Analysis of the Embedding Generator

Current Design:

Used the Gensim word2vec library for generating the word embeddings. To save upon computational efforts and training time, currently we've used the pre-trained word embeddings generated by Google. Link: https://code.google.com/archive/p/word2vec/

Approach for preprocessing includes the following steps:

1. Removing Punctuations
2. Converting to lowercase
3. Removing stopwords from the text. The stopwords used are imported from the Gensim library.
4. Tokenization
5. Lemmatization

Given a file, extract the text, and preprocess it to get only the "important" words describing the gist of the document. Then get the word embeddings for every word in the above mentioned list and take the mean over all the word embeddings obtained.

Results Obtained:

Sample Text used:

Well done, Javier, on all your hard work this term. You are always active in speaking activities, and this is great. I always enjoy reading your writing too; you have very good ideas. You can see from your marks that reading is not a problem for you.
Sometimes you still make small grammar mistakes, and I think you can improve your vocabulary. I recommend you review many of the language points we studied this term. There is extra language practice in your online workbook.
The area you need to work on the most is listening. I know this is difficult for you. I recommend more practice at home. Listen to English TV shows, podcasts and radio as much as possible. I can give you a list of things to listen to.
I've enjoyed working with you. Have a nice holiday, and good luck for next term!

Words most similar to the embedding generated by our code:
('KIM_CLIJSTERS_Yeah', 0.6645171642303467), ('GREG_POTTER_Yes', 0.6537461280822754), ('pioneer_LaLanne', 0.6527040004730225)

Analysis:

The results obtained are far from expectations in the sense that, they do not represent the overall meaning of the text.