EECS 767

INFORMATION RETRIEVAL

Progress Report

By: FiniteLoop Squad

Ron Andrews, Nidhi Midha, Blake Bryant

Table of Contents

[1. Introduction 3](#_Toc508453004)

[2. Programming Platform and Version Control Selection 4](#_Toc508453005)

[3. Data Structures 4](#_Toc508453006)

[3.1. Pre-Processing to Processing Interface Data Structures 4](#_Toc508453007)

[3.2. Processing to Query Processing Interfaces Data Structures 5](#_Toc508453008)

[3.3. Query Processing to HMI Data Structures 6](#_Toc508453009)

[4. Ingest 6](#_Toc508453010)

[5. Preprocessing 7](#_Toc508453011)

[5.1. Tokenization 7](#_Toc508453012)

[5.2. Indexing 7](#_Toc508453013)

[6. Processing 8](#_Toc508453014)

[7. Query 8](#_Toc508453015)

[8. Appendix 9](#_Toc508453016)

[8.1. Group Log 9](#_Toc508453017)

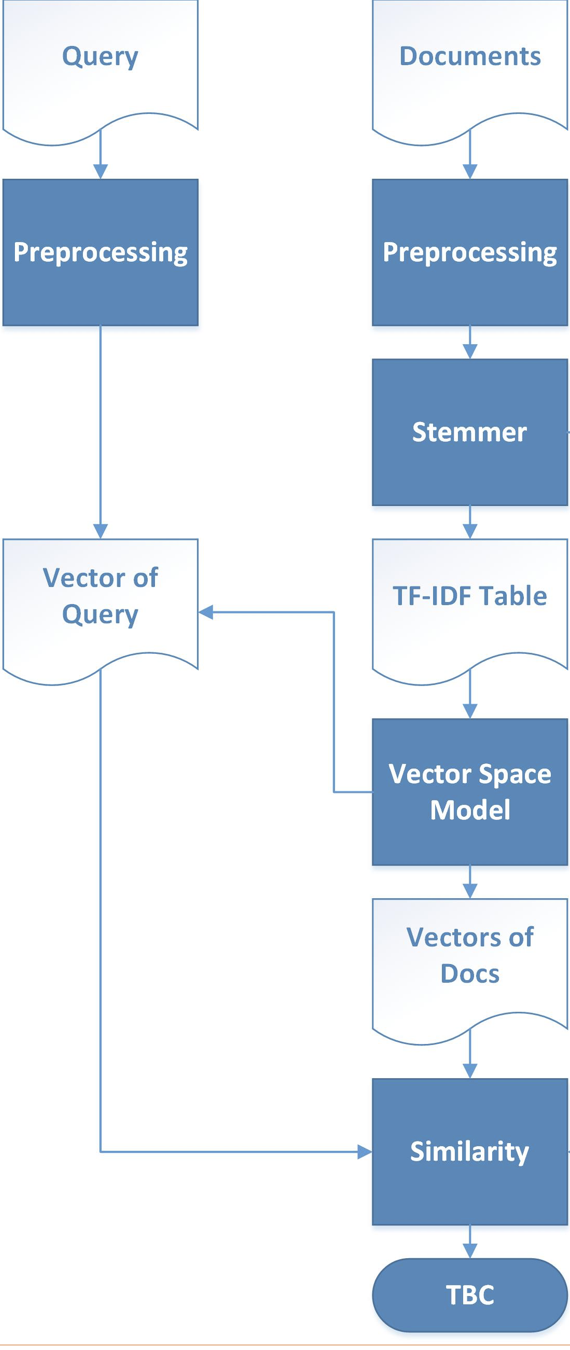
[8.2. GitHub Contributions 10](#_Toc508453018)

[8.3. GitHub ChangeLog 10](#_Toc508453019)

# Introduction

Our project goal is to create a simple Information Retrieval System (Search Engine) using Vector Space Model. We would add optimization techniques such as term proximity, Relevance Feedback, and test the product on a practical data set obtained by crawling a specific domain with Niche Crawler.

In phase 1, we have finished the implementation of document processing, inverted index and vector space model. We are currently working over adding term proximity as a consideration of displaying the results. Figure, below, shows our current progress.



# Programming Platform and Version Control Selection

In review of the various programming options available, we focused on those languages which were most capable, natively for this project. Specifically, looking at those languages which supported complex functions such as cosine similarity, web compatibility (Common Gateway Interface, CGI, or apache server module based), and of course, familiarity. After consideration of various options, such as *R*, *Perl*, *C++*, and *Python*, we selected *Python* as our language of choice.

With the current versions available for *Python*, we initially selected to go with version 3.6, being the latest available. As we worked through the various modules of our search engine (pre-processing, processing, query, and human machine interface (HMI), we ran into a few challenges. The Natural Language Toolkit (NLTK) that we selected to facilitate the stop list and lemmer was compatible with *Python* 3.5, not 3.6. Additionally, the Electrical Engineering and Computer Science (EECS) student web server currently provides access to *Python* 2.7 and 3.5. Our conclusion was to go forward with Python 2.7 as it was common to our individual environments as well as the web server. Additionally, we elected to use the EECS web server CGI capability for hosting our search engine.

For our collaboration environment, we set up a GitHub repository specifically for our **FiniteLoop** **Squad** to work and share. In the environment, we are able to coordinate our code development efforts as well as documentation.

# Data Structures

In order to pass the data structures between our modules, we are leveraging a *Python* module called *shelve*. This native module enables us to pass the raw data structures by way of a binary file stored on the file server. The following sub-sections provide the data structures passed between the modules.

# Pre-Processing to Processing Interface Data Structures

Pre-Processing provides three data structures to the Processing module:

1. Document Key Matrix
2. Term Incidence Matrix, with Frequency
3. Term Proximity Matrix

The Document Key provides a listing of the details for each document as a *Python* dictionary where the document name is the key and the details is a list of values. The details for each document include the document identifier and current location of the document. The order of the list of documents matches the order of the documents in the Term Incidence Matrix.

The generalized data structure looks like the following, in *Python* terms:

doc\_key = [

{ DocName1: [DocID1, DocLocation1] },

{ DocName2: [DocID2, DocLocation2] },

…,

{ DocNamen: [DocIDn, DocLocationn] }

]

The Term Incidence Matrix provides a listing of each term and its occurrence in the corpus as a *Python* dictionary where the term is the key and the document incidence with frequency is a list. Each document incidence list is aligned in order with the list provided in the doc\_key data structure.

The generalized data structure looks like the following, in *Python* terms:

index = [

{ Term1: [tf1, tf2, …, tfn] },

{ Term2: [tf1, tf2, …, tfn] },

…,

{ Termm: [tf1, tf2, …, tfn] }

]

The Term Proximity Matrix provides a dictionary of each term, as the key, and a list of tuples as the value. The tuples identify the document and offset from the beginning of the document. Offsets are based on word distance from the beginning of the document after the tokenization and stop word parsing is complete.

The generalized data structure looks like the following, in *Python* terms:

proximity = {

Term1: [ [DocID, Prox], [DocID, Prox], … [DocID, Prox] ],

Term2: [ [DocID, Prox], [DocID, Prox], … [DocID, Prox] ],

…,

Termm: [ [DocID, Prox], [DocID, Prox], … [DocID, Prox] ]

}

# Processing to Query Processing Interfaces Data Structures

Processing provides three data structures to the Query Processing module:

1. Document Key Matrix
2. Normalized Vector Space Model (VSM)
3. Term Proximity Matrix (InWork)
4. Term Index Look-Up Dictionary (InWork)

The Document Key Matrix is forwarded, unaltered from what was received from the Pre-Processing module.

The Normalized VSM provides an alphabetically sorted list of vectors (lists). Each vector is in order as identified by the Term Index Look-Up Dictionary and each vector is in order of the Document Key Matrix. The vectors provide the normalized Term Frequency – Inverted Data Frequency (TF-IDF) weight of the term for each document.

The generalized data structure looks like the following, in *Python* terms:

docVector = [

[WT1,D1, WT1,D2, …, WT1,Dn],

[WT2,D1, WT2,D2, …, WT2,Dn],

…

[WTm,D1, WTm,D2, …, WTm,Dn]

]

The Term Proximity Matrix provides a listing similar to the Normalized VSM where instead of the weights, it provides a list of the proximities.

The generalized data structure looks like the following, in *Python* terms:

proxVector = [

[ [P1T1,D1, P2T1,D1, …, P*i*T1D1], [P1T1,D2, P2T1,D2, …, P*i*T1D2], …, [P1T1,D*n*, P2T1,D*n*, …, P*i*T1Dn] ],

[ [P1T2,D1, P2T2,D1, …, P*i*T2D1], [P1T2,D2, P2T2,D2, …, P*i*T2D2], …, [P1T2,D*n*, P2T2,D*n*, …, P*i*T2Dn] ],

…,

[ [P1T*m*,D1, P2T*m*,D1, …, P*i*T*m*D1], [P1T*m*,D2, P2T*m*,D2, …, P*i*T*m*D2], …, [P1T*m*,D*n*, P2T*m*,D*n*, …, P*i*T*m*Dn] ]

]

The Term Index Look-Up Dictionary provides a dictionary of each term, as the key, and an index into the Normalized VSM and Term Proximity Matrix lists for efficient look-up of the vectors needed in searching for results based on a query.

The generalized data structure looks like the following, in *Python* terms:

termIndex = {

Term1: i1,

Term2: i2,

…,

Term*m*: i*m*

}

# Query Processing to HMI Data Structures

The data passed to the HMI is done so directly with the CGI script importing the Query module directly for dynamic processing and results. The Query module provides the following data and data structures to the HMI for display to the user:

* Total Number of Results found
* Time taken to process query and return results
* List of results, ordered by relevance

The list of results is provided to the HMI as a list of document entries. Each entry contains a list of parameters to display to the user.

The generalized data structure looks like the following, in *Python* terms:

results = [

[ DocName1, DocLocation1, Rank1, Summary1],

[ DocName2, DocLocation2, Rank2, Summary2],

…,

[ DocName*n*, DocLocation*n*, Rank*n*, Summary*n*]

]

# Ingest

Ingestion is performed by functions contained within the ingest.py file which operate on files stored within a local directory. The path to the directory is currently configured with a static path hard coded within the ingestion function. A function was created to allow for user entry of a different path location, however this functionality is currently deemed unnecessary and has been disabled via inline comments.

The native python library “urllib2” is used to read files and provide compatibility for processing various document formats including html. Each document within the specified directory is loaded into memory as a stream of raw characters and stored in an array, called “data,” with each document representing a single index within the array.

A dictionary called “doc\_key” is created to store the filename, document ID and file path for each document ingested. The document filename is used as the key for the dictionary while the value for each key is represented as an array containing the document ID and file path.

# Preprocessing

# Tokenization

Each document, stored as an index within the data array created during ingestion, is processed by the function “func\_tokenize” within the ingest.py file. The Python Natural Language Tool Kit (NLTK) (<https://www.nltk.org/>) is used to create stop word lists and a stemming function within func\_tokenize. This requires the installation of NLTK as well as downloading the stop word list prior to running the ingestion program. NLTK version 3.2.5 (the latest version as of this report) requires either Python version 2.7 or 3.5. Python 3.5 is not the most current release of Python 3 and may require some systems to install an earlier version of Python.

Preprocessing within func\_tokenize consists of 4 steps executed on each index of the data array:

1. HTML tags are removed from the data stream via regular expression pattern matching.
2. The data stream is converted to lower case, punctuation is removed, and the stream is split into tokens via the Python string.split() method.
3. Stop words contained within the NLTK stop word list are removed from the list of tokens.
4. Each token within the data stream is processed by the Porter stemmer provided by the NLTK library. There is a known issue when processing

Unicode characters with the NLTK Porter stemmer. Because of this, words containing Unicode characters are currently dropped in the stemming phase.

# Indexing

The number of documents processed during ingestion is used to determine the dimensionality of arrays within the terms dictionary. Each data stream, corresponding with a separate document, is parsed for unique terms. If a new term is discovered in a document, the term is added as a key to the dictionary called “terms” with an initial value of an array with multiple indices corresponding with the number of documents processed. The index corresponding to the document, wherein the term was observed, is then incremented by 1. Subsequent observations of the term within the same document, or future documents, results in incrementing the value stored in the term array at the index corresponding with the id of the document being parsed. This results in an array reflecting the term frequency for each term observed within each document.

Concurrently, an additional dictionary called “proximity” is created to record positional data pertaining to each occurrence of a term within the documents parsed. Similar to the “terms” dictionary, unique terms are used as key values in the dictionary. Values within the “proximity” dictionary are represented as an array containing tuples consisting of the document ID and the position the term appeared within the document. A new tuple representing the document id and term position is appended to the term array each time the term is observed within the data stream. This data structure is described in further detail within the “Data Structures” section of this report.

Finally, data is exported to a file called “ingestOutput.db” using the Python shelve library. The “terms” dictionary is exported as “index”, the “doc\_key” dictionary is exported as “doc\_key” and the “proximity” dictionary is exported as “proximity” within the output file.

# Processing

The processing module reads in the output file from the ingest function in order to acquire the pre-processing data structures. It then generates the TF-IDF, normalizes the vectors and stores the VSM for the query module. Additionally, processing generates a proximity matrix, similar to the VSM data structure and also stores it, along with the doc\_key structure into the output shelve file for use by the query function.

To accomplish this, the index data structure (the term incidence matrix) is sorted and then walked through to evaluate for the Document Frequencies of each term in order to generate the VSM. It does so by first calculating the non-zero indices in the document arrays along with calculating the idf (log n/df), where n is the length of the array (# of documents). With this data, the non-normalized weights are calculated, |Wi| = sqrt(sum(idf2)), and then added to the VSM (docVector). Finally, the module goes through the VSM and normalizes the document vectors and stores the data structure in the shelve output file for the query function.

The generation of the proximity file is currently in work.

# Query

The dictionaries “doc\_vector” and “doc\_key” are passed to query module using a file processingOutput.db which is Python shelve library. “doc\_vector” dictionary provides the alphabetically sorted list of vectors (lists). Each vector is in order as identified by the Term Index Look-Up Dictionary and each vector is in order of the Document Key Matrix. The vectors provide the normalized Term Frequency – Inverted Data Frequency (TF-IDF) weight of the term for each document. Weights in this dictionary are compared with another vector “query\_vector”.

Search query inputted by user is first preprocessed by removing the stop words and processing by the Porter Stemmer provided by NLTK library. Once processed, query tokens are stored in the query vector, whose weights are compared with the document vectors to calculate the cosine similarity. The cosine similarities are calculated using the method similarity() which sorts the similarity, and ranks the results in descending order on the basis of doc\_key associated with each document vector. The ranks are stored in queryOutput.db, which would be passed to CGI to display the results.

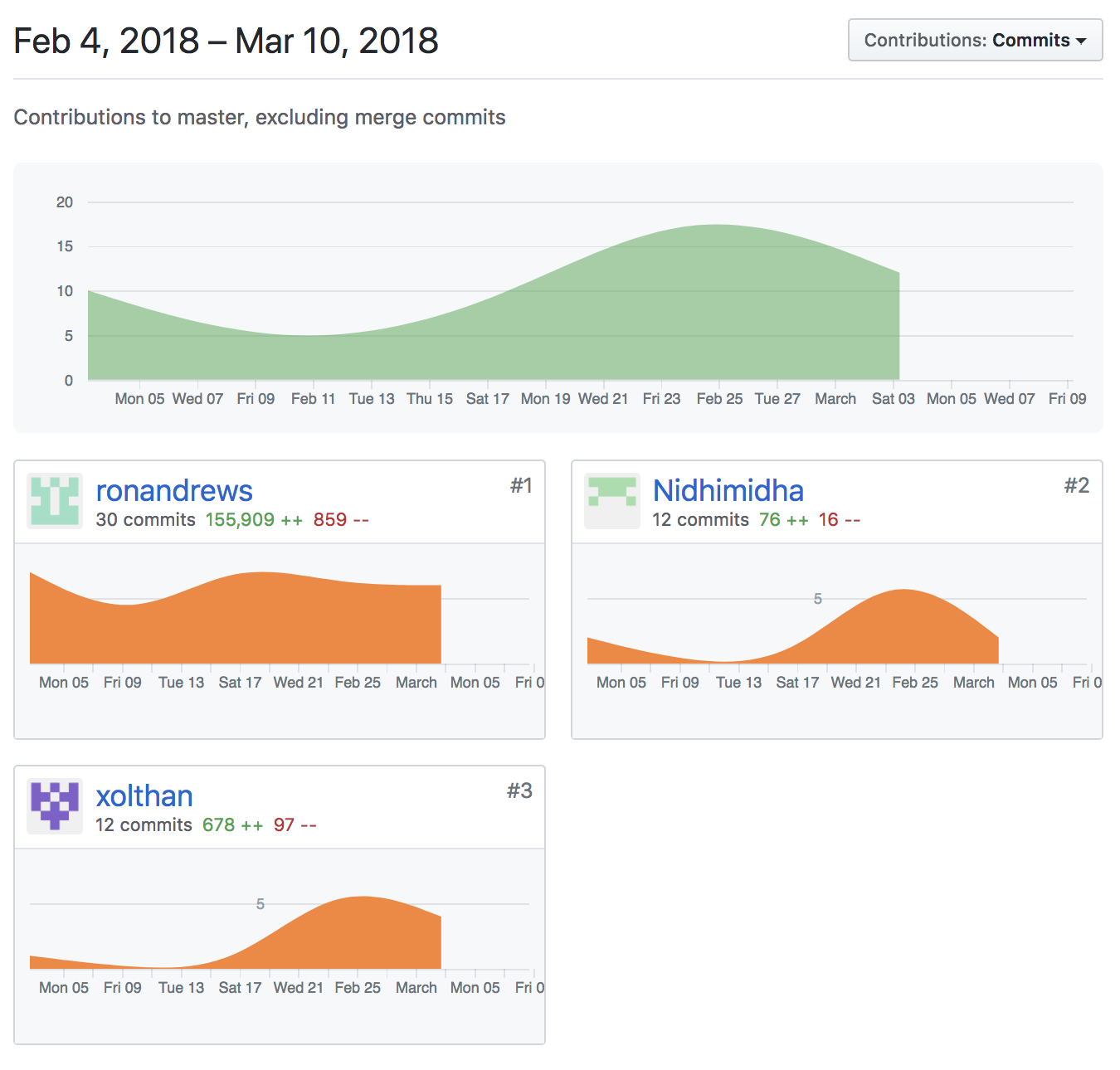
Current work which is going on includes getting the proxVector from processing module, and re-rank the top 10 documents obtained from cosine similarity, on the basis of proximities.

# Appendix

# Group Log

* 29 Jan 2018: First meeting between Nidhi and Ron - discussed separation of functionality and preferred coding language for project. Reviewed language alternatives (java, python, R, etc.) - looking for the language with the most to offer natively
* 01 Feb 2018: Notified professor on formation of group
* 05 Feb 2018: First meeting of group as three members, discussed approach to project, discussed initial outline of code (draft created by Blake). Coordinated module responsibilities and discussed I/O between modules. Selected Python for language. Started GitHub repository for collaboration environment
* 15 Feb 2018: Notified professor of additional team member and team name: FiniteLoop Squad
* 19 Feb 2018: Met to review initial code stubs and discuss next steps - working to complete simple search to reflect (and be able to check against) steps outlined in class quiz (taking in simple docs and matching with query)
* 26 Feb 2018: Met to go over ingest functionality and output - need to modify processing to accept minor change in ingest output format. Discussed front end GUI - coordinated CGI script and environment on EECS *peoples* server
* 5 Mar 2018: Met to go over current status, the progress report and functionality. Discussed implementation for term proximity ranking and feedback relevance. Team discussed data structures to pass for proximity and how feedback relevance should work in our search engine. Also assigned sections for working on the progress report.

# GitHub Contributions



# GitHub ChangeLog

commit a721d2f522ca97e7d36157a094b08e7b02bf5e65 (HEAD -> master, origin/master, origin/HEAD)

Author: Ron <ron@merehuman.net>

Added termIndex and proxVector to Processing

commit a255d61ad5aafa0064315d40929fc10441d4ede4

Merge: be5e8be 46e6696

Author: Ron <ron@merehuman.net>

Merge branch 'master' of https://github.com/ronandrews/eecs767

commit be5e8be73360f6f19ece98b344fcf860a3a0529c

Author: Ron <ron@merehuman.net>

Formatting, seeShelve, and Template

commit 46e669667f0b0d930e7bf426352619cb4bd6dfeb

Author: Nidhimidha <36167384+Nidhimidha@users.noreply.github.com>

Add files via upload

commit e89295ef9d3cf1612a262c140aa077fa2d5f5f20

Author: Ron <ron@merehuman.net>

Updated Report - Processing

commit a366aed0cacadf00fe280893448b8b52b18324ed

Author: Nidhimidha <36167384+Nidhimidha@users.noreply.github.com>

Add files via upload

commit 2781573d1709f8f1e926c46282e5ae82c820fa7f

Author: xolthan <xolthan@gmail.com>

Blake updated ingest and preprocessing sections

commit 85734d01e8da44db0aa2eef0829945d07ea80791

Author: xolthan <xolthan@gmail.com>

Now exports data via shelve to ingestOutput.db

commit 5de02bf3a740bec963c999ce34cf7e9e0260b188

Author: Ron <ron@merehuman.net>

Added Data Structures Info

commit 745d3dd3c2a64ec7d6e67b730c0b62c764bfce28

Author: Ron <ron@merehuman.net>

Notes in the template

commit 87ecf7b5e105b182f87f98a9dcd47e3cb10bb124

Merge: 53280a4 62732de

Author: Ron <ron@merehuman.net>

Merge branch 'master' of https://github.com/ronandrews/eecs767

commit 53280a435c4ea6f231fb5c72de19b0d7530e80f0

Author: Ron <ron@merehuman.net>

Update Assignments in Report

commit 62732de61d23c605b2742dda99f41529a1cb379b

Author: xolthan <xolthan@gmail.com>

deals with unicode exception and added proximity

commit fc78449ddb2e071a8c4b2a14166fa6e02c5d70d2

Author: xolthan <xolthan@gmail.com>

index.py rewritten for Python 2.7

commit b307fc441c18720e717c93a0f17519f2969d43da

Author: xolthan <xolthan@gmail.com>

Removed decoding in Ascii

commit 0d88284104221772834c420fd77dd9c245fe18e9

Author: xolthan <xolthan@gmail.com>

Add files via upload

commit 6b13d0628fb0d142247cf87bf88cc08f5c4a2cc7

Author: Ron Andrews <35875316+ronandrews@users.noreply.github.com>

Removing Duplicate file

commit a3a4db30ac695daa923e88b95070eeedd522000c

Author: Ron <ron@merehuman.net>

Updated Processing to Match Input format

commit 5c48fb0d503168a8e1677d2788e09b640dcce0f5

Author: Ron Andrews <35875316+ronandrews@users.noreply.github.com>

26 Feb Meeting Update

commit 16b72200d3828d5a791ee3fd31d2fca605a6ae62

Author: xolthan <xolthan@gmail.com>

Add files via upload

commit c8f8d6be309c92574f8fe2d60ed66398be24dc29

Author: xolthan <xolthan@gmail.com>

Add files via upload

commit 60969adee0988b86a7b4168fd6a7f84cfe5018f1

Author: Ron <ron@merehuman.net>

d

commit aa1a3f14db9555106af3fd58fb7dc024b9049b9e

Author: Ron <ron@merehuman.net>

oops

commit 408df3b350596ed86c3ac2bbd808891219001df3

Author: Ron <ron@merehuman.net>

updates

commit 441578f442f55a74ed745b06bb4cfb32193dea70

Author: xolthan <xolthan@gmail.com>

Add files via upload

commit 27e9f904aad15f3daa5c8d5343bd44ad8a256f57

Author: Nidhimidha <36167384+Nidhimidha@users.noreply.github.com>

Update query.py

commit 70dc48c19ebc57532a794520265560864acc02df

Author: xolthan <xolthan@gmail.com>

Add files via upload

commit c0a95246cc14d60d038cabb23433195e1514e623

Author: xolthan <xolthan@gmail.com>

Add files via upload

commit 7cdad81e5c0197e0ea2048008b64d50e02a30679

Author: Nidhimidha <36167384+Nidhimidha@users.noreply.github.com>

Update query.py

commit 6b118fef8efb19fd4e274393ae7f09bcef9dae78

Author: Nidhimidha <36167384+Nidhimidha@users.noreply.github.com>

Update query.py

commit e11905626a621d97c32532b2243d9c220aad0351

Author: Nidhimidha <36167384+Nidhimidha@users.noreply.github.com>

Add files via upload

commit c254401baee72274194b53c7563ea6c658203231

Author: Nidhimidha <36167384+Nidhimidha@users.noreply.github.com>

Update query.py

commit afbe2f079b4fc71d545ca1b98b51e68a7bb8b71d

Author: Nidhimidha <36167384+Nidhimidha@users.noreply.github.com>

Delete FinalReport\_v1.docx

commit 7c9b6e45d96b064a7f162c57fc02d7e762e4a2b5

Author: Nidhimidha <36167384+Nidhimidha@users.noreply.github.com>

Create FinalReport\_v1.docx

commit ab28e88c5757c0cc61c85cc98911ffc1a7b52a33

Author: Nidhimidha <36167384+Nidhimidha@users.noreply.github.com>

Update query.py

commit 1a6744815b08a3d570376f029e9e7d144e28c171

Author: Ron Andrews <35875316+ronandrews@users.noreply.github.com>

Create GroupLog file

commit 48d89594777dcc0a660d01200ecb0ced442a648f

Author: Ron <ron@merehuman.net>

Updated processing output

commit 26c0046b5d285063e789cc7f211b76dc5d005062

Author: Ron <ron@merehuman.net>

Changed to LL

commit ca30d784f2d87a30f2941038f2293eba99894bfb

Author: Ron Andrews <35875316+ronandrews@users.noreply.github.com>

Minor update - some notes

commit 7c6fdd439365195f1430fc904b566836a1e14155

Author: Ron Andrews <35875316+ronandrews@users.noreply.github.com>

Added text from blackboard assignment

commit a67949ab63086b16805aeb97f96d26082b135140

Author: Ron Andrews <35875316+ronandrews@users.noreply.github.com>

oops - useless

commit 3f07b2589313ba430107c2b9124f86968513553e

Author: Ron Andrews <35875316+ronandrews@users.noreply.github.com>

oops - useless....

commit 51c067a63eab2f2b0ee008ff2af360346b2f676f

Author: Ron <ron@merehuman.net>

Added initial CGI script

commit c8b9469b4cc84b07a749ff70bd25fd5c1ea65506

Author: Ron <ron@merehuman.net>

Ported processing to python3

commit 222ff5ed7e47bdda4f1f5fb1c911e0f54c30a8a0

Author: Ron <ron@merehuman.net>

Expanded the ingest files

commit 5b2102a07c6c5730308adf68b816b5574edf2ffa

Author: Ron <ron@merehuman.net>

Adding docsnew.zip

commit e960ae2a0c677cb42444fbde866390a7f1ff7f6e

Author: Ron <ron@merehuman.net>

Processing Update

commit a18ee3be8edb86e7fba1caa9db153bbd602c69e9

Author: Ron <ron@merehuman.net>

Processing Update

commit 4bf8f08391b78c07a08fa14a74b69726c1e950fc

Author: Ron <ron@merehuman.net>

Updates

commit 45af2b3ed7190b2f04d24226fa01e186a13617d9

Author: Ron <ron@merehuman.net>

Initializing CODE folder

commit d4b458a48d076f02d1a88fc9a4a14fc4b8341691

Author: Ron <ron@merehuman.net>

Presentation Material

commit 522ca614cb7719f7b5e06d1333bec9d466f4f151

Author: xolthan <xolthan@gmail.com>

Add files via upload

commit e5cfc6ca08ea434a7f4c4fbf763debd124020fd5

Author: Nidhimidha <36167384+Nidhimidha@users.noreply.github.com>

Update References-Links

commit c928cfed29d90805c1c87c06f553210ab8f6aecc

Author: Ron Andrews <35875316+ronandrews@users.noreply.github.com>

Create References-Links

commit 32692d6249f56fcbbac7581fdb77326d18a74a84

Author: Nidhimidha <36167384+Nidhimidha@users.noreply.github.com>

Create HomePage

commit 8b7eeb65ff87929abb8347b2fd00727698a85bf0

Author: Ron Andrews <35875316+ronandrews@users.noreply.github.com>

Initial commit