



UNIVERZITET U NIŠU
ELEKTRONSKI FAKULTET
Katedra za računarstvo



DOCUMENT FINDER

- Kriptografija -

Studenti:

Andrija Petrović

Aleksandar Kostić

Andrej Rakić

Aleksandar Randelović

Jovan Pešić

Katarina Randelović

Đorđe Čikić

Petar Đorđević

Uroš Milivojević

Svetlana Mančić

Mentor:

Prof. dr. Vladan Vučković

Sadržaj

Sadržaj	2
Uvod	3
Opis aplikacije	3
Funkcionalnost	4
Dodatna funkcionalnost	7
Korišćene biblioteke	8

Uvod

Kako Windows operativni sistem nema odgovarajuću podršku za pretraživanje sadržaja fajlova, neophodno je u tu svrhu koristiti softver za pretragu fajlova, čija je funkcija, uglavnom, pronalaženje fajlova koji sadrže zadati termin. Različiti softveri za pretragu fajlova imaju različite ciljeve i skupove funkcija. Aplikacije ovog tipa obično imaju polje za unos teksta, koji se pretražuje, i polje za prikaz rezultata pretrage, a u zavisnosti od aplikacije mogu imati dodatna polja i funkcionalnosti. Dobro poznati primeri softvera za pretragu fajlova uključuju Agent Ransack, FileSeek, Wise JetSearch, Quick Search, Duplicate File Finder, SearchMyFiles, Everything.

U ovom radu biće reči o aplikaciji koja je implementirana za potrebe predmeta Kriptografija na master akademskim studijama Elektronskog fakulteta u Nišu, modula Računarstvo i informatika. Biće diskutovan opis aplikacije i njene funkcionalnosti. Na kraju rada biće navedene korišćene biblioteke.

Opis aplikacije

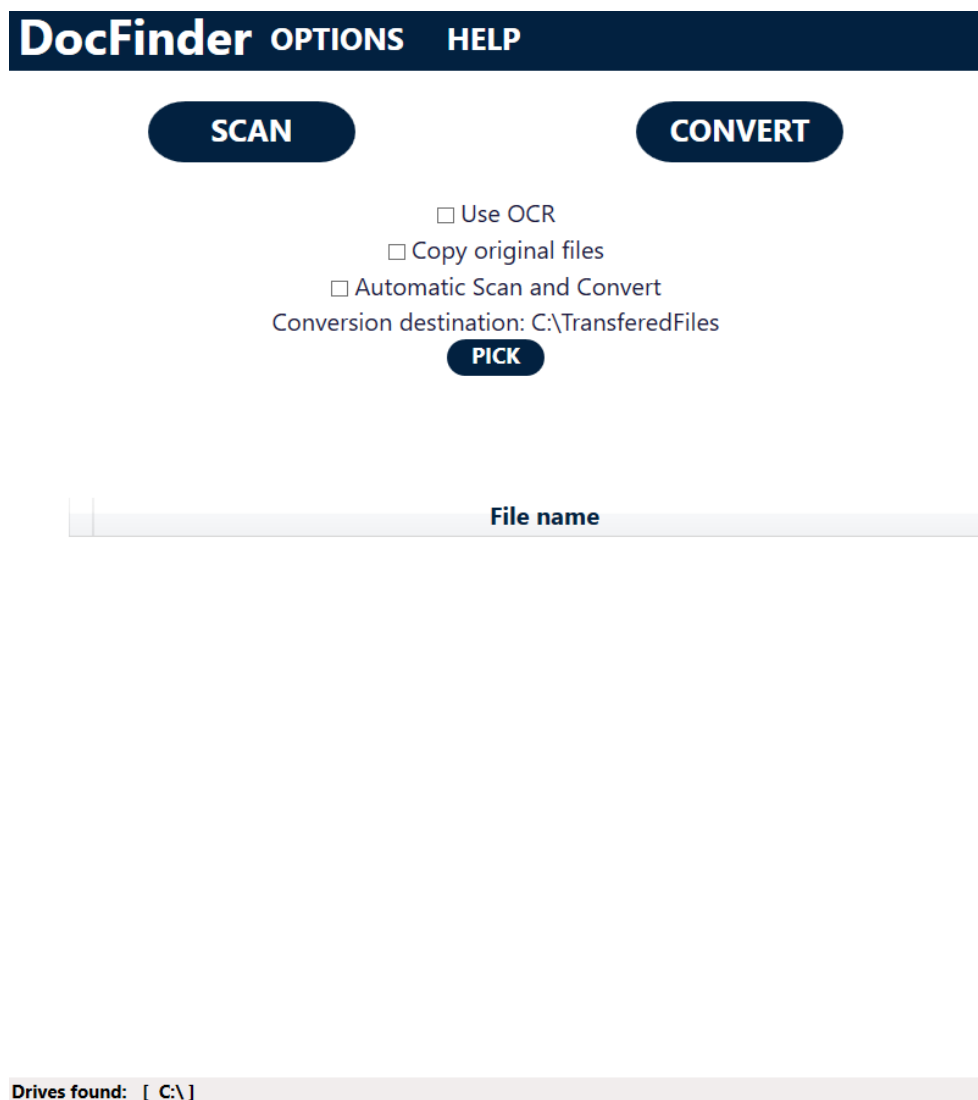
Aplikacija **DocumentFinder** omogućava pronalaženje fajlova koji poseduju sledeće extensione: **.txt .doc .docx .png** i **.pdf**, a zatim sve fajlove koji nisu tekstualnog tipa, konvertuje u **.txt** format. Fajlovi se objedinjuju u jedan direktorijum, nakon čega je moguće vršiti pretragu teksta unutar svih fajlova, po proizvoljno zadatom terminu.

The image shows the initial interface of the DocumentFinder application. It features a dark blue header bar with the application name 'DocFinder' and menu options 'OPTIONS' and 'HELP'. A red 'EXIT' button is located in the top right corner. Below the header, there are two primary action buttons: 'SCAN' and 'CONVERT'. To the right of the 'CONVERT' button is a large, empty rectangular text input field for entering search terms. Under the 'SCAN' button, there are three checkboxes: 'Use OCR', 'Copy original files', and 'Automatic Scan and Convert'. Below these checkboxes, the text 'Conversion destination: C:\TransferredFiles' is displayed, followed by a 'PICK' button. To the right of the checkboxes, there is another checkbox labeled 'Case sensitive' and a 'SEARCH' button. At the bottom of the interface, there is a table with two columns: 'File name' and 'Terms found'. The table is currently empty. At the very bottom, a status bar indicates 'Drives found: [C:\]'.

Slika 1. DocumentFinder početni interfejs

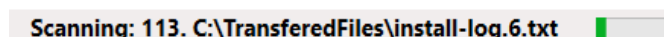
Funkcionalnost

Pri dnu dvodelnog interfejsa se nalazi informaciona traka, na čijoj su levoj strani prikazane sve detektovane particije na sistemu. Novododate particije (*primer USB*), nakon pokretanja aplikacije, bivaju detektovane u fazi skeniranja. Informaciona traka na desnoj strani sadrži informacije o trenutno obrađivanom fajlu, sa indikatorom napretka celokupnog posla. Na levoj strani interfejsa aplikacije mogu se uočiti opcije **SCAN** i **CONVERT**. U fazi skeniranja, sve detektovane particije se pretražuju rekurzivno, pri čemu se pamte izvorne putanje do fajlova, koji poseduju sledeće extenzije: **.txt .doc .docx .pgn i .pdf**.



Slika 2. Levi deo interfejsa

Tekstualni fajlovi automatski bivaju kopirani u direktorijum **C:\TransferredFiles**. Izvorne putanje pronađenih fajlova se upisuju unutar istog direktorijuma, u log fajl **_TransferredFilesPaths.txt**. Treba još napomenuti predefinisane putanje: **C:\Windows**, **C:\Recovery**, **C:\Program Files**, **C:\ProgramData**, **C:\\$Recycle.Bin**, koje se zaobilaze tokom skeniranja. To su putanje koje nazivamo sistemskim, zbog čega na njima ne očekujemo da pronađemo fajlove od esencijalne važnosti za krajnjeg korisnika.



Slika 3. Faza skeniranja

Prilikom konverzije **.doc**, **.docx** i **.pdf** fajlova, kreiraju se tekstualni fajlovi unutar već pomenutog direktorijuma, imenovani identično izvornom fajlu, ali sa **.txt** ekstenzijom i tekstualnom **ASCII** sadržinom.

Converting: 12. Alberto Fernandez Villan - Mastering OpenCV 4 with Python.pdf

Slika 4. Faza konverzije

Na korisniku je da odabere način izvršavanja pomenutih procesa. Naime, moguće je koristiti optičko prepoznavanje znakova, izborom **Use OCR** opcije, prilikom konverzije **.pdf** dokumenata koji nisu u potpunosti pretraživi. Biranjem opcije **Copy original files**, izvorni **.pdf**, **.doc** i **.docx** fajlovi bivaju kopirani u direktorijum, na predefinisanoj putanji **C:\TransferredFiles**, koja je lako izmenljiva, izborom direktorijuma klikom na dugme **PICK**. Objedinjavanje procesa skeniranja i konverzije (*automatizacija*), postiže se izborom **Automatic Scan and Convert** opcije, kao i **Scan and Convert** opcije iz padajućeg menija.

Desni deo interfejsa aplikacije služi za pretragu **.txt** fajlova, dobijenih nakon faze skeniranja i konvertovanja.

The screenshot shows a dark blue header bar with a red 'EXIT' button on the right. Below the header is a large white rectangular input field for text. Underneath the input field is a checkbox labeled 'Case sensitive'. Below the checkbox is a dark blue button with the word 'SEARCH' in white capital letters. Below the 'SEARCH' button is a light gray rectangular box with the text 'Terms found' in bold. At the bottom of the interface is a light gray horizontal bar with a small gray square on the right side.

Slika 5. Desni deo interfejsa

Na vrhu je polje za unos teksta koji je predmet pretrage, a ispod njega opcija **Case sensitive**, koja ukazuje na to da li se veličina unetih karaktera ignoriše ili ne. Pretraga počinje klikom na dugme **SEARCH** i može se stopirati u bilo kom trenutku klikom na dugme **STOP**, koje se nalazi na vrhu.

Na samom početku se proverava da li je obavljeno skeniranje ili konverzijam, kao i da li je unet tekst, jer je to preduslov za pretragu. Ako su navedeni uslovi ispunjeni, pretražuju se svi **.txt** fajlovi, koji se nalaze u direktorijumu **C:\TransferredFiles**. Takođe, proverava se izbor opcije **Case sensitive**. Primer uticaja ove opcije na rezultat pretrage može se uočiti na slikama 6 i 7, gde se vidi da se broj pronađenih fajlova razlikuje.

Finished search for: **Dark**. Found 4 files.

Slika 6. Pretraga "Dark"

Finished search for: **dark**. Found 8 files.

Slika 7. Pretraga "dark"

Provera se vrši za svaki fajl. Ukoliko tekst koji se pretražuje sadrži više reči, za svaku od njih se vrši provera.

Searching: 2107/3352 Snort Intrusion Detection and Prevention Toolkit by Caswell B., Baker A., Beale J. (z-lib.org).txt

Slika 8. Faza pretrage

Lista pronađenih fajlova sortira se najpre po broju pronađenih reči iz unetog teksta, a zatim se rezultati pretrage predstavljaju tabelarno. Na krajnje levoj strani, nalazi se ikonica, koja predstavlja tip fajla. Prilikom pozicioniranja miša, na nekom od rezultata, pojavljuje se nagoveštaj (*hint*), koji sadrži informaciju o originalnoj putanji fajla.

DocFinder

OPTIONS

HELP

EXIT

SCAN

CONVERT

opencv python

☐ Use OCR
☐ Copy original files
☐ Automatic Scan and Convert
 Conversion destination: C:\TransferredFiles

☐ Case sensitive

SEARCH

PICK

File name	Terms found
Alberto Fernandez Villan - Mastering OpenCV 4 with Python	opencv, python
AppCache132931369752459579	opencv, python
AppCache132931454019467170	opencv, python
AppCache132931551745366673	opencv, python
AppCache132931551791819324	opencv, python
Gabriel Garrido, Prateek Joshi - OpenCV 3.x with Python By Example	opencv, python
Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow	opencv, python
Joseph Howse, Joe Minichino - Learning OpenCV 4 Computer Vision	opencv, python
OCR with OpenCV, Tesseract, and Python by Adrian Rosebrock (z-lib.org)	opencv, python
vocab	opencv, python
_58_58_nuget_58_58_	python
appsglobals	python
appssynonyms	python

Drives found: [C:\]

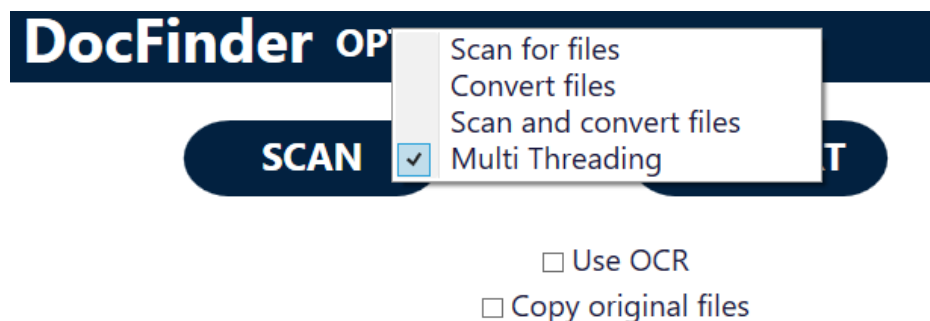
Searching: 310/310 ~\$RD0000.txt

Slika 9. Rezultat pretrage

Dvoklikom na stavku iz liste rezultata, otvara se **.txt** i originalni fajl, ukoliko **.txt** nije format originalnog fajla.

Dodatna funkcionalnost

U standardnom režimu rada, aplikacija koristi isključivo jedno procesorsko jezgro. Izborom **Multi Threading** opcije, iz padajućeg menija pri vrhu, omogućava se funkcionisanje aplikacije na svim dostupnim procesorskim jezgrima.



Slika 10. Multi Threading

U više-procesorskom režimu rada, proces konverzije se znatno brže izvršava. Ubrzanje, posebno dolazi do izražaja pri korišćenju **OCR** funkcije, koja je veoma procesorski intenzivna.

Prethodno opisane procese, skeniranja, konverzije i pretrage, moguće je zaustaviti u bilo kom trenutku izvršenja, jednostavnim klikom na dugme **STOP**. Uspešno odrađeni deo celokupnog posla, do tog trenutka, biva sačuvan.

Od dodatnih funkcionalnosti, treba još napomenuti, mogućnost konverzije i pretrage iz **log** fajla. U slučaju da korisnik, pre procesa skeniranja, odabere proces konverzije ili pretrage, aplikacija će potražiti log fajl, na predefinisanoj putanji. Nakon uspešno učitano log fajla, nastavlja se izvršenje odabranog procesa.

Korišćene biblioteke

1. <https://www.nuget.org/packages/iTextSharp/>
2. <https://www.nuget.org/packages/tesseract/>
3. <https://www.nuget.org/packages/Microsoft.Office.Interop.Word/>