In this project, we designed it in the following way:

First we take the ngrams and the gold standard data set, then with mapreduce we create a vector for each pair of words in the gold standard data set from features that we calculated using the ngrams. The next step is to run Weka on the vectors while using 90% for classifying and 10% for testing.

In the map reduce part we have 6 steps:

1. Word count: in this step, we count each word with its type and count the total amount of words in the ngrams
2. Word indexing: in this step, we create an index for each word where the index is the words spot in the feature vectors
3. Word association: in this step, we create a list of words that have a co-occurrence with a specific word in the gold standard data set for each word. In this step we load the gold standard data set into the cache to enable screening of words that we don’t need.
4. Feature vectors: in this step, we take the information from the previous step to create 4 vectors for each word in the gold standard data set according to features. In this step we load the output of the indexing step to merge the features with the indexes.
5. Vector pairing: In this step, we merge 2 words that got paired in the gold standard data set and connect them to their vectors. In this step we load the gold standard data set to check what pairs we have. (the file is already in the cache)
6. Final vector: in this step, we check the vector similarities using 5 methods on all 8 vectors (4 for each word) and create a vector of size 20 with the results.

After we get the results of the map reduce we run Weka as described before and analyze the results. During the whole process when the step input is the ngrams we create about 3 times the size of ngrams pairs during the other steps we create (thanks to screening) | gold standard data set| pairs for the vectors.

In a few of the steps we use the cache for loading the indexes for the vectors and calculating the similarity and the gold standard data set for screening un necessary vectors.