# IR Assignment-2

## Design Document

1. **Translation Model:**

   - **Pre-Processing:** The model is trained with the given corpus. For preprocessing, Map-Reduce Technique is implemented because we observed that the size of the corpus is too big (1.8 million sentences and around 100,000 different words for English as well as French document. So it's impossible to import all the data at once because if we do so, it exceeds the memory limit of laptop or PC (around 8 GB). That's why we processed the corpus in batch of 18000 files each containing 100 sentences and then the Reduction Part comes into play. These 18000 files are merged into single file by using the following heuristic:

     - The contents of the files are compared, for every respective (e,f) pair's probability then maximum among them is taken else we simply take into account.

     At the end we are left with one file containing all the translations predicted by IBM model 1.

2. **Data Structures:** We have used Python 3 for the purpose of implementation and for storing the (e,f) pairs, multi-level HashMaps (Dictionaries) are used.

3. **Cosine Similarity:** The 2 documents are processed and their tf vector is calculated and then the vectors are normalized and the angle is computed between them.