PDFTracker

# Reference Guide

## Table of Contents

1. Introduction
2. Application Overview
   1. Key Features
   2. System Architecture
   3. Configuration
   4. System Requirements
3. Installation
4. Usage
5. Acknowledgements

  
Data Mining and Machine Learning Laboratory

Dr. Huan Liu

School of Computing, Informatics and Decision Systems Engineering

Arizona State University

Tempe, AZ

E-mail: huan.liu@asu.edu

Shamanth Kumar

School of Computing, Informatics and Decision Systems Engineering

Arizona State University

E-mail: shamanth.kumar@asu.edu

## 1. Introduction

*PDFTracker* helps a user to parse, store, and analyze PDF (Portable Document Format) documents. PDF is a very flexible and convenient format that is widely used for storing documents. PDF Tracker makes searching and analyzing PDF documents easier.

## 2. Application Overview

### 2.1 Key Features

The key features of *PDFTracker* are:

* Summarizes text from the PDF documents.
* Provides an interface to search the text content of PDF documents.
* Realizes convenient side-by-side viewing of text from two PDF documents.
* Uses SQLLite3 as the backend database (zero configuration setup)

### 2.2 System Architecture



### 2.3 Configuration

The tool *pdftracker* uses a configuration file named “pdftracker.conf”. This file contains configuration parameters essential for running the tool. Following is a list of the parameters that must be specified in the configuration file:

#### 2.3.1 workingdir

This parameter specifies the directory used by the tool for storing the error report generated during the parsing of PDF documents.

#### 2.3.2 indexdir

This specifies the location of the index. This index is created using the documents in the database and forms the backbone of the search feature in *pdftracker.*

#### 2.3.3 connectionstring

This parameter specifies the location and the name of the database file that is used to store the converted text files.

The tool also uses a file containing stop words named “stopwords.txt”. This file is essential for building the indices on the database and for the generation of word clouds.

The stop words and configuration file must be placed in the same directory as the tool.

### 2.4 System Requirements

1. JRE 6 or higher.
2. CD-ROM drive.
3. At least 512mb RAM
4. SQLLite3(Microsoft SQL Server 2008 also supported)

## 3. Installation

The demonstration can be run directly from the CDROM. The demonstration uses the configuration file *pdftracker.conf* to identify the working directory. By default, the current working directory is set to  *\pdftrak* which can be modified by copying the files from the CD-ROM onto the hard disk and editing the *workingdir* parameter in the configuration file.

### 3.1 Recommended Usage

For full functionality and unhindered usage of the tool, the application files should be copied to a directory on the hard disk. Please note, that the directory must be write enabled.

## 4. Usage

### 4.1 Parsing PDF documents

To parse a set of PDF documents, place all the documents in a folder. Use the parse window of the tool to select the folder and enter a category to associate with these documents. *PDFTracker* will then parse all the documents and store them in a database file in the working directory. By default, the database is titled “*pdfdocs.db*” and is created in the working directory. If any errors are found during the parsing operation, *PDFTracker* reports the errors by writing them to the file *“error.txt”* in the working directory. Each error message in the file consists of three parts: the filename, location of the file, and the cause of failure during parsing. *PDFTracker* parses PDF documents using the Open Source PDFBox library.

### 4.2 Generating Word Clouds

The comparison window is just a click away from the main window. The comparison window lists all the document categories currently in the database. Select any category to load all the documents for that category. Clicking on a document of interest will display its text in the side-by-side viewer. Once the text has been loaded, the word cloud for the two documents can be generated by clicking on the generate word cloud button. A word cloud is a text visualization technique. The word cloud displays high frequency words to the user. In a word cloud, the font size is proportional to the frequency of the word in the text. The more frequent a particular word, the larger the text.

In the future, word clouds will be generated from bigrams and trigrams to identify domain specific popular keywords.

### 4.3 Searching PDF documents

To search for PDF documents that contain a specific keyword, enter the keyword in the search window. The tool uses Lucene search and indexing library to index the parsed PDF documents and retrieves documents containing the search term.

### 4.4 Summarizing directories

*PDFTracker* is also capable of summarizing directories in the form of word clouds. The top 25 frequent words are also stored in a text file in the working directory. This feature cannot be used while the files are on the CD-ROM. In order to summarize directories, the application must be copied onto a hard drive directory with write permissions.

## 5. Acknowledgements

This work is supported, in part, by the Office of Naval Research.