Info Retrieval - Plagiarism Checker

Checks documents for plagiarism with respect to a corpus. The Algorithm used for the same is Bag of Words along with Cosine Similarity for score calculation.

Installation

Requirements

- Python 3.8+
- Natural Language ToolKit
- NLTK Popular Model Collection
- Source Code

Note: Source Code can be `git clone`ed if not already present.

Usage

- `$ cd "Path/to/cloned/repo"`
- `$ python main.py -h`

Indexing

- Make a folder with only the files that are meant to be included in the corpus
- `$ python main.py index "path/to/folder/made/above"`
- Index can be updated by running the same command with updated corpus folder.

Querying

- Make sure the index has been previously made

Stdin

- `$ python main.py query -`
- Type in the query document in the terminal now

Single File

- `$ python main.py query "path/to/file"`

Multiple Files

- `$ python main.py query "path/to/file1" "path/to/file2"`

Documentation

- Make sure you have the `Documentation` folder.
- Open `Documentation/index.html` in your browser.