README

Plagiarism Detector

This is a plagiarism detection tool built using the locality sensitive hashing (LSH) algorithm with minhash.
The tool uses Python datasketch library for calculating MinHash and MinHashLSH.

Prerequisites:

- NLTK Package
- Time Package
- Text Distance Package
- Datasketch
- Redis

Usage

Environment setup:
Specify the corpus location in `lsh_main.py`
Run `lsh_main.py` to preprocess the corpus

Testing:
Open the project. In `query.py` script specify `check_file` for the file to be checked for plagiarism, and run the script. Output: message about whether the document is plagiarized or not. If the document is plagiarized, the list of source document names is printed. Also, distance of query document with each of the documents in the list using different measures is printed.

Project structure

Data - folder with data files. Subfolders and files:
    true_positives - contains documents that are correctly marked as plagiarised.
    true_negetives - contains documents that are correctly marked as not plagiarised.
eval.py - evaluation script and precomputed variants of LSH for the document database.

lsh_Main - script for calculating LSH for the document database and precomputed variants of LSH.
query.py - script for checking new document for plagiarism.
utils.py - file with useful functions and global variables that are (or may potentially be) in different parts of the project.