FALL 2019: CAP 5768 – Intro to Data Science

[Exam Review]

Problems

1. (Lec 11) Define *Jaccard Similarity* for two sets, S and T.
2. (Lec 11) Define the "bag of words" concept.
3. (Lec 11) *Bloom Filters* can test for membership. Write down the formula for false positive rate. Assume that n is the length of the bit array used for the fliter; the size of the set (universe) whose membership is being tested by the filter is of size $m < n$; and k is the number of hash functions being applied by the bloom filter.
4. (Lec 17) Compute one entire iteration of K-means assuming that at the start of the iteration the cluster centers are at (1,0), (0,0) and (0, -1), and that the input points are given by: $S = \{(-1,1),(-2,2),(-2,1),(1,1),(2,2),(2,1),(1,-2),(2,-1),(1,-3),(-1,-4),(-4,-1)\}$.
5. (Lec 17) In hierarchical clustering, identify the pair of points that will be clustered together. If more than one pair is possible, report any one of them. You can assume that the input consists of the point set S given above.
6. (Lec 20) Given the following time series, compute the *moving average* and the *error* for window size = 4: $S = \{-1,-2,-2,1,2,2,1,2,1,-1,-4,-1\}$.
7. (Lec 21) What transformation should I use to achieve near normality if my original data is a collection of counts?
8. (Lec 22 and 25) Write down the equation obtained by decomposing a matrix using the *Singular Value Decomposition* (SVD). Describe the matrices you have in the equation.
9. (Lec 23) Explain what happens on a *random walk* when the graph has two dense subsets of vertices with only one vertex in common.
10. (Lec 23) If the probability vector of being in any vertex at time t is given by the vector $\mathbf{v}$ and if the transition probability matrix is given by a $n \times n$ matrix M, explain what happens to the probability vector of being in any vertex at time $t+1$.
11. (Lec 24) Using *Bayes' Rule* for random variables A and B, write $Pr(A|B)$ in terms of $Pr(B|A)$?
12. (Lec 24) Explain the difference between *common cause* and *common effect*.
13. (Lec 25) Given the 4 documents below, compute the term document matrix. List the stop words you are using.

    D1: You only live once, but if you do it right, once is enough. Mae West.
    D2: Life shrinks or expands in proportion to one's courage. Anais Nin.
    D3: I could die for you. But I coudn't, and wouldn't, live for you. Ayn Rand
    D4: Moral courage is to die for. Don't need no shrink to tell me that. Anon.

14. Compute the TF $\times$ IDF for the terms in the above documents that are not stop words.