DATA.STAT.840 Statistical Methods for Text Data Analysis

Exercises for Lecture 12: Paragraph embedding

Exercise 12.1: Paragraph embedding of four newsgroups.
Using the same data as in exercise 6.1 (four newsgroups), use paragraph embedding to create a vector for each document, using the Python implementation discussed on the lecture (lecture 12, slides 12-13), or using a different language/library of your choice.
Then find, by Euclidean distance, the closest other documents for the following documents: 101551 (part of rec.autos), 103118 (part of rec.motorcycles), 98657 (part of rec.sport.baseball) and 52550 (part of rec.sport.hockey). Are the closest documents from the same newsgroup? Does their content seem to match that of the original document?
Report your code, results, and discussion.