Зборник радова Факултета техничких наука, НовиСад

УДК: 4.738
ДОИ: https://doi.org/10.24867/34BE16Malinovic

Развој OCR модула за форензички алат Autopsy

Developing OCR Module for Autopsy Forensics Tool

Немања Малиновић, *Факултет техничких наука, Нови Сад*

Студијски програм – РАЧУНАРСТВО И АУТОМАТИКА

Кратак садржај – *У оквиру овог рада развијен је OCR модул интегрисан у форензички алат Autopsy ради аутоматске екстракције текста из слика током дигиталне форензичке истраге.*
Кључне речи (три до пет): *OCR, индексирање, препроцесирање, модул, Autopsy*

Abstract – *Within this thesis, an OCR module integrated into the forensic tool Autopsy was developed for automatic text extraction from images during digital forensics investigation.*
Keywords: (three to five): *OCR, indexing, preprocessing, module, Autopsy*

НАПОМЕНА: Овај рад проистекао је из мастер рада чији ментор је био др Стеван Гостојић, ред. проф.

1. УВОД

Савремене дигиталне истраге подразумевају анализу великих количина података различитих формата. Један од изазова је идентификација и екстракција текстуалних информација из слика, које нису у машински читљивом облику. У пракси, то отежава претрагу и повезивање релевантних информација, што може довести до споријег и мање ефикасног форензичког процеса. Недостатак модерних уграђених алата за препознавање текста у оквиру Autopsy окружења представља значајан изазов у свакодневном раду форензичара.

У овом раду проблем је решен развојем посебног модула за Autopsy који користи технологију оптичког препознавања карактера (OCR). Као основа примењен је Tesseract OCR, у комбинацији са техникама препроцесирања слике ради побољшања тачности препознавања. Развијени модул омогућава аутоматизовано извлачење текста, његово индексирање и интеграцију са постојећим системом за претрагу (keyword search) и анализу података унутар Autopsy окружења. Мотивација за овакав приступ произилази из потребе да се унапреди ефикасност дигиталних форензичких истрага. Екстракција текста из слика омогућава бржу идентификацију релевантних информација, смањује ризик од превиђања доказа и проширује могућност анализе. На тај начин, овај рад доприноси не само практичној примени у истражним процесима, већ и даљем развоју алата отвореног кода који се користе у дигиталној форензици.

Структура рада организована је на следећи начин. У другом поглављу описан је софтвер Autopsy и могућност његовог проширења кроз Java и Python модуле. Треће поглавље бави се технологијом оптичког препознавања карактера, индексирањем текста и прегледом других форензичких алата који обављају сличне функције. Четврто поглавље садржи спецификацију функционалних и нефункционалних захтева, као и дизајна развијеног софтвера. Пето поглавље обухвата кључне елементе његове имплементације. Шесто поглавље демонстрира примену модула у пракси, анализира резултате и пружа техничке коментаре о раду софтвера. На крају, закључно поглавље даје резиме извршених активности, упоређује предности и мане развијеног решења у односу на претходно описане методе и предлаже могућности за даље унапређење.

2. СТАЊЕ У ОБЛАСТИ

Autopsy [1] је један од најраспрострањенијих алата отвореног кода за дигиталну форензику. Првобитно развијен као графички интерфејс за The Sleuth Kit (TSK) библиотеку, а данас се користи као самосталан систем за анализу дигиталних доказа. Подржава анализу система датотека, преглед метаподатака, екстракцију артефаката из оперативних система, имејл клијената и интернет прегледача. Захваљујући архитектури заснованој на модуларности, Autopsy омогућава проширивање функционалности кроз додатке писане у Java и Python програмским језицима. Врсте модула у Autopsy алату су Ingest, Report и General модули.

2.1. Autopsy – Ingest модули

Ово су најважнији модули јер се извршавају током увоза и анализе података у случај. Постоје две подврсте. Data Source Ingest модули, раде над читавим извором података (нпр. Hash Lookup – упоређује све фајлове са познатим hash сетовима, keyword search – индексира цео извор ради претраге итд). Углавном се покрећу једном по додавању извора података у случај. File ingest модули, раде над појединачним фајловима (нпр. EXIF парсер – извлачи метаподатке из слика итд.) извршавају се појединачно у оквиру процеса учитавања. OCR модул који је тема овог рада је типа

File ingest модул. Анализира сваку слику, врши претпроцесирање и затим екстрахује текст и индексира их у бази података.

2.2. Java и Python модули

Једна од најважнијих предности Autopsy алата у контексту дигиталне форензике је његова висока проширивост. Java модули интегришу нове форензичке анализе директно у платформу, док Python модули омогућавају експерименталну или скриптовану обраду података, што је идеално за тестирање.

Java модули додају се одабиром опције Plugins унутар Autopsy алата где је потребно одабрати и инсталирати .nbm фајл модула који је претходно развијен.

Python модули [2] пружају већу флексибилност и брзину развоја софтвера, што их чини идеалним за истраживаче и академске пројекте. Лако их је написати за аутоматизацију одређених задатака као што су претрага, филтрирање, претварање или нормализација, извлачење метаподатака и интеграција са другим форензичким алатима. Autopsy платформа користи тзв. „Jython“ имплементацију Python програмског језика на Java виртуелној машини. То значи да Python код није изворно извршен од стране класичног Python интерпретера, већ се преводи у Java бајт-код унутар Autopsy платформе. Предности развијања ове врсте модула су брзина развијања и једноставност интеграције у окружење. Развијање је могуће у било ком текст едитору. Мане развијања ове врсте модула су ограничена подршка за Python библиотеке (максимално Python верзија 2.7.18) и потреба за познавањем Autopsy Java API-а.

2.3. Поређење са сличним алатима

У области дигиталне форензике постоји више алата који се користе за анализу података са компјутера, мобилних уређаја и других дигиталних медија. Сваки од ових алата има своје предности и мане, а избор зависи од потреба корисника, буџета и типа истраге.

EnCase је комерцијалан алат који представља индустријски стандард у дигиталној форензици. Омогућава дубинску анализу података, анализу имејлова, као и подршку за оптичко препознавање карактера из слика и докумената. Највећа предност је његова широко прихваћена употреба у судским процесима.

FTK (Forensic Toolkit) је још један комерцијални алат, познат по снажним могућностима за претрагу кључних речи и анализу дигиталних доказа са компјутера, мобилних уређаја и cloud извора. Издваја се по брзини индексирања и кориснички пријатељском интерфејсу, што омогућава ефикасно претраживање великих скупова података.

3. OCR И ИНДЕКСИРАЊЕ ТЕКСТА

3.1. Примена OCR технологија

Оптичко препознавање знакова (енг. optical character recognition – OCR) [3] представља технологију која омогућава претварање текста са слика или скенираних докумената у машински читљив и обрадив формат. Основна идеја јесте да се визуелни приказ знакова, који човек може да прочита, преведе у дигиталну репрезентацију погодну за даљу обраду, претраживање и чување у базама података. Овај процес је од кључне важности у контексту дигитализације докумената, јер омогућава елиминацију ручног прекуцавања и значајно убрзава приступ великој количини информација.

Иако је OCR данас изузетно напредовао, остају бројни изазови. Слаба резолуција, искошени документи, рукопис или текстови на језицима са сложеним писмима и даље представљају препреку. Међутим, трендови показују да интеграција OCR технологије са модерним техникама вештачке интелигенције [4] отварају нове могућности за високо прецизну, брзу и скалабилну обраду текста из визуелних извора.

Основни кораци у OCR процесу су претпроцесирање (нпр. припрема слике да буде погодна за даље анализе применом метода претпроцесирања као што су бинаризација, филтрирање шума, исправљање искошења, нормализација резолуције итд.), сегментација (подразумева раздвајање текста на мање целине, редове, речи и појединачне знакове и представља најизазовнији корак), препознавање знакова (шаблонско препознавање, статички модели, неуронске мреже и технике дубоког учења), постпроцесирање (циљ је исправљање грешака и унапређење квалитета добијеног текста користећи речнике како би се извршило упоређивање препознате речи са базом података познатих речи). Све наведене фазе заједно чине један логички ланац који омогућава да се визуелни садржај претвори у дигитални текст. Уколико једна од фаза не функционише како треба, коначни резултат ће бити значајно мање употребљив. Зато је OCR често посматран као интеграција више дисциплина као што су обрада слике, машинско учење и обраде природног језика.

Примене OCR технологија се проналази у дигитализацији књига и архива где се скениране књиге или новински чланци или историјски документи претварају у дигитални текст који се може претраживати. Претрага текста у административним системима налази примену OCR технологија за аутоматско извлачење релевантних података (нпр. број фактуре, датум, ПИБ компаније). Препознавање регистрационих таблица, банкарство, медицинска документација су још неки од области у којима се OCR користи.

Савремени системи за оптичко препознавање карактера значајно су напредовали последњих деценија и данас представљају један од кључних елемената у процесима дигитализације докумената, обраде текста и аутоматизације административних послова.

Tesseract OCR је један од најпознатијих OCR алата отвореног кода, развијен од стране компаније Google. Овај систем подржава велики број језика, укључујући и ћирилицу, што га чини погодним за примену у различитим културним и језичким контекстима. Једна од највећих предности је могућност тренирања сопствених модела, односно прилагођавање систему специфичним фонтовима или рукописима. Иако је у

прошлости био ограничен по питању тачности, интеграција са техникама дубоког учења и употребом вештачке интелигенције донела је значајно побољшање у прецизности препознавања знакова и карактера. Предност је што је алат бесплатан и отвореног кода.
Handwriting OCR (ICR – интелигентно препознавање карактера) је алат који је високо специјализован у препознавању рукописа. За разлику од штампаног текста, рукописи се карактеришу великом варијабилношћу у облику слова, неуједначеном величином и нагибом, као и честим спајањем карактера. Савремене технике дубоког учења и вештачке интелигенције омогућавају овом алату да буде међу бољима у области препознавања текста писаног рукописа. Мана је скупа лиценца.
Cloud OCR решења последњих година нуде такође висок ниво квалитета услуга. Најпознатији представници су Google Vision API, Microsoft Azure OCR и Amazon Textract. Ове услуге заснивају се на инфраструктури великих cloud провајдера, што им омогућава високу скалабилност, константно унапређивање модела и интеграцију са другим сервисима који користе вештачку интелигенцију. Ова решења елиминишу потребу за локалним хардверским ресурсима, што их чини погодним за организације које желе брзо и флексибилно увођење OCR-а без значајних почетних инвестиција.

3.2. Индексирање у форензичким алатима

Модерни форензички софтвери препознају да велики део доказа није у класичном текстуалном облику, већ унутар слика, PDF докумената и скенираних прилога. Управо због тога су у своје оквире интегрисали модуле за оптичко препознавање карактера, као и механизме за индексирање текста, чиме истражитељима и корисницима омогућавају брзо претраживање и анализу.
Инвертовани индекс је најчешће коришћена техника за индексирање и представља процес креирања помоћних структура података које омогућавају да се одређена реч или фраза пронађе унутар великог скупа докумената у врло кратком року.
Индексирање обухвата кораке као што су токенизација (раздвајање текста на токене или речи), нормализација (претварање свих речи у мали регистар), стеминг и лематизација (свођење речи на њихов корен или инфинитив), уклањање стоп речи (елиминација честих и мање значајних речи), генерисање мапе (креирање инвертованог индекса у форми: реч – [документ1,позиција1]).
Autopsy алат користи Apache Solr алат за претрагу који се надограђује на алат Lucene за индексирање и претрагу. Тиме је омогућена претрага по тексту који је претходно извучен из датотека.

4. СПЕЦИФИКАЦИЈА

4.1. Функционални захтеви

Функционални захтеви дефинишу шта систем треба да ради. Развијени модул омогућава корисницима одабир конфигурације модула (постављање параметара за подржане формате слика, језик и опције претпроцесирања), покретање OCR-а над сликама (извлачење текста из подржаних фајлова и складиштење резултата), преглед OCR логова (праћење статуса обраде и грешака приликом извршења модула) и претрагу извученог текста (брзо пронаћи конкретан текст или кључне речи).

4.2. Нефункционални захтеви

У контексту развијеног модула, ови захтеви се односе на перформансе, стабилност, употребљивост и интеграцију у постојећи софтверски оквир. Од модула се очекује да може ефикасно да обради велики број датотека у разумном времену, као и да пружи брзе и тачне резултате претраге над индексираним текстом. Интерфејс мора бити довољно једноставан да га форензичари могу користити без потребе за додатним техничким знањем, док интеграција у саму платформу мора бити без нарушавања других функционалности. Безбедност података је важна. Сав извучени текст мора се чувати унутар Autopsy база података како би се очувао интегритет доказа. Модул треба да буде развијен тако да омогући лако одржавање и проширивање, како би се накнадно могли додати нови алгоритми или технике без већих измена постојеће структуре.

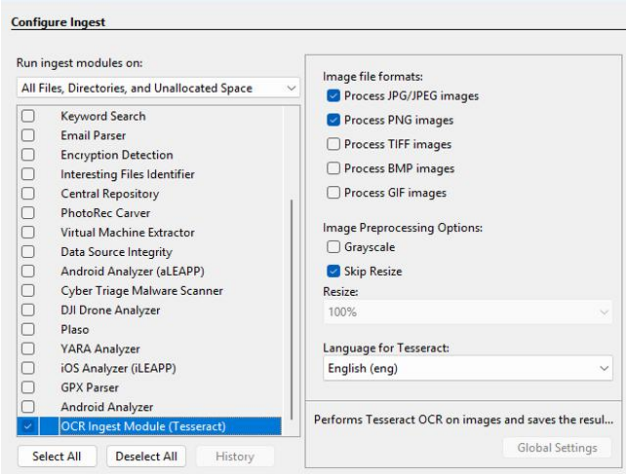
5. ИМПЛЕМЕНТАЦИЈА

Основна структура модула обухвата тзв. фабрику (енг. factory pattern) која је одговорна за креирање инстанце главне класе модула и његово повезивање са Autopsy платформом. Платформа захтева овај образац ради регистрације сваког новог модула. Главна класа модула садржи функционалности чувања подешавања, иницијализацију контекста и одређивање типова фајлова које модул обрађује на основу корисничког избора. Метода startup се позива пре обраде било ког фајла и служи за филтрирање по типу слике. Подржани формати слике су JPEG/JPG, PNG, TIFF, BMP и GIF.
ImageMagick алат отвореног кода се користи за технике претпроцесирања пре OCR процеса. Подржане технике су претварање слике у црно-белу слику (енг. grayscale) ради изоштравања читљивости текста и опција промене резолуције слике (енг. resize). Након обраде слика, модул покреће Tesseract OCR као спољашњи процес на процесираној слици. Излазни ток процеса садржи текст са слике. Уколико извршење није успешно, програм евидентира упозорење у логовима, али наставља са обрадом других фајлова, како би се обезбедила робусност целог модула. Након успешног извршења, излазни текст се декодира у стринг и резултат се анализира. Ако је пронађен текст, креира се артефакт који ће бити сачуван у Autopsy бази података ради касније индексирања.
Структура кода модула је организована у складу са принципима модуларности. Фабрика је одговорна за креирање инстанци модула и његово повезивање са Autopsy окружењем, главна класа управља логиком обраде, док је панел за подешавања издвојен као

засебна компонента која омогућава кориснику да конфигурише начин рада.

6. ДЕМОНСТРАЦИЈА

У овом одељку је кроз пример приказана употреба модула. На слици 1 приказан је изглед корисничког интерфејса који представља подешавања за модул.

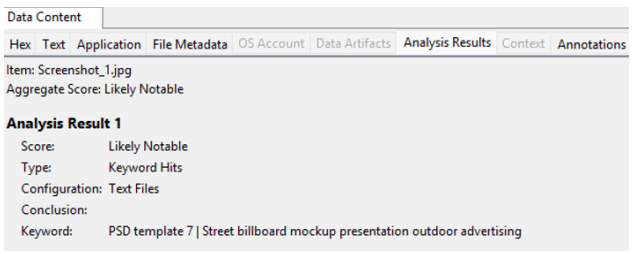


Слика 1. Кориснички интерфејс

Корисници бирају формат слика, опције претпроцесирања као и језик у којем је текст на слици описан. На слици 2 налази се слика жутог билборда са црним текстом на сивој металној конструкцији, са замућеном позадином. Из слике 2 ће бити извучен текст.

Слика 2. Слика из које се извлачи текст

Корисничка подешавања модула у овом примеру су одабир формата JPG и PNG, за језик је одабран енглески и коришћена је операција претварање слике у црно-белу. На слици 3 приказан је резултат извлачења текста.



Слика 3. Резултат анализе – Жути билборд

У овом случају извлачење текста је изузетно тачно. Сви читљиви текстуални елементи, осим стрелице (која није текст) и броја „7“ (који није на слици), су правилно препознати и спојени у један низ. Осим тих грешака, текст је изузетно тачно препознат и спојен, при чему су све речи правилно извучене упркос различитим фонтовима и позицијама.

7. ЗАКЉУЧАК

У овом раду анализирана је потреба за аутоматизованом обрадом и препознавањем текста из дигиталних слика у контексту форензичких истрага. Специфицирано решење и реализовани модул, доносе значајне предности у односу на слична постојећа решења. Пре свега, интеграција у Autopsy платформу омогућава да корисник из једног окружења управља целокупним процесом обраде и анализе слика, без потребе за додатним алатима. Ипак, имплементирани модул има и своја ограничења. OCR резултати могу бити непоуздани у случајевима када су слике ниске резолуције, имају сложену позадину или текст не одговара изабраном језику. Алат такође не може сам да изабере параметре за претходну обраду. Обрада великих или бројних слика захтева значајне хардверске ресурсе, посебно када су укључене операције претпроцесирања. Ово представља простор за побољшање у наредним верзијама. Ове надоградње учиниле би модул још кориснијим у форензичким истрагама и истраживању дигиталних доказа.

8. ЛИТЕРАТУРА

[1] Autopsy User Documentation [Online]. Доступно: https://sleuthkit.org/autopsy/docs/user-docs/4.18.0/ (приступљено септембар 2025.)

[2] Python module development [Online]. Доступно: https://sleuthkit.org/autopsy/docs/api-docs/4.19.3/mod_dev_page.html (приступљено септембар 2025.)

[3] OCR and Indexing [Online]. Доступно: https://support.filevine.com/hc/en-us/articles/360034968272-OCR-and-Indexing (приступљено септембар 2025.)

[4] C.M. Bishop, Pattern recognition and machine learning. New York, NY, USA: Springer, 2006.

Кратка биографија:

Немања Малиновић рођен је у Новом Саду 2000. год. Мастер рад на Факултету техничких наука из области Електротехнике и рачунарства – дигитална форензика одбранио је 2025.год.
Контакт: malinovicnemanja6@gmail.com