

Part I: Word and Sentence embeddings with Word2vec:

1. The total number of raw words find in the corpus is 85026035.