UNIVERZITET U NIŠU

ELEKTRONSKI FAKULTET

Katedra za računarstvo

DOCUMENT FINDER

- Kriptografija -

Studenti:

Andrija Petrović

Aleksandar Kostić

Andrej Rakić

Aleksandar Ranđelović

Jovan Pešić

Katarina Ranđelović

Đorđe Čikić

Petar Đorđević

Uroš Milivojević

Svetlana Mančić

Mentor:

Prof. dr. Vladan Vučković

Niš, mart 2022. godina

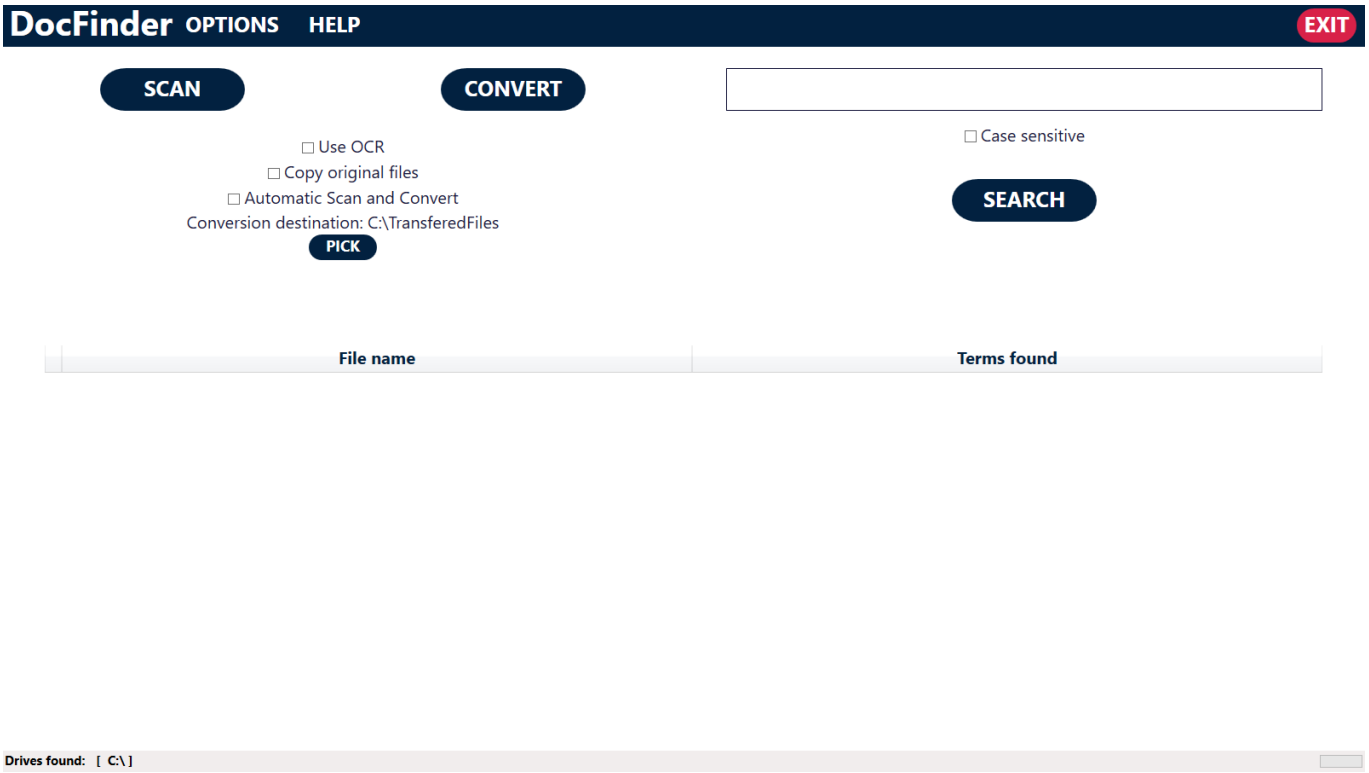
Sadržaj

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 extenzije: **.txt .doc .docx .pgn** 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.

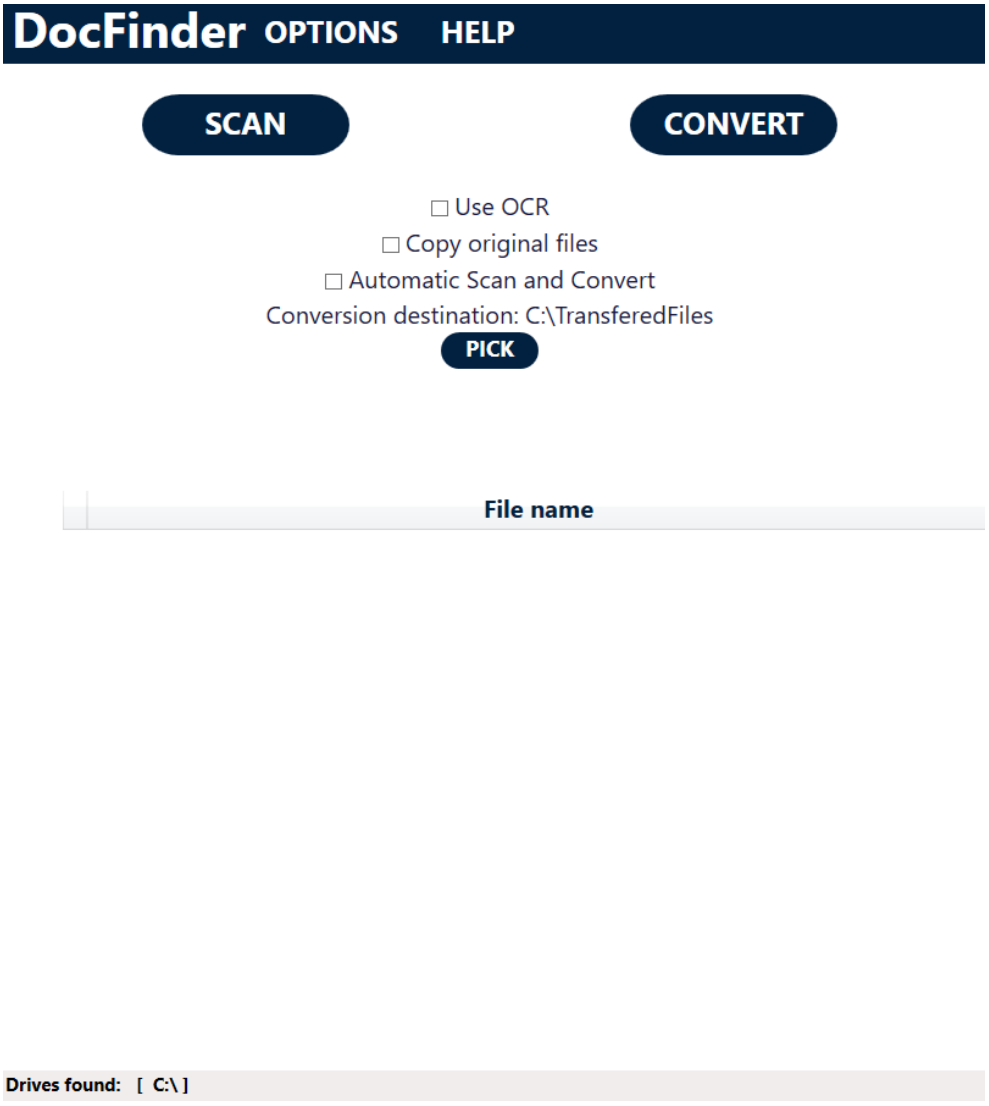


Slika 1. DocumentFinder početni interfejs

Niš, mart 2022. godina

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**.



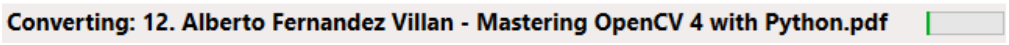
Slika 2. Levi deo interfejsa

Tekstualni fajlovi automatski bivaju kopirani u direktorijum **C:\TransferedFiles**. Izvorne putanje pronađenih fajlova se upisuju unutar istog direktorijuma, u log fajl **_TransferedFilesPaths.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

Niš, mart 2022. godina

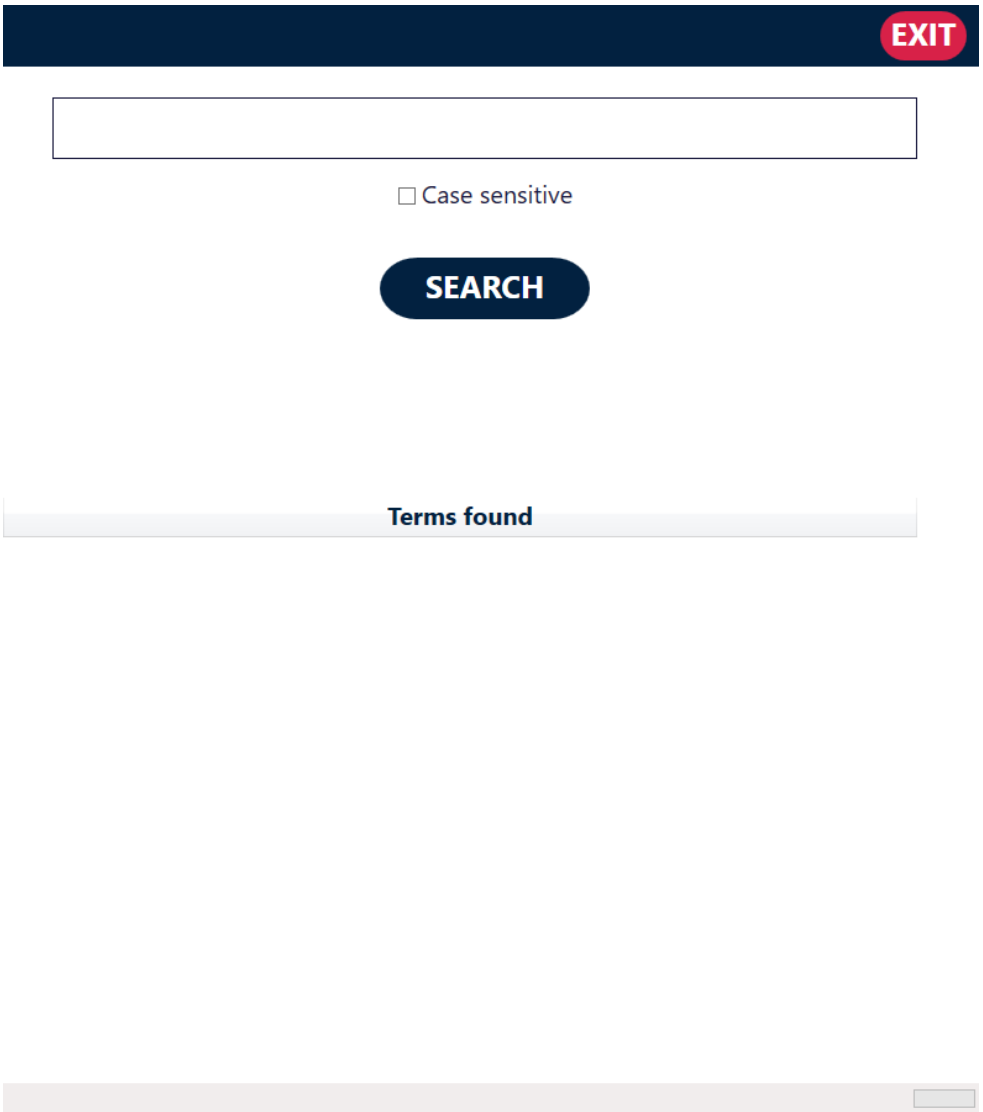
Prilikom konverzije **.doc .docx** i **.pdf** fajlova, kreiraju se tekstualni fajlovi unutar već pomenutog direktorijuma, imenovani identično izvornom fajlu, ali sa **.txt** ekstenzijom i tekstualnom **ASCII** sadržinom.



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 konverije **.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:\TransferedFiles**, 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.



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:\TransferedFiles**. 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.

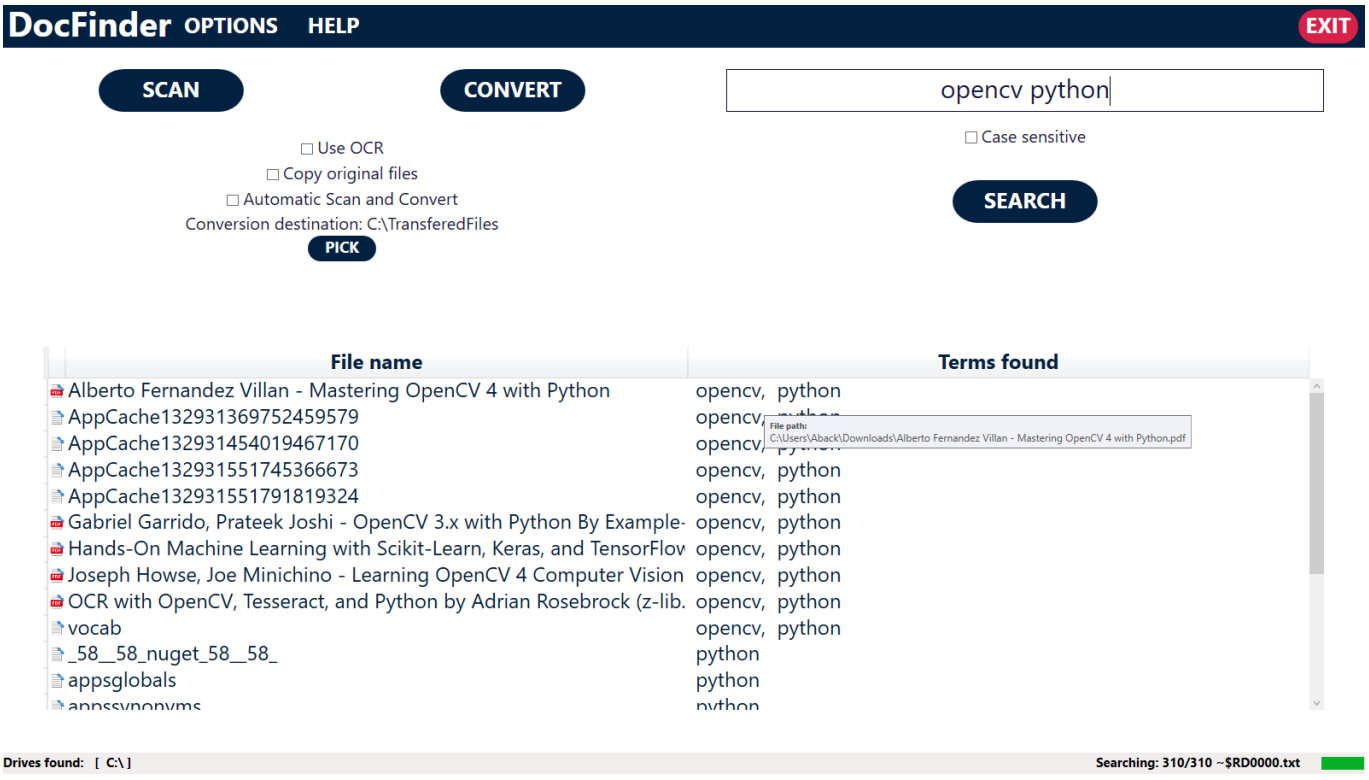
Slika 6. Pretraga "Dark"

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.

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 orignalnoj putanji fajla.

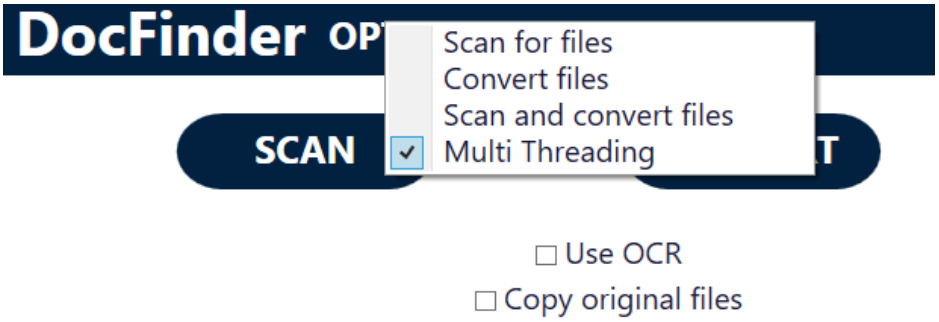


Slika 9. Rezultat pretrage

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 učitanog 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/