Python ML Machine Learning Project

Text Plagrarism

Overview

Plagiarism detector to be able to load students'/persons work from files in .txt format and then compute the similarity to determine if there is a chance that the students have copied.

Value output from 0-1 with 1 having a higher likelihood of copying.

How will the text be analysed?

So, first of all we need to convert our text input into an output of 0,1 s by using some sort of computation on textual data.

Word embedding

So, we can use word embedding to convert textual data into an array of numbers.

Detecting similarity in .txt files?

Using the concept of vector dot product to determine how closely the two different .txt files are related by computing the value of cosine similarity between vectors representations of student's text assignments.

So, why am I using Cosine Similarity?

So the basic concept is that we have two vectors and we work out the angle between them. For example, two vectors with the same orientation have a cosine similarity of 1, two vectors oriented at right angle relative to each other have a similarity of 0, and two vectors diametrically opposed have a similarity of -1.

Project Review

The program runs successfully and I am given outputs based upon the similarity of the .txt files.

One thing I would like to extend this concept to is to work out similarity between images, much similar to that used in google reverse image search.

And, I can use TensorFlow again to do this.