0375	0.012236	0.221429	0	0	0	0.200177
12		0	0.017556	0	0	0	0.016892	0.19602	0.265382	0.250778	0.264915	0.194877	0.011551	0	0	0	0.010716	0.0196	0.005169	0.038596	0.1	0.021192	0.221513	0	0	0	0.198378
13		0	0.004727	0.010417	0.076923	0	0.006757	0.19602	0.265382	0.250778	0.264915	0.194877	0.003603	0.010845	0.094651	0	0.004742	0.01372	0.001216	0.010526	0.0625	0.014144	0.175048	0.296006	0.373852	0	0.196582
14		0	0.005402	0	0	0	0.00473	0.19602	0.265382	0.250778	0.264915	0.194877	0.003939	0	0	0	0.003112	0.005831	0	0.003509	0.075	0.00602	0.217332	0	0	0	0.196148
15		0	0.010804	0.010417	0	0	0.011486	0.19602	0.265382	0.250778	0.264915	0.194877	0.007575	0.010828	0	0	0.007751	0.012544	0.000912	0.003509	0.075	0.012872	0.211602	0.340865	0	0	0.195263
16		0	0.017556	0.010417	0.076923	0	0.019595	0.19602	0.265382	0.250778	0.264915	0.194877	0.011249	0.010828	0.090283	0	0.012386	0.016219	0.000912	0.014035	0.0375	0.016543	0.209161	0.340865	0.308685	0	0.194833
17		0	0.027684	0	0	0	0.027027	0.19602	0.265382	0.250778	0.264915	0.194877	0.026243	0	0	0	0.025846	0.019796	0.004257	0.014035	0.05	0.020703	0.213086	0	0	0	0.194235
18		0	0.01553	0	0.038462	0	0.015541	0.19602	0.265382	0.250778	0.264915	0.194877	0.009278	0	0.03687	0	0.008724	0.014112	0.000608	0.021053	0.05	0.014536	0.214602	0	0.099361	0	0.194201
19		0	0.008103	0.005208	0.038462	0	0.008784	0.19602	0.265382	0.250778	0.264915	0.194877	0.006203	0.00454	0.03687	0	0.006067	0.018081	0.002432	0.017544	0.0875	0.018892	0.210694	0.080887	0.111131	0	0.194074
20		0	0.035787	0	0	0	0.035135	0.19602	0.265382	0.250778	0.264915	0.194877	0.034058	0	0	0	0.033794	0.067817	0.000912	0.003509	0.1	0.068177	0.204127	0	0	0	0.187353
21		0	0.005402	0.005208	0.038462	0	0.006081	0.19602	0.265382	0.250778	0.264915	0.194877	0.005135	0.006305	0.05778	0	0.005672	0.005292	0.000608	0.014035	0.0375	0.005579	0.200668	0.234446	0.197556	0	0.185885
22		0	0.014855	0.010417	0.115385	0	0.017568	0.19602	0.265382	0.250778	0.264915	0.194877	0.008625	0.010828	0.127153	0	0.010049	0.020384	0.00152	0.035088	0.0625	0.021192	0.181821	0.260852	0.220343	0	0.184978
23		0	0.010128	0	0	0	0.009459	0.19602	0.265382	0.250778	0.264915	0.194877	0.00966	0	0	0	0.008975	0.019943	0.000912	0.021053	0.05	0.020409	0.199436	0	0	0	0.182043
24		0	0.004727	0.010417	0	0	0.005405	0.19602	0.265382	0.250778	0.264915	0.194877	0.004539	0.012576	0	0	0.005093	0.005978	0.001824	0.021053	0.05	0.006607	0.195232	0.217523	0	0	0.181249
25		0	0.009453	0.005208	0.038462	0	0.010135	0.19602	0.265382	0.250778	0.264915	0.194877	0.009106	0.006288	0.053413	0	0.009738	0.009947	0.002736	0.010526	0.0375	0.010523	0.195263	0.103088	0.210969	0	0.178718
26		0	0.047265	0	0	0	0.046622	0.19602	0.265382	0.250778	0.264915	0.194877	0.027204	0	0	0	0.026199	0.058653	0.000608	0.014035	0.05	0.058927	0.195073	0	0	0	0.177653
27		0	0.004727	0	0	0	0.004054	0.19602	0.265382	0.250778	0.264915	0.194877	0.004539	0	0	0	0.003766	0.006174	0.000304	0.010526	0.0625	0.00646	0.194426	0	0	0	0.176321
28		0	0.008778	0	0	0	0.008108	0.19602	0.265382	0.250778	0.264915	0.194877	0.006828	0	0	0	0.006041	0.026558	0.000912	0.021053	0.05	0.027016	0.193442	0	0	0	0.175702
29		0	0.019581	0	0	0	0.018919	0.19602	0.265382	0.250778	0.264915	0.194877	0.018653	0	0	0	0.018123	0.050372	0.00152	0.010526	0.1	0.050949	0.191015	0	0	0	0.174921
30		0	0.01688	0	0	0	0.016216	0.19602	0.265382	0.250778	0.264915	0.194877	0.016142	0	0	0	0.015568	0.04655	0.00152	0.017544	0.0875	0.047181	0.190982	0	0	0	0.174796
31		0	0.009453	0	0	0	0.008784	0.19602	0.265382	0.250778	0.264915	0.194877	0.005622	0	0	0	0.00475	0.009604	0.000608	0.010526	0.0625	0.009935	0.194146	0	0	0	0.174461
32		0	0.004727	0	0.038462	0	0.00473	0.19602	0.265382	0.250778	0.264915	0.194877	0.004525	0	0.05778	0	0.004401	0.010584	0.002128	0.024561	0.075	0.011404	0.184834	0	0.1408	0	0.173741
33		0	0.004727	0	0	0	0.004054	0.19602	0.265382	0.250778	0.264915	0.194877	0.003286	0	0	0	0.002449	0.007889	0.002128	0.003509	0.075	0.008418	0.193936	0	0	0	0.173074
34		0	0.003376	0.005208	0.038462	0	0.004054	0.19602	0.265382	0.250778	0.264915	0.194877	0.003248	0.006288	0.053413	0	0.00378	0.002597	0.000912	0.007018	0.1	0.003083	0.186809	0.19795	0.249733	0	0.17304
35		0	0.01418	0	0.076923	0	0.014865	0.19602	0.265382	0.250778	0.264915	0.194877	0.00957	0	0.090283	0	0.009738	0.02499	0.000304	0.017544	0.075	0.025401	0.186442	0	0.272121	0	0.172512
36		0	0.002701	0.020833	0	0	0.00473	0.19602	0.265382	0.250778	0.264915	0.194877	0.002568	0.025203	0	0	0.004373	0.007595	0.003649	0.007018	0.0625	0.008369	0.153931	0.275237	0	0	0.17247
37		0	0.010128	0	0	0	0.009459	0.19602	0.265382	0.250778	0.264915	0.194877	0.009646	0	0	0	0.008961	0.024843	0.001824	0.031579	0.075	0.025695	0.187788	0	0	0	0.170817
38		0	0.055368	0	0	0	0.05473	0.19602	0.265382	0.250778	0.264915	0.194877	0.052683	0	0	0	0.05274	0.15832	0.001216	0.007018	0.1	0.158673	0.185865	0	0	0	0.170733
39		0	0.004727	0	0.076923	0	0.005405	0.19602	0.265382	0.250778	0.264915	0.194877	0.004539	0	0.106825	0	0.005093	0.009212	0.000608	0.035088	0.1	0.010033	0.182816	0	0.172729	0	0.170626
40		0	0.011479	0	0	0	0.010811	0.19602	0.265382	0.250778	0.264915	0.194877	0.010922	0	0	0	0.01026	0.027783	0.000912	0.014035	0.1	0.028338	0.18688	0	0	0	0.170552
41		0	0.002701	0.005208	0.038462	0	0.003378	0.19602	0.265382	0.250778	0.264915	0.194877	0.002596	0.006288	0.053413	0	0.003116	0.002695	0.000608	0.007018	0.1	0.003132	0.180699	0.233809	0.249733	0	0.170332
42		0	0.010804	0	0	0	0.010135	0.19602	0.265382	0.250778	0.264915	0.194877	0.006611	0	0	0	0.005745	0.017346	0	0.003509	0.075	0.017522	0.187788	0	0	0	0.170059
43		0	0.01553	0	0.076923	0	0.016216	0.19602	0.265382	0.250778	0.264915	0.194877	0.014795	0	0.111193	0	0.015512	0.049245	0.001216	0.014035	0.0625	0.049677	0.179965	0	0.380178	0	0.16955
44		0	0.004051	0	0	0	0.003378	0.19602	0.265382	0.250778	0.264915	0.194877	0.003239	0	0	0	0.002423	0.008967	0.001824	0.007018	0.0875	0.009544	0.189374	0	0	0	0.169237
45		0	0.026334	0	0	0	0.025676	0.19602	0.265382	0.250778	0.264915	0.194877	0.025121	0	0	0	0.024702	0.084232	0.000608	0.010526	0.1	0.084622	0.184419	0	0	0	0.169235
46		0	0.002026	0.005208	0.038462	0	0.002703	0.19602	0.265382	0.250778	0.264915	0.194877	0.001943	0.006288	0.053413	0	0.002453	0.002205	0.000912	0.010526	0.0875	0.002692	0.174616	0.19795	0.210969	0	0.169114
47		0	0.002026	0.005208	0.038462	0	0.002703	0.19602	0.265382	0.250778	0.264915	0.194877	0.001943	0.006305	0.05778	0	0.002425	0.005096	0.002736	0.010526	0.0375	0.005677	0.155989	0.103369	0.228221	0	0.168664
48		0	0.010128	0	0	0	0.009459	0.19602	0.265382	0.250778	0.264915	0.194877	0.009688	0	0	0	0.009003	0.029743	0.000608	0.038596	0.0625	0.030442	0.184405	0	0	0	0.168127
49		0	0.002701	0.005208	0	0	0.002703	0.19602	0.265382	0.250778	0.264915	0.194877	0.002596	0.006288	0	0	0.002453	0.002352	0.001824	0.014035	0.05	0.002888	0.182905	0.135572	0	0	0.168083
50		0	0.002701	0.005208	0	0.047619	0.003378	0.19602	0.265382	0.250778	0.264915	0.194877	0.002568	0.006305	0	0.065112	0.00306	0.003185	0.000304	0.014035	0.05	0.003475	0.176991	0.286311	0	0.325985	0.167874
51		0	0.007427	0.005208	0.038462	0	0.008108	0.19602	0.265382	0.250778	0.264915	0.194877	0.007162	0.006288	0.053413	0	0.007761	0.005929	0.000912	0.010526	0.1	0.00646	0.183981	0.19795	0.210969	0	0.167624
52		0	0.003376	0	0	0	0.002703	0.19602	0.265382	0.250778	0.264915	0.194877	0.002601	0	0	0	0.001773	0.007497	0.000608	0.007018	0.05	0.007733	0.185941	0	0	0	0.167188



	QID	n	no. 1s	no.0s
	170	1000	3	997
	176	1000	13	987
	177	1000	23	977
	178	1000	11	989
	185	1000	10	990
	189	1000	14	986
	191	1000	13	987
	192	1000	1	999
	194	1000	32	968
	197	1000	3	997
	200	1000	26	974
	204	1000	13	987
	207	1000	7	993
	209	1000	12	988
	221	1000	9	991
TOTAL	15000		190	14810

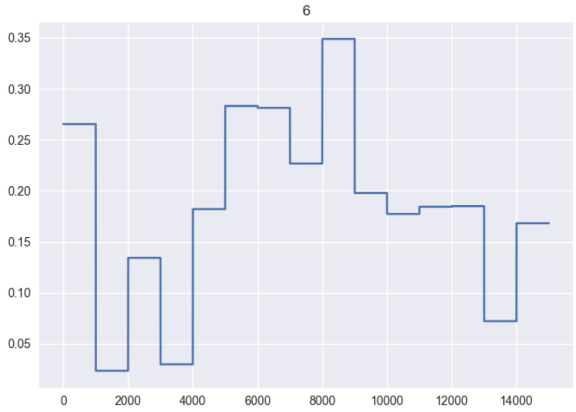
Observation: Preliminary observations show that the Query’s are organized into 1000 documents each. All together 15000 documents output, on 15 different query’s. Relatively few relevant documents across the query’s. Qid:192. Only has one relevant documets.

Motivation: plot to see obvious overall structure.
Method: plots the horizontal mean vs document indexes

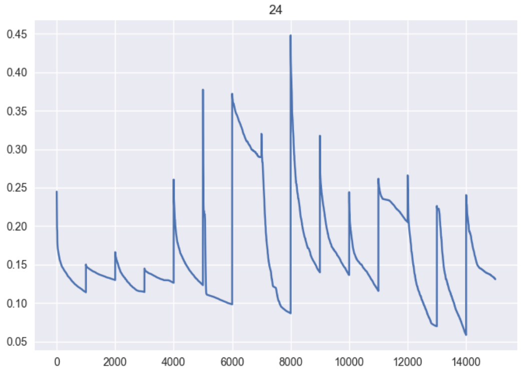


Observation: Quite obvious mean structural change between querys, similar level of signal strength across the querys, with the 7th query (qid:191) show noticeably high strength. A curious decay in mean across the documents indexing within each query. This suggest that the documents indexing has a informative ordering.

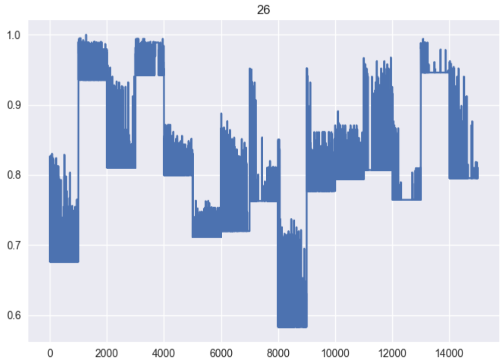
Motivation: find out the source of this decay, and mean structure change, and documents signal
Method: plot the individual features vs documents
Observations: By looking at the 64 plots, it looks like we can seperate into 6 class



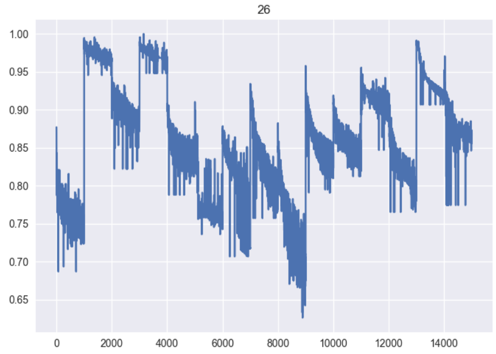
Class1: [6,7,8,9,10] these features show zero document signal, changes only when query changes. These are certainly query based only features.



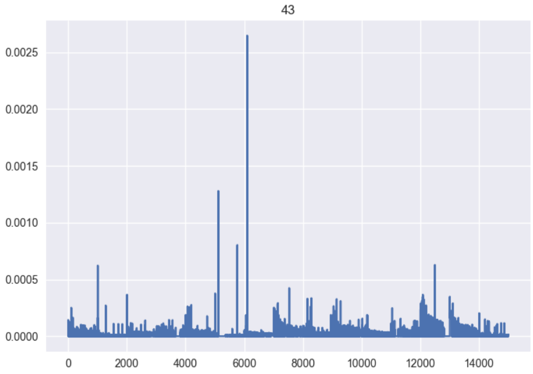
Class2: [25] This singular feature, capture no documents signal, respond to query change only and has a decay across the document indexing. We can use this feature to learn the decay rate of signal strength for each Querys or the average decay rate of all Querys. This learnt decay rate can be used to offset the signal decay of other features.



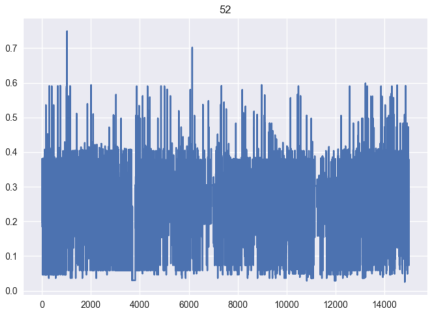
Class3: [27,28,29,32,33,34,37,38,39,45,61,62,63,64] Captures everything and no decay. These are likely to be the most informative assessment features for the interactions between query and documents in-order to learn relevance.



Class4: [21,26,30,31,35,36,40,41,42,44,46,47,48] captures everything and included the decay. We can offset this decay by the decay rate learnt from class 2.



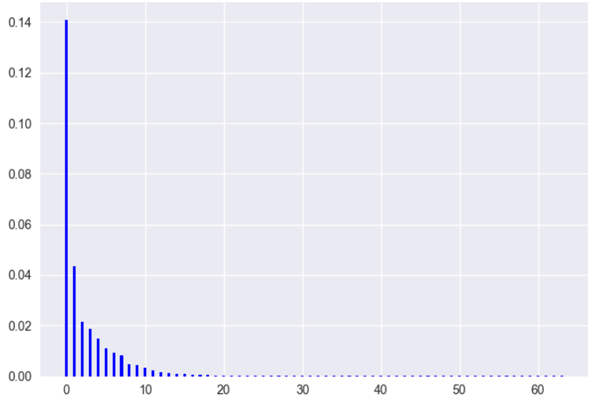
Class5: [1,2,5,11,12,13,14,15,16,20,43,49,50,51,53,54,55,56,60] These features captures the documents, but only very minor level of decay and query (if at all)



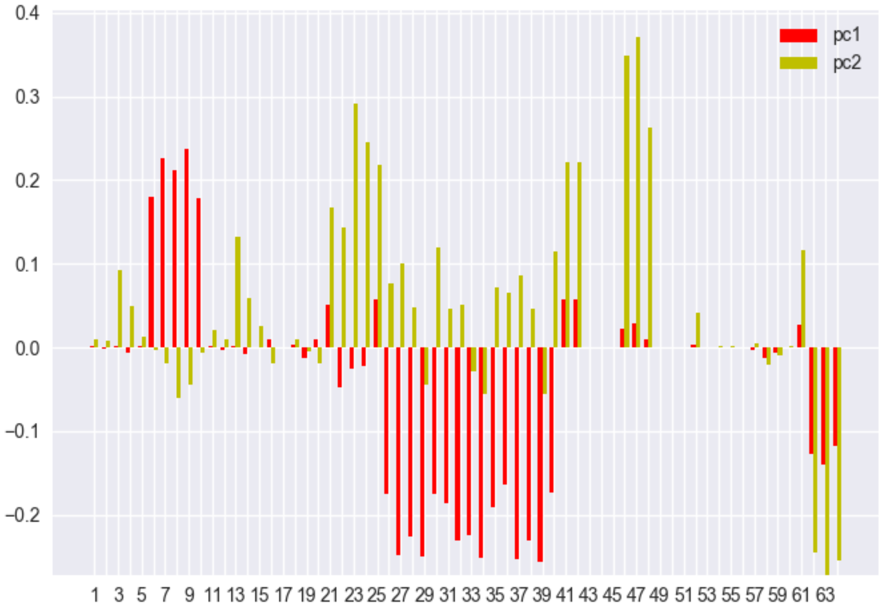
Class6: [3,4,17,18,19,22,23,24,52,57,58,59]. These features doesn’t seem to capture any query information or decay at all. It is very likely that these are documents based only features. However, the precise set member of this class are harder to tell than class 1, since every documents are unique, we can’t hold the documents constant and check if they really respond to query for sure.

Observation: It is very possible that class member of Class 5 and Class 6 get mixed up. In addition, we can deduce that, of the features that is based on both Documents and Querys, with the exception of class 3; the more they react to query, the more they are affected by decay.

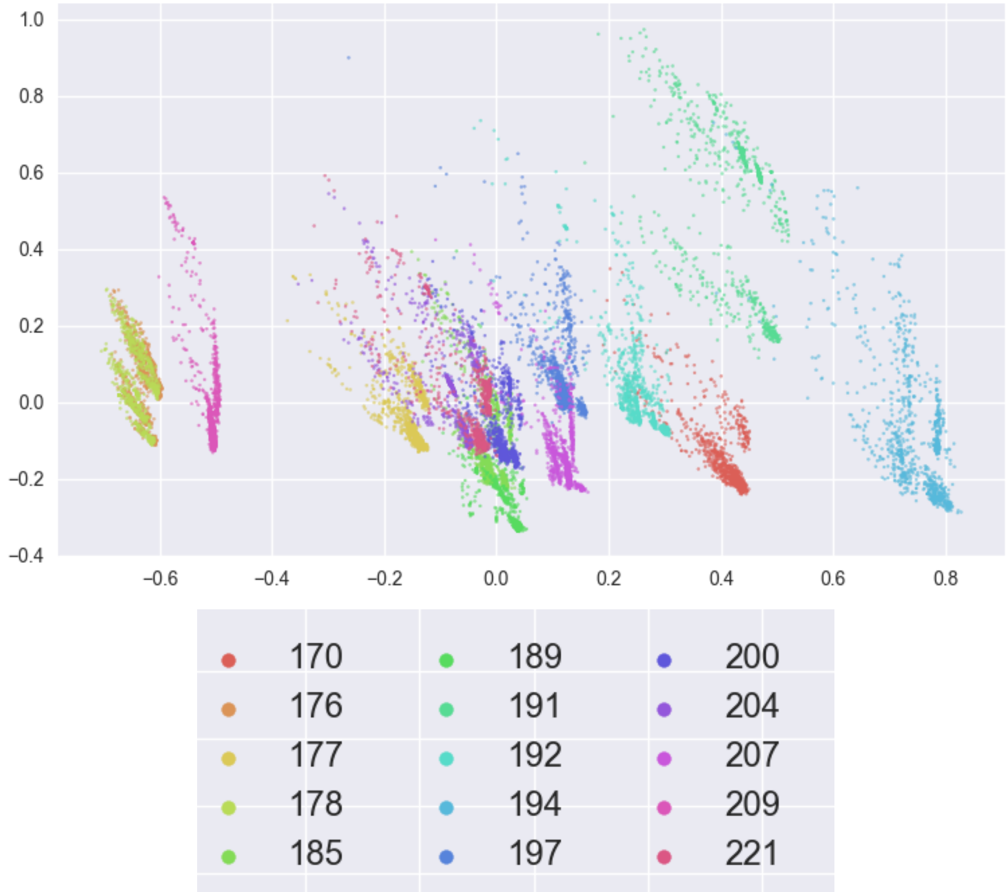
Motivation: Use Dimensionality reduction to visualize data
Method: PCA



These are the eigenvalues of the 64 PCs: The first 10 principle component almost explains almost all the variance. The First and 2nd principle components combined is sufficient to explain 63% of all the variance.

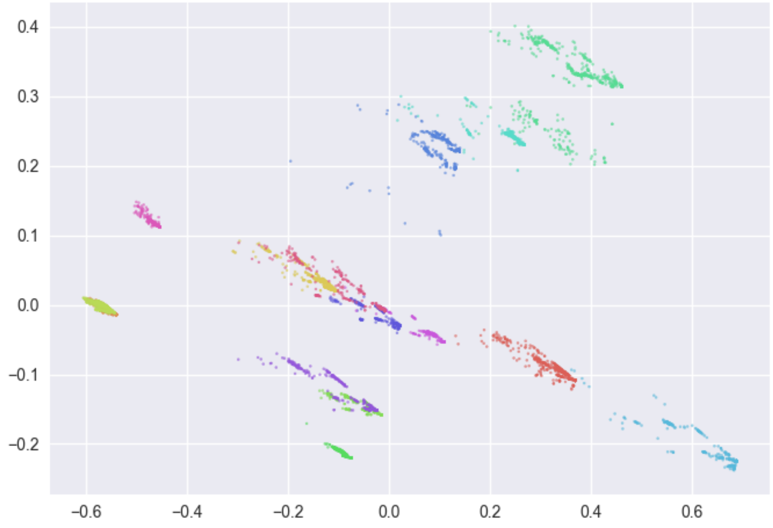


The direction of projection of the 1st and 2nd principle component. As expected from previous plots, class 1,2,3,4 has dominant presence. With Class 1,2,3 mostly in the PC1, and class 4 mostly in PC4.

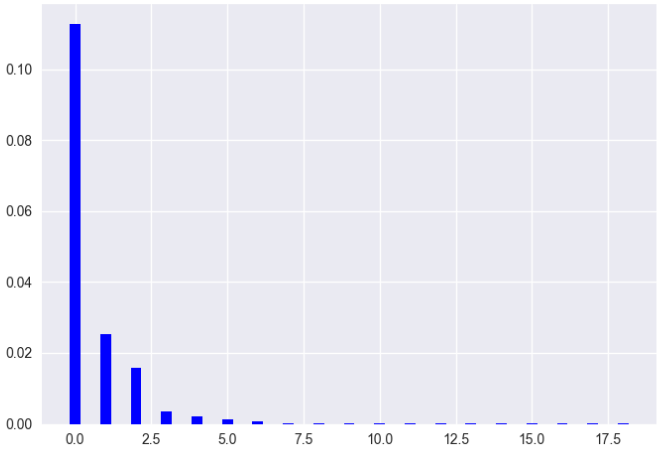


Scatterplot of all data points projected into a plain span by PC1 (x-axis) and PC2 (y-axis). Notice that qid176 and qid178 is almost identical in shape and position. And the Furtherest away.

Motivation: Lets focus on the features most closely related to Qids, ie: class1 (query based only without decay), class3 (strong query based relative to DocBased without decay). To to better differentiate between the Qids.

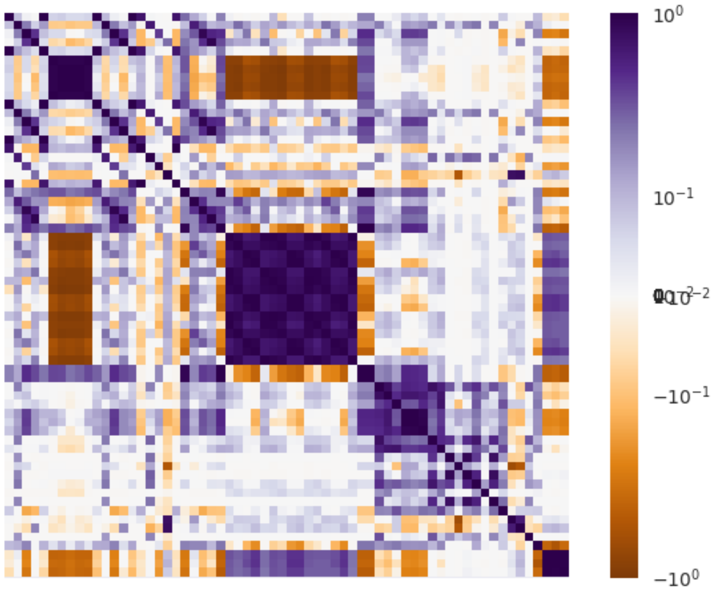


Incase, if I want to do query specific supervised learning and there isn't sufficient relevancy count, I can group them based on the above clustering.

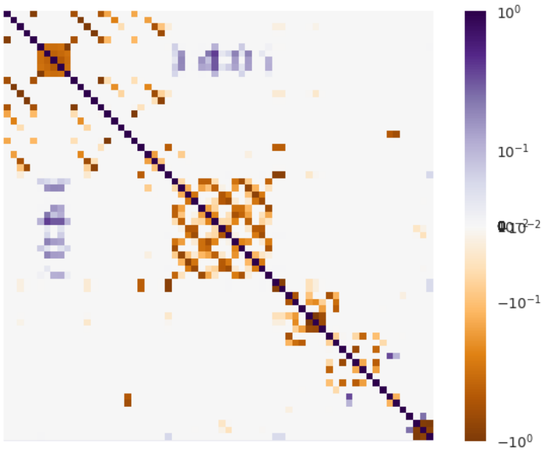


The eigenvalue decay from 1st PC on the left to 19th PC. With features from class1 and class 3 only. Observation: Only 3 orthogonal component is enough to describe the qids type. With 1st PCs contributing to 96% of all the variances.

Motivation: discover more relationship betweens features
Methods: Normalized the entire feature spaces, and make covariance matrix and then use GraphicalLasso to discover sparsed percision matrix.

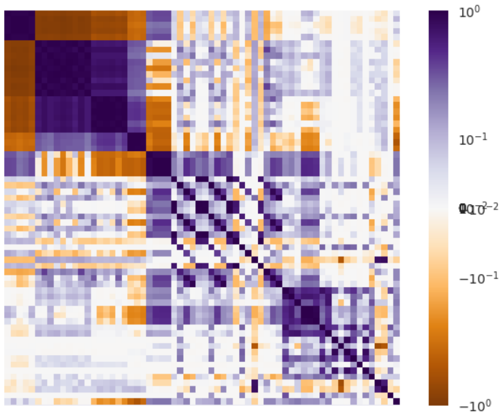


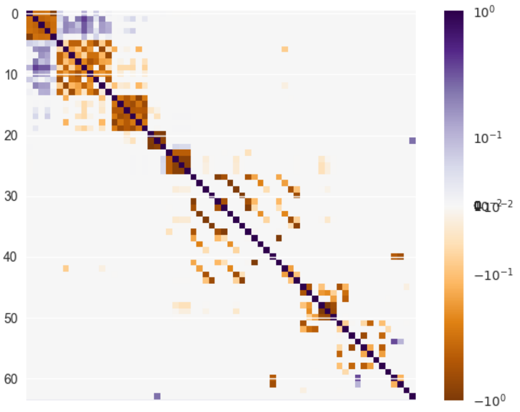
Covariance matrix: Observation: Since we know that 6,7,8,9,10. Belong to class 1 and has the cleanest query based features we can deduce from that, block [25 to 39] has the closest interaction with the query (which made sense because they are majority class 3) Follow with block [61 to 64]. Notice that there is a checker board pattern in black 25 to 39. This mean that they are ordered in an interesting phase oscillation pattern, and they can be further divided into 2 subgroub. Subgroub 1: [25,26,29,30,31,34,35,36,39], subgroup 2: [27,28,32,33,37,38]. Subgroub 1 has slightly stronger tide with the query than subgroup 2. They each have stronger relationship with features within the subgroup, less but still strong with the other subgroup.

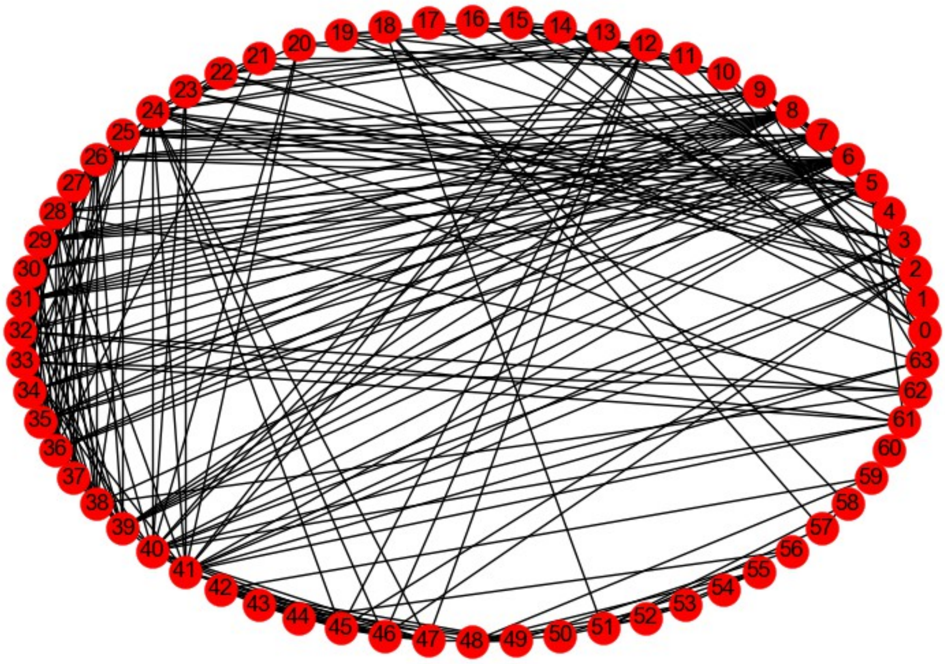


Sparsed precision matrix from Graphical Lasso: Observation: Notice that the there a repetitive pattern between in a period of 5 indexes between [10,11,12,13,14] → [15,16,17,18,19]-→ [20,21,22,23,24]. This could be another oscillatory relationship therefore related subgroups. Also notice that block 61-64 is now conditionally independent to block 6 to10 and the big middle block. Therefore 61-64 could be capturing the redundant part of the query. Also, block[49-59] do is conditionally independent on almost everything else. This could be a strong indicator that they are capturing the document based only information.

Motivation: Move the coulombs around to better visualize relationship
Methods: Manually move block [6-10] to left most. Follow with subgroup from the middle block, [25,26,29,30,31,34,35,36,39], then [27,28,32,33,37,38], and then [61-64].



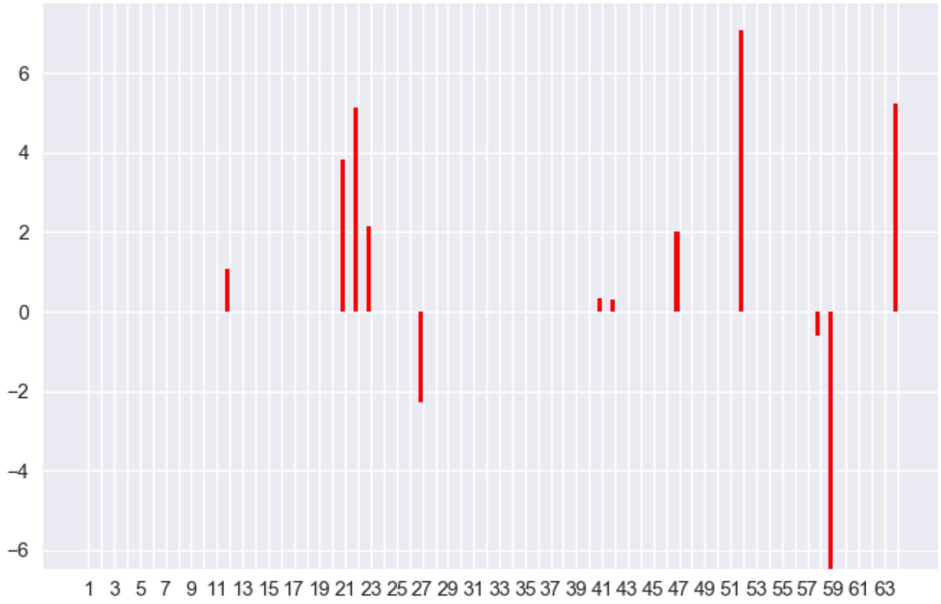




Above is a network built by the Precision Matrix as a weighted adjacency matrix. If given more time. I could use Hierarchical clustering with linkage to better group the features.

Motivation: Try to predict the relevancy by the features. See which features contributes the most to a documents relevancy.

Method: Logistic regression using 'L_1' penalty to force sparsity in the prediction co-efficient. Use 5-fold cross validation to tune the regularization parameter so as to not overfit or under-fit.



Above shows the co-efficient of the features in the decision function. Learnt by the logistic regression solver. It has the average score of 0.98 according to my 5-fold cross validation algorithm. However, this may not be a high score at all since the there are far too many zero's and not enough ones in the relevancy column. Interestingly, the majority of the dominant predictor are what I considered class 5 or class 6; the classes of features which I believe are almost all documents based only. This could mean that, the Queries given in this data set are not sufficiently different to make a difference to the probability of relevancy.

If I have more time, I will try to use a different function approximator to predict relevancy. I would approximate the decay using feature 25, use it to mediate the decay found in class 2,4,5 before using any supervised learning algorithm. I would try centering and normalizing the data set before applying the learning algorithm. I would use Hierarchical clustering to better group the features.