
RH

REVA HACK</> 2021

Elevator Pitch

AFTERMATH

Abhay Pratap Singh
Sudhanva M

DOCUMENT SIMILARITY

12th November, 2021

Overview

Document similarity is one of the central themes in Information Retrieval. Usually documents are treated as similar if they are semantically close and describe similar concepts. On the other hand “similarity” can be used in the context of duplicate detection.

So, to overcome complexity in matching documents and making the process easier, document similarity algorithms are used in programming.

Goals

1. Checks if a document is plagiarized and to what extent.
2. Can be used to fact check news and related documents.
3. Can be used by educational faculties to evaluate answer scripts.
4. Can be used in information retrieval scenarios.

Working Methodology

File Uploading :-

- The user opens the page containing a HTML form featuring a text files, a browse button and a submit button.
- The user clicks the browse button and selects a file to upload from the local PC.
- The selected file is sent to the temporary directory on the server.
- The PHP script that was specified as the form handler in the form's action attribute checks that the file has arrived and then copies the file into an intended directory.
- The PHP script confirms the success to the user.

Files Comparison :-

- Cosine similarity is a metric used to measure how similar the documents are irrespective of their size.
- The cosine similarity is advantageous because even if the two similar documents are far apart by the Euclidean distance (due to the size of the document), chances are they may still be oriented closer together.

Specifications

Server-Side

Operating System : Windows 7 or newer / OS X / BSD / Debian and Arch

Web Server : Apache HTTP Server (2.4.46)

Programming Languages : PHP (7.3.21) for frontend ; Python (2.7) for backend

Client-Side

Browser which supports php , html , css , javascript

Links and other Information:

- https://www.loom.com/share/a1b525b8f6714a27abe0e2aed534a421 (Visual Representation)
- https://github.com/sud0x00/DS-Checker (Github)