Gensim is a Python library for topic modelling, document indexing and similarity retrieval with large corpora. For this project the similarity retrieval is used. The vectors of length 300 are trained corresponding to each tweet. Each tweet is treated as a document or sentence with a list of words.

The input to this code is the text file which contains tweets in a simple format: one document = one line = one TaggedDocument object. Words are expected to be already pre-processed and separated by whitespace; tags are constructed automatically from the document line number. Training is done using Distributed Memory Model (give link to the research paper titled Distributed Representations of Sentences and Documents). The trained model is saved for further use.