User Manual
for
PDF-Convert-Search *version 1.0*

CS 208 : Software Engineering

Group 10

Himanshu Dogra

Utkarsh Saxena

Apurv Goyal

Piyush Verma

Rathlavath Santosh

Abstract

The goal of PDF-Convert-Search is to automate the process of conversion of the Patent Data (provided by the Patent Office) from pdf file format to csv (Excel compatible) data format. The patent data is provided in particular format. The output csv data can then be used as it is or can be further fed to the software to execute a complex search query on the data. The output format of the searched data is also csv. The software provides a graphical user interface, so that the user can quickly and easily input a file for conversion and/or searching.

This document is the user manual for PDF-Convert-Search software version 1.0. It is designed to allow users to quickly access information on the various functions of PDF-Convert-Search by the function name. Each function has its own sub-section with information including the purpose of the function, materials needed, preparations for executing the function, inputs to the function, cautions and warnings regarding the function, invocation of the function, how to suspend or quit the function, output from the function, error conditions associated with the function, and any information that is related to the function.

Table of Contents

1. Introduction

1.1 Audience

The intended audience of this User Manual is the user who wishes to use the software for the purpose of conversion of the Patent Data provided by the Patent Office in PDF file format into a more portable/search efficient CSV data format and/or wants to search on the CSV data. This manual describes the functions of the software and helps user in using this software in a quick and effective manner. It also helps user to get familiar with this software and its various functionalities.

1.2 Applicability

The Software is applicable to the patent data provided by the Patent Office in pdf format (Strict Format only).

1.3 Purpose

The main purpose of the software is to convert Patent Data in pdf format into csv format. Search can be performed on this tabular data with using multiple filters and fields. The main pdf input file "XYZ.pdf" is decomposed into three files (if corresponding data is available) namely "**XYZ.in Early Publication.csv**","**XYZ.in Publication After 18 months.csv**","**XYZ.in Granted.csv**". Individual searches can be performed on respective csv files. The required output will be displayed in a separate csv file.

1.4 User Manual Usage

This manual can be used as an installation guide. Also it will help you out in step by step conversion of the input patent file into three csv files. It will guide you how to perform search on the whole data.

1.5 Conventions and Terms

JDK - The Java Development Kit (JDK) is an implementation of either one of the Java SE, Java EE or Java ME platforms released by Oracle Corporation in the form of a binary product aimed at Java developers on Solaris, Linux, Mac OS X or Windows.

Java - Java is a set of several computer software and specifications developed by Sun Microsystems, later acquired by Oracle Corporation, that provides a system for developing application software and deploying it in a cross-platform computing environment.

GitHub - GitHub is a web-based Git repository hosting service, which offers all of the distributed revision control and source code management (SCM) functionality of Git as well as adding its own features.

Oracle Database - Oracle Database (commonly referred to as Oracle RDBMS or simply as Oracle) is an object-relational database management system produced and marketed by Oracle Corporation.

C++ - It is a general-purpose programming language. It has imperative, object-oriented and generic programming features, while also providing the facilities for low-level memory manipulation.

Compiler - A compiler is a computer program (or set of programs) that transforms source code written in a programming language (the source language) into another computer language (the target language, often having a binary form known as object code).

.pdf - pdf stands for Portable Document Format. Each PDF file encapsulates a complete description of a fixed-layout flat document, including the text, fonts, graphics, and other information needed to display it.

.csv - csv stands for Comma (or Character) Separated Values. This format can be read by any spreadsheet program. As it is a plain text file, it can also be read by word processor or simple notepad programs.

.exe - .exe is a common filename extension denoting an executable file (the main execution point of a computer program) for DOS, OpenVMS, Microsoft Windows,Symbian or OS/2.

Software - Computer software or simply software is any set of machine-readable instructions that directs a computer's processor to perform specific operations.

Patent - A patent is a set of exclusive rights granted by a sovereign state to an inventor or assignee for a limited period of time in exchange for detailed public disclosure of an invention.

Adobe Reader - The most commonly used pdf reader.

Conversion - Conversion, here ,refers to the conversion of .pdf file to .csv file

JAR - JAR (Java Archive) is a package file format typically used to aggregate many Java class files and associated metadata and resources (text, images, etc.) into one file to distribute application software or libraries on the Java platform.

Windows Command Prompt - Command Prompt, also known as cmd.exe or just cmd (after its executable file name), is the command-line interpreter on OS/2 and eComStation, Windows CE and Windows NT operating systems. Command Prompt interacts with the user through a command-line interface.

Command Line Interface - A command-line interface or command language interpreter (CLI), also known as command-line user interface, console user interface, and character user interface (CUI), is a means of interacting with a computer program where the user (or client) issues commands to the program in the form of successive lines of text (command lines).

1.6 If You Have A Problem

In case of any queries regarding the product, please mail the Developer Team at: cse130001015@iiti.ac.in . Our team will get back to you ASAP.

Thank You for buying our PRODUCT.

2. Installing PDF-Search-Convert

2.1 Before you begin

Before you begin, you should install Java Developer's Kit (JDK) version and Oracle 11g XE for the use of this software. Also make sure that you have Microsoft Excel (or any other software that supports CSV file format). Availability of a PDF reader is also recommended (Adobe Reader is a good option). Developers additionally need a C++

compiler and development environment for both C++ and Java set up on their computers.

2.2 Unpacking PDF-Search-Convert

The repository for the project is maintained on the GitHub. The link to the same can be found here. Developers can fork the repository and modify it as per their requirements. The users should unpack the zip from this link. Extract the zip file into any folder of your choice.

2.3 Compiling PDF-Search-Convert

The above extracted folder contains all the source files for the project. Developers can compile the file named "**C++ Converter.cpp**" using a C++ compiler. For compiling the modules that run a search query on the CSV data, Java Development Kit (version 7 and above) is required.

3. Using PDF-Search-Convert

3.1 PDF to CSV conversion

Users should use the file named "module.exe" to convert the data from PDF to CSV. The program generates three CSV files as output, for Early Publication Patents, Patents after 18 Months and Patent Granted, separately.

The usage of the programme is as follows (in Windows Operating System):

- Open a **command prompt** window in the **project folder**.
- use the command: java -jar pdfbox-app-1.8.8.jar ExtractText "input_file.pdf" "a.in" (replace input_file with the name of the file to be given as input for the conversion). Make sure the file to be converted **lies in the project folder**.
- The above command produces a file named **"a.in"** in the the project folder itself. Now in the command prompt window, type command C++ Converter.exe a.in
- This creates the three required files "**a.in Early Publication.csv**","**a.in Publication After 18 months.csv**","**a.in Granted.csv**".

3.2 Searching on CSV data

Give the input file path of the csv files. Fill the options of the required filters. Click the search button to generate the required new csv file containing the search results.

4. LIMITATIONS

The PDF-Search- Convert can be used to convert only a specific format namely .pdf format to the .csv format.It is designed to detect only text in the pdf file as .csv file stores tabular data in PLAIN TEXT format. Hence the images remain undetected.The Software can not be used to convert any other format except .pdf. If tried, the Software will terminate with an error message