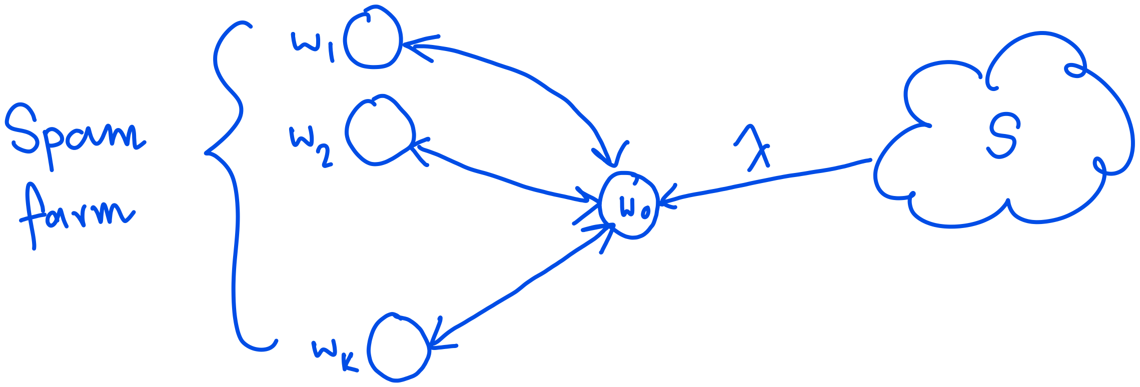
CS328: Homework 2

The usual agreement with respect to collaboration being okay (mention names of collaborators) buy not copying. Other aspects of honor code are also to be followed. All code needs to be written by self. Code to be submitted using colab as usual. Also give names of collaborators.

1. Suppose that youve already computed the personalized PageRank vectors of a set of users (denote the computed vectors V). What is the set of all personalized PageRank vectors that you can compute from V without accessing the web graph?

2. Consider the design of a spam farm as shown below. The farm has pages $\{w_1, \ldots, w_k\}$ and the target page is w_0. The rest of the web is S. λ denotes the amount of pagerank that flows in through edges from S to p_0, i.e. $\lambda = \sum_{j \in S} \frac{p_j}{d_j}$, where p_j is the pagerank of page j in S and d_j is its outdegree. Let $N = |S| + k + 1$, i.e. the number of pages in entire web. Write down pagerank p_0 of the target page in terms of λ, k, N as well as α (the teleportation parameter). Notice that λ does not include the teleportation, but p_0 should.



3. Suppose you have a turnstile stream of n distinct items. Suppose frequency distributions follow a power law with exponent 3, i.e. number of items with frequency k is C/k^3. What should roughly C be (order notation)? If you fix w and d, does CM or CS give a better guarantee for this distribution.

4. Implement CM sketch, CS sketch and Misra-Gries sketch, all should use the same size. For CM and CS, fix $d = 5$. Choose a k for MG and set $w = k/d$ for CM and CS. Try all $k \in \{100, 200, 500, 1000, 2000\}$.

   Compare their performance as follows. The input data is should be the entire train.data obtained from http://qwone.com/ jason/20Newsgroups/. Use the Matlab file format (actually just a text-file), throwing away the docIdx. That is, the stream is just tuples (wordIdx, count).

   For preparing the query data, find out the exact frequencies. Then choose 100 words at random out of the top 1000 most frequent words. These are your query set. Find out the relative error aeveraged over the 100 queries i.e.

   $$\frac{1}{100} \sum_{x \leq 100} |f_x - \hat{f}_x| / f_x,$$

   where f_x is true frequency and $\hat{f}_x$ is estimated by the sketch. There should be 3 curves, one for each of the sketches. Report the numbers in a table and in a plot.

   For each of the sketches, also report the minimum w that you need for the average error over the 100 queries is less than 1%.

5. Consider the https://grouplens.org/datasets/movielens/ We will use a low-rank approximation of to estimate the entries that are queried for in . Follow the procedure below.

```
    Train = 80% of set of labeled triplets. Denoted as (i, j, T(i,j))
    Test = 20% of set of labeled triplets. Denoted as (i, j, S(i, j))
```

```
Represent Train as a matrix and find a rank-k approximation to Train.
Let this matrix be named Pred.
For every triplet (i, j, S(i,j)) in Test:
   err = err + (S(i, j) - Pred(i, j))^2
Return err
```

Plot k versus the error for k in a range $(0, 100)$. Compare this with the following baseline algorithm: every test entry $(i.j)$ is predicted as $\alpha * \mu_i + \beta * \eta_j$ where μ_i is the average rating of user i, and η_j that of movie j over all ratings in , α and β are fitted using training data.