**Pairwise Document Similarity in Large Collections with MapReduce**

Computing pairwise similarity on large document collections is a task common to a variety of problems such as clustering and cross-document coreference resolution. This paper considers a large class of similarity functions that can be expressed as an inner product of term weight vectors.

**Pairwise Document Similarity:**

The work focuses on the document similarity metrics. document d is represented as a vector Wd of term weights wt, d, which indicate the importance of each term t in the document. We take a term that contributes to each pair that contains it. For example, if a term appears in documents x, y, and z, it contributes only to the similarity scores between (x, y), (x, z), and (y, z). The list of documents that contain a particular term is exactly what is contained in the postings of an inverted index. Thus, by processing all postings, we can compute the entire pairwise similarity matrix by summing term contributions. Two efficient solutions has been proposed to Pairwise Document Similarity.

1. **Indexing:**

Each term is associated with a list of docid’s for documents that contain it and the associated term weight. Mapping over all documents, the mapper, for each term in the document, emits the term as the key, and a tuple consisting of the docid and term weight as the value. The MapReduce runtime automatically handles the grouping of these tuples, which the reducer then writes out to disk, thus generating the postings

1. **Pairwise Similarity:**

Mapping over each posting, the mapper generates key tuples corresponding to pairs of docids in the postings: in total, 1 2m(m−1) pairs where m is the posting length. These key tuples are associated with the product of the corresponding term weights—they represent the individual term contributions to the final inner product. The MapReduce runtime sorts the tuples and then the reducer sums all the individual score contributions for a pair to generate the final similarity score.