Information Retrieval Assignment 2

Submitted by: Shubhank Bhandarkar
Roll No: 17BM6JP49
Date: 26th March'18

Analysis:

Code submitted consists of modules for all important functions required for generating summaries. It is mainly divided into following parts:

1. Sentence extraction module:
    - Extracting sentence
    - Reading all the files together
    - Text pre-processing
    - Appending all the sentences together
2. Similarity Matrix module:
    - Creating tf-idf matrix
    - Creating idf-modified-cosine matrix
3. Degree Centrality module
    - Calculating centrality given a particular threshold
    - Generating summaries
4. Text Rank module
    - Power Method
    - Generating text rank matrix
5. Summary generating module

Some findings:

1. Topic 1 ground truth is top 250 words of file named '*APW20000901.0039*' from Topic 1 files and when I generated the 250 words from only this one file I was getting results as follows:
    - Rouge-1: {'f': 0.88, 'p': 0.93, 'r': 0.85}
    - Rouge-2: {'f': 0.87, 'p': 0.90, 'r': 0.84}
    - Rouge-L: {'f': 0.92, 'p': 0.95, 'r': 0.89}

Results generated from the code are below:

Topic	Threshold	Metric	Degree Centrality			TextRank		
			Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L
Topic 1	0.1	F-1	0.227	0.035	0.136	0.231	0.031	0.124
		Precision	0.234	0.036	0.142	0.247	0.032	0.129
		Recall	0.221	0.034	0.132	0.217	0.029	0.120
	0.2	F-1	0.248	0.035	0.116	0.233	0.031	0.128
		Precision	0.266	0.036	0.120	0.247	0.032	0.133
		Recall	0.233	0.033	0.112	0.221	0.029	0.124
	0.3	F-1	0.243	0.030	0.140	0.230	0.039	0.140
		Precision	0.260	0.032	0.146	0.240	0.041	0.146
		Recall	0.229	0.029	0.136	0.220	0.038	0.136
Topic 2	0.1	F-1	0.148	0.010	0.094	0.168	0.015	0.094
		Precision	0.186	0.013	0.135	0.212	0.020	0.135
		Recall	0.122	0.008	0.084	0.140	0.012	0.084
	0.2	F-1	0.172	0.015	0.098	0.141	0.010	0.098
		Precision	0.220	0.020	0.142	0.178	0.013	0.142
		Recall	0.141	0.012	0.088	0.117	0.008	0.088
	0.3	F-1	0.172	0.015	0.098	0.154	0.010	0.094
		Precision	0.220	0.020	0.142	0.195	0.013	0.135
		Recall	0.141	0.012	0.088	0.128	0.008	0.084
Topic 3	0.1	F-1	0.194	0.014	0.131	0.208	0.032	0.135
		Precision	0.204	0.015	0.141	0.218	0.034	0.145
		Recall	0.185	0.013	0.124	0.199	0.030	0.128
	0.2	F-1	0.192	0.023	0.122	0.222	0.045	0.131
		Precision	0.204	0.025	0.132	0.231	0.049	0.141
		Recall	0.182	0.021	0.116	0.214	0.042	0.124
	0.3	F-1	0.186	0.018	0.127	0.208	0.036	0.135
		Precision	0.197	0.020	0.136	0.218	0.039	0.145
		Recall	0.177	0.017	0.120	0.199	0.033	0.128
Topic 4	0.1	F-1	0.132	0.033	0.089	0.125	0.022	0.089
		Precision	0.195	0.050	0.160	0.184	0.034	0.160
		Recall	0.099	0.024	0.080	0.094	0.016	0.080
	0.2	F-1	0.141	0.033	0.098	0.141	0.033	0.098
		Precision	0.207	0.050	0.176	0.207	0.050	0.176
		Recall	0.107	0.025	0.088	0.107	0.025	0.088
	0.3	F-1	0.137	0.023	0.093	0.151	0.034	0.107
		Precision	0.195	0.034	0.168	0.218	0.050	0.192
		Recall	0.106	0.017	0.084	0.116	0.026	0.096
Topic 5	0.1	F-1	0.231	0.035	0.122	0.230	0.022	0.122
		Precision	0.244	0.037	0.131	0.244	0.023	0.131
		Recall	0.220	0.033	0.116	0.217	0.020	0.116
	0.2	F-1	0.233	0.031	0.122	0.171	0.017	0.114
		Precision	0.244	0.032	0.131	0.171	0.018	0.122
		Recall	0.223	0.029	0.116	0.172	0.017	0.108
	0.3	F-1	0.222	0.022	0.131	0.163	0.009	0.110
		Precision	0.226	0.023	0.140	0.165	0.009	0.118
		Recall	0.219	0.021	0.124	0.162	0.008	0.104