

Senior Project File Directory and Info

Kevin VanEmmerik

- **convertPDF/**

- **pdfLibrary/** This folder contains the pdfbox JAR file which is the library used for extracting plain text from the PDF files.
- **pdfProjectData/** This folder contains all of the articles as PDF files.
- **txtFolder/** This folder contains the text files of the articles which are created after running my JAVA program.
- **GET LIST INSTRUCTIONS.txt** Short instructions on how I got a list of the file names.
- **list.txt** List of file names used in my program
- **MainProgram.java** The program I designed in order to convert the PDF files into TXT files. Check the file for more details and documentation.
- **MainProgram.class** The class file from running the java program.

- **SQLInfo/**

- **projectdata.sql** This is the SQL file that will load the database. The csv files in this folder were used to load the data to a WAMPSEVER. From the server I was able to export this SQL code.
- **DocNames.csv, TermIDs.csv, TopicDocDist.csv, TopicsIDs.csv, TopicTermDist.csv**

- **StanfordTopicModel/**

- **lda-0897d0aa-30-d818a922/** This folder contains the output from the Stanford Topic Model Program.
- **script1.scala | script2.scala | script3.scala** These scripts were used with the program to create the outputs.
- **tmt-0.4.0.jar** This is the executable JAR file of the Stanford Topic Model Program.

- **VisualizationProgram/**

- **d3/** - This folder contains the D3 library.
- **dataFiles/**
 - **newFile.csv** - This is a temporary CSV file that is loaded with data from an SQL query
 - **search.php** - This is the PHP file which runs queries in the databases and exports the results to the csv file above.
 - **topTerms.csv** - This file is a CSV file that contains a list of the top 100 keywords from the database. It is used to give suggestions on the search page
- **otherFiles/**
 - **jsFunctions.js** - This is the most important file and contains all the Javascript functions and D3 functions used for displaying the data.
 - **style.css** - This file is a simple CSS file for styling my HTML page.
- **pdfProjectData/** - This folder contains all of the articles as PDF files.
- **index.html** - This is my main HTML page.
- **startPage.pdf** - This is a PDF that is initially loaded into the PDF Viewer.