Homework #8

Bloom Filters, Time Warp, Graphs I

Anastasiia Konoplina

---

1. Bloom filters - generate 1M random 64 bit integers and store these in a Bloom filter of size 16M bits (2 bytes per long integer). Create multiple hash functions to store these keys in 16M bit array. Now create 100M new random integers (not in the original 1M) and estimate what is the rate of "false alarms" with different Bloom filter settings. Experiment with the nr of different hash functions - 1, 2, 3, X (the more you use, the higher is the ratio of bits that are set to 1 in the Bloom filter; but also the higher the chances to hit a 0 for false keys). Report your findings and argue what is "the optimal" number of hash functions for such Bloom filters.

---

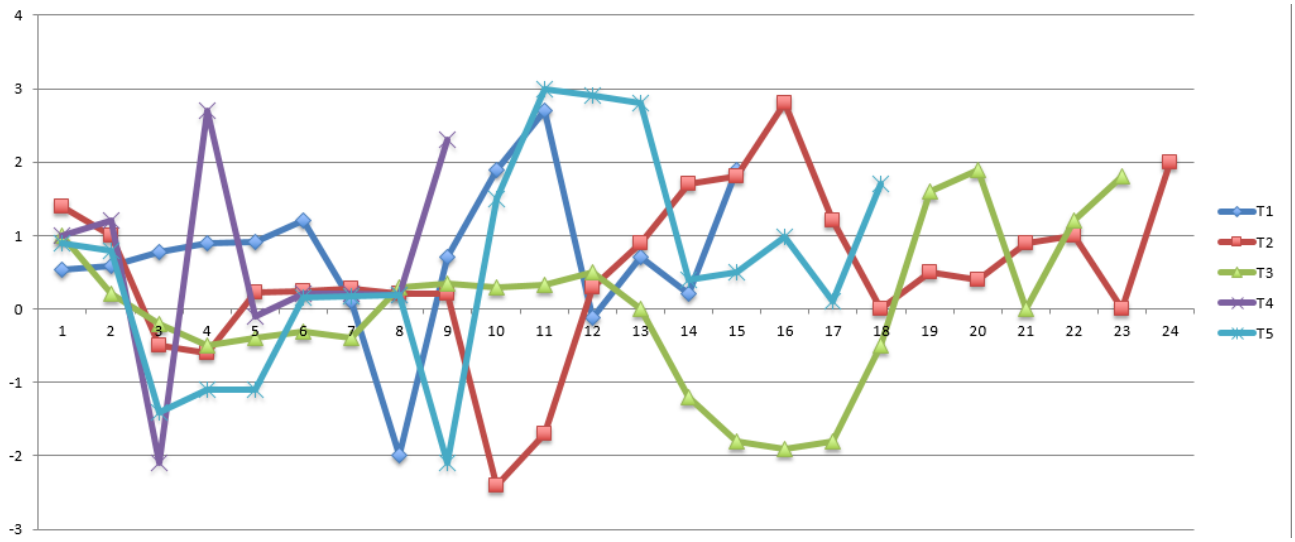
2. Compare the five series T1..T5 between each other using Dynamic Time Warping (DTW) algorithm (but you can experiment with the minimization function). Calculate all pairwise similarities. Which series are most similar?

The results are below:

	T1	T2	T3	T4	T5
T1		8.29	10.56	6.22	8.53
T2			10.32	18.33	7.96
T3				19.24	13.92
T4					13.65
T5					

The most similar are T1 and T4.

3. Visualize the time series as line graphs (as in above Excel screenshot) and respective DTW table of most similar pairs (use Excel to highlight the DTW table as was done for edit distance).



	0.54	0.58	0.78	0.89	0.92	1.2	0.1	-2	0.7	1.9	2.7	-0.12	0.7	0.2	1.9
1	0.46	0.42	0.22	0.11	0.08	0.2	0.9	3	0.3	0.9	1.7	1.12	0.3	0.8	0.9
1.2	0.66	0.62	0.42	0.31	0.28	0	1.1	3.2	0.5	0.7	1.5	1.32	0.5	1	0.7
-2.1	2.64	2.68	2.88	2.99	3.02	3.3	2.2	0.1	2.8	4	4.8	1.98	2.8	2.3	4
2.7	2.16	2.12	1.92	1.81	1.78	1.5	2.6	4.7	2	0.8	0	2.82	2	2.5	0.8
-0.1	0.64	0.68	0.88	0.99	1.02	1.3	0.2	1.9	0.8	2	2.8	0.02	0.8	0.3	2
0.2	0.34	0.38	0.58	0.69	0.72	1	0.1	2.2	0.5	1.7	2.5	0.32	0.5	0	1.7
0.2	0.34	0.38	0.58	0.69	0.72	1	0.1	2.2	0.5	1.7	2.5	0.32	0.5	0	1.7
0.19	0.35	0.39	0.59	0.7	0.73	1.01	0.09	2.19	0.51	1.71	2.51	0.31	0.51	0.01	1.71
2.3	1.76	1.72	1.52	1.41	1.38	1.1	2.2	4.3	1.6	0.4	0.4	2.42	1.6	2.1	0.4

	0.54	0.58	0.78	0.89	0.92	1.2	0.1	-2	0.7	1.9	2.7	-0.12	0.7	0.2	1.9
1	0.46	0.88	1.1	1.21	1.29	1.49	2.39	5.39	5.69	6.59	8.29	9.41	9.71	10.51	11.41
1.2	0.66	1.08	1.3	1.41	1.49	1.29	2.39	5.59	5.89	6.39	7.89	9.21	9.71	10.71	11.21
-2.1	2.64	3.34	3.96	4.29	4.43	4.59	3.49	2.49	5.29	9.29	11.19	9.87	12.01	12.01	14.71
2.7	2.16	4.28	5.26	5.77	6.07	5.93	6.09	7.19	4.49	5.29	5.29	8.11	10.11	12.61	12.81
-0.1	0.64	1.32	2.2	3.19	4.21	5.51	5.71	7.61	5.29	6.49	8.09	5.31	6.11	6.41	8.41
0.2	0.34	0.72	1.3	1.99	2.71	3.71	3.81	6.01	5.79	6.99	8.99	5.63	5.81	5.81	7.51
0.2	0.34	0.72	1.3	1.99	2.71	3.71	3.81	6.01	6.29	7.49	9.49	5.95	6.13	5.81	7.51
0.19	0.35	0.73	1.31	2	2.72	3.72	3.8	5.99	6.5	8	10	6.26	6.46	5.82	7.52
2.3	1.76	2.07	2.25	2.72	3.38	3.82	5.92	8.1	7.59	6.9	7.3	8.68	7.86	7.92	6.22

The excel file with calculations are available here on the second tab.

4. Study what are these data about. See readme files and original sources. Note that they are in slightly different formats. Better reformat first. And you can first ignore the weights or types of edges.

In "students" there is directed graph which shows social media interaction between 180 students while working on assignments.

In "karate" the data represents ties between karate club members.

In “Email-Enron” directed graph representing e-mail exchange among users.

In “usairport” directed graph showing flight between airports in USA, where weight – number of passengers.

---

5. *Install and use Cytoscape - http://www.cytoscape.org/ - try to visualise some graphs (from Cytoscape or even better, if from task 4) yourself. Show some screenshot from your computer. Identify key features and capabilities of Cytoscape.*

I have tried to visualize “karate” graph (screenshot is below).

Key features:

- Import table/network from file/URL
- Analyze graph (shortest path, distributions of degrees, number of shared neighbors, neighborhood connectivities, average clustering coefficients etc)
- Merge graphs/tables
- Apply different layouts