CME 323: Distributed Algorithms and Optimization, Spring 2020
`http://stanford.edu/~rezab/dao`.
Instructor: Reza Zadeh, Matroid and Stanford.

Lecture 18, 6/4/2020. Scribed by Robin Brown.

18 DIMSUM

We'd like to compute entries for $A^T A$ for which $\cos(r_i, r_j) \geq s$ for some threshold s. Columns of A are vectors, and vectors can have similarities. We need the following notion of similarity of two vectors [Cosine Similarity] The *cosine similarity* between two columns c_i and c_j is defined as

$$\cos(c_i, c_j) = \frac{c_i c_j}{c_i c_j}.$$

Algorithm 1 DIMSUMMapper (r_i)

```
for all pairs $(a_{ij}, a_{ik})$ in $r_i$ do
    With probability $\min\left\{1, \frac{}{c_j c_k}\right\}$, emit $((j,k) \to a_{ij} a_{ik})$
```

Algorithm 2 DIMSUMReducer $((i,j), v_1 \ldots, v_R)$

```
if $\frac{}{c_i c_j} > 1$ then
    return $b_{ij} \to \frac{1}{c_j c_j} \sum_{i=1}^{R} v_i$
return $b_{ij} \to \frac{1}{} \sum_{i=1}^{R} v_i$
```

The *Dimension Independent Matrix Square using MapReduce* (DIMSUM) algorithm is described in (1) and (2). The DIMSUM algorithm outputs the cosine similarities (in fact probabilistic estimates of the cosine similarities). Also note that you need to compute the norms of columns beforehand (which requires all-to-all communication).

References

[1] MapReduce-Combiners. Retrieved from `http://www.tutorialspoint.com/map_reduce/map_reduce_combiners.htm`.

[2] Broadcast Variables. Retrieved from `https://jaceklaskowski.gitbooks.io/mastering-apache-spark/content/spark-broadcast.html`.