Analysis of Embedding Generator

12.05.2022

—

Dhairya Parekh

Naman Singh Rana

Proshort.AI

Overview

We are trying to create an algorithm to create/generate D-dimensional vector embeddings for videos. Current approach takes into consideration only the transcript of the video and the context and embeddings are generated based on the text of the transcript.

Current Design

1. Given a file containing the transcript, Extract the text
2. Preprocess the Text to get only the “important” words describing the gist of the document.
3. The 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
4. Get the word embeddings for every word in the above mentioned list
5. 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/

Results Obtained

Sample Text Used:

Well done, 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!

Output:

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.

I. Alternatives Tried

Used different libraries, namely NLTK and Spacy for removing stop words in text. Used stemming instead of lemmatization in step 5 of preprocessing. Tried on other measures of central tendency like median and mode but didn't get any better results.

II. Suggestions and Other approaches

Instead of using the whole transcript for generating video embeddings, we can take video tags and generate the video embeddings from the word2vec embeddings of the tags. Also, we can find some other libraries to find the vector embeddings for the audio (Voice) and maybe integrate both the vector embeddings.