Preprocessing:

Runtime: 227 seonconds

Additional processing step:

There were some weird issues with BeautifulSoup not able to find <body> tag not matter what function I use, such as .body or find('body'). So the way I determine if an article contains body tag is by first replacing all word "BODY" with word "CONTENT" and then use BeautifulSoup to detect "content" tag. This increase the runtime a bit, but it's the only way I could solve this issue.

Clustering

The more cluster, the higher the criterion value in general.

L2 have lower runtime than SSE in general (used cosine similarity in l2)

model	criterion function	number cluster	criterion value	entropy	purity	runtime
freq	l2	20	843.1211754	-2.578977346	3735	421382
freq	l2	40	843.1211754	-2.578977346	3735	611574
freq	l2	60	843.1211754	-2.578977346	3735	636500
freq	SSE	20	337.3943456	-5.157954693	7470	895910
freq	SSE	40	316.0392376	-2.578977346	3735	1937754
freq	SSE	60	307.8216189	-2.578977346	3735	1687987
sqrtfreq	SSE	20	258.803252	-5.157954693	7470	647441
sqrtfreq	SSE	40	256.0724034	-5.157954693	7470	632037
sqrtfreq	SSE	60	251.8892141	-5.157954693	7470	1002119
log2freq	SSE	20	293.8911044	-2.578977346	3735	868918
log2freq	SSE	40	280.0900066	-5.157954693	7470	941225
log2freq	l2	40	802.6069605	-2.578977346	3735	529093
log2freq	l2	60	802.6069605	-2.578977346	3735	668986

*Data for E1 is not collected here, but E1 was implemented

Note:

Format of output

- `output file:` file stored the same as the name in command parameter

- cluster distribution file: a two dimensional matrix of dimensions in the format of

	class 1	class 2	…	class 20
cluster 1	value	value		value
cluster 2	value	value		value
...				

- evaluation file: print the value of the criterion function, entropy, purity for the best trial