Report

Preprocess:

To preprocess the data, I used bs4.BeautifulSoup package. After this process, around 8090 articles are generated.

Clustering algorithm:

In this code, I implemented incremental k-means algorithm. The relevant two centroids are updated every time the point assignment changes. And the termination condition is when the average centroid moves is below a user given threshold.

Crition_fun	model	clusterNum	objective_value	entropy	purity	runtime(s)/trial
SSE	freq	20	6229.477372	1.6663437	0.665184	162.7278782
SSE	freq	40	6047.203583	1.525876831	0.684705	129.98882
SSE	freq	60	5917.87562	1.29427298	0.726711	139.8542298
SSE	sqrtfreq	20	6540.850469	1.938630724	0.641216	153.5665787
SSE	sqrtfreq	40	6449.127377	1.894609084	0.644675	113.835056
SSE	sqrtfreq	60	6351.500977	1.698007978	0.690882	88.93364667
SSE	log2freq	20	6333.350006	1.668440501	0.662219	148.9144364
SSE	log2freq	40	6215.832506	1.549629871	0.691747	89.03486
SSE	log2freq	60	6086.442036	1.447193135	0.691994	100.1180476
I2	freq	20	5118.412172	1.300545093	0.722511	151.0304796
I2	freq	40	5074.952308	0.984773257	0.787126	147.7004182
I2	freq	60	5220.805144	0.861857715	0.825303	181.562588
I2	sqrtfreq	20	5005.208656	1.178696569	0.75315	119.4061262
I2	sqrtfreq	40	4744.192827	0.874139398	0.830121	132.6501151
I2	sqrtfreq	60	4875.076169	0.735282861	0.853842	141.5522058
I2	log2freq	20	5172.661939	1.164513529	0.749691	140.0015658
I2	log2freq	40	4947.54562	0.861184205	0.831109	154.5413413
I2	log2freq	60	5093.3051	0.7624464	0.850507	204.0816182
E1	freq	20	6208.935763	1.962729836	0.570299	733.8658996
E1	freq	40	5652.95416	1.561294284	0.662837	1258.441885
E1	log2freq	20	6201.428852	1.855002429	0.59192	739.8172964
E1	log2freq	40	6041.920611	1.723408257	0.620583	1404.856147

Didn't have enough time to take other data for E1.

Clustering analyses:

In the table it shows all the result of performance for all combinations of criterion function, vector model and cluster number. Based on this result, we can conclude that

1. **Entropy.** The total entropy within a kind of criterion function are close to each other. And generally I2 has the lowest entropy, around 0.9, while E1 has the highest entropy, around 1.8. I2 is the best. Also note that the max entropy is -log(20) = 4.32, meaning uniform distribution. Compared to that, even E1 with 1.8 entropy is not bad.

2. **Purity.** Again we can see I2 criterion has the best performance in terms of purity. It has purity of around 0.8, very close to the maximum ie 1.0. E1 has the lowest purity, around 0.6, which again is not bad because the lowest purity is 1/20 = 0.05. And purity are close to each other within a kind of criterion.
3. **Run time.** As the run time increases, the cluster number increases. Furthermore, I2 and SSE have similar runtime, and when cluster# increases, they don't slow down very much. However, E1's run time greatly increases as the cluster# increases. Besides it's 5 times slower than the rest two. This is because the similarity calculation of E1 is much more complex than the others; it's based on the objective function change. That's why it's much slower.
4. **Vector model.** Based on the entropy and purity, we can see the log2freq has better result than sqrtfreq, and sqrtfreq has better performance than freq.