Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.9
DOI: https://doi.org/10.24867/24BE14Knezevic

PREPOZNAVANJE TEKSTA POMOĆU KOGNITIVNIH SERVISA U AMAZON VEB SERVISIMA

TEXT RECOGNITION USING COGNITIVE SERVICES IN AMAZON WEB SERVICES

Mia Knežević, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIKA I RAČUNARSTVO

Kratak sadržaj – *Ovaj rad se zasniva na istraživanju upotrebe mašinskog učenja za prepoznavanje i ekstrakciju teksta i podataka iz setova dokumenata u različitim formatima, koji sadrže slike tekstova napisanih rukom ili na različitim jezicima, štampanim slovima u digitalnom formatu. Istraživanje je fokusirano na upotrebu Amazon Textract-a, kognitivnog servisa koji pružaju Amazon veb servisi, za automatsku obradu dokumenata. Cilj istraživanja jeste da se upotrebom različitih ulaza prikupe rezultati koji će se iskoristiti za testiranje i analiziranje tačnosti prepoznavanja teksta i performansi obrade dokumenata od strane Amazon Textract servisa, kao i prilagodljivosti na različite ulazne dokumente i kvalitet ulaznih dokumenata. Za tehničku implementaciju zadatka korišćeni su sledeći Amazonovi servisi: Simple Storage Service. Lambda funkcija i Textract servis.*

Ključne reči: *Amazon veb servisi, kognitivni servisi, mašinsko učenje, Textract, Lambda funkcija*

Abstract – *This paper is based on an exploration of the use of machine learning for recognition and extraction of text and data from sets of documents in different formats, which contain images of texts written by hand or in different languages, printed letters in digital format. The research is focused on the use of Amazon Textract, a cognitive service provided by Amazon Web Services, for automatic document processing. The goal of the research is to use different inputs to collect results that will be used to test and analyze the accuracy of text recognition and document processing performance by the Amazon Textract service, as well as adaptability to different input documents and the quality of input documents. The following Amazon services were used for the technical implementation of the task: Simple Storage Service. Lambda function and Textract service.*

Keywords: *Amazon Web Services, cognitive services, machine learning, Textract, Lambda function*

1. UVOD

Poslednjih nekoliko decenija, sve više računara se oslanja na mnoštvo tehnika veštačke inteligencije i njenu primenu u ljudskom životu. Replikacija ljudskih aktivnosti i funkcija je sve učestalija i pruža mnoštvo benefita u savremenom svetu. Sve više se susrećemo sa sistemima koji za svoje upravljanje koriste veštine i znanje čoveka. Prepoznavanje optičkih znakova postalo je jedna od najuspešnijih primena tehnologije u polju veštačke inteligencije i prepoznavanja uzoraka, a koja se koristi za konvertovanje rukom napisanih znakova, slova ili reči u digitalni format. Primena ove tehnologije omogućava automatizaciju procesa izdvajanja podataka iz dokumenata i fotografija čak i ako su oni napisani rukom ili su u nečitkom obliku, što može da uštedi vreme i smanji greške u odnosu na ručno izdvajanje podataka. Osim toga, omogućeno je da se izdvojene informacije elektronski uređuju, pretražuju i skladište kompaktnije i efikasnije.

Ovaj rad se zasniva na istraživanju upotrebe mašinskog učenja za prepoznavanje i ekstrakciju teksta i podataka iz setova dokumenata u Joint Photographic Experts Group (JPEG) formatu i Portable Document Format (PDF) formatu, koji sadrže slike tekstova napisanih rukom ili na različitim jezicima, štampanim slovima u digitalnom formatu. Istraživanje je fokusirano na upotrebu Amazon Textract-a, kognitivnog servisa koji pruža Amazon Web Services (AWS), za automatsku obradu dokumenata.

Cilj istraživanja jeste da se upotrebom različitih ulaza prikupe rezultati koji će se iskoristiti za testiranje i analiziranje tačnosti prepoznavanja teksta i performansi obrade dokumenata od strane Amazon Textract servisa, kao i prilagodljivosti na različite ulazne dokumente i kvalitet ulaznih dokumenata.

2. OPIS PROJEKTNOG ZADATKA

Cilj ovog projekta je ispitivanje sposobnosti i performansi Textract servisa u automatskom prepoznavanju teksta na osnovu ulaznih slika u JPEG ili PDF formatu. Kako bi se dobili što precizniji rezultati, testiranje će obuhvatiti četiri različita skupa podataka sa različitim izazovima.

Prvi skup podataka sadrži slike ručno pisanog teksta na engleskom jeziku, drugi skup čine slike ručno pisanog teksta na francuskom jeziku, treći skup su uslikane fotografije računa i četvrti, poslednji skup, čine fotografije skeniranih računa. Svaki od ovih skupova imaju različite izazove sa kojima se servis za obradu susreće, a od tih izazova zavise i rezultati performansi datog servisa.

Rezultati testiranja i njihova analiza će dati jasnu sliku o performansama Textract servisa i njegovoj primenljivosti u realnom svetu. Da bi se došlo do rezultata, neophodno je pripremiti skupove podataka, postaviti arhitekturu i iskonfigurisati servise da adekvatno obrade ulazne podatke i zapišu dobijene rezultate u odgovarajući format.

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bio dr Srđan Vukmirović, red. prof.

3. TEHNOLOGIJE

3.1 Prepoznavanje teksta pomoću kognitivnih servisa u amazon veb servisima

AWS kognitivni servisi [2] su set alata i tehnologija koji se koriste za razvoj aplikacija sa sposobnošću da razumeju, razgovaraju, pročitaju i interpretiraju prirodni jezik. Ovi servisi koriste mašinsko učenje i druge kognitivne tehnologije za obradu i analizu podataka, u svrhu pružanja mogućnosti za razvoj aplikacija koje mogu da razumeju tekst i govor, prepoznaju emocije i ton u tekstu, sinhronizuju se sa chatbot-ovima i drugim vrstama interakcija sa korisnicima, kao i da prepoznaju slike i objekte na njima.

Amazon Textract [3] je servis za automatsko prepoznavanje teksta. Koristi mašinsko učenje da prepozna tekst, tabele, formulare i slične objekte u dokumentima. Ovaj servis omogućava korisnicima da jednostavno ekstraktuju i obrađuju podatke iz različitih tipova dokumenata. Takođe može da prepozna tabele, koristi automatsku klasifikaciju za prepoznavanje tipova dokumenata, takođe i ima funkciju za obradu više jezika.

Lambda funkcija [4] je platforma za izvršavanje koda bez administracije servera. Koristi se za izvršavanje koda u odgovoru na događaje. Korišćenjem Lambda funkcija, korisnici mogu da kreiraju skalabilne aplikacije, a da pritom ne moraju da brinu o upravljanju serverima i infrastrukturom. Potpuno su automatizovane i skaliraju se u zavisnosti od potreba aplikacije. Funkcije mogu da se pišu u više programskih jezika, uključujući NodeJS Javascript, Java, C#, Python.

Simple Storage Service [5] je Amazonova usluga skladištenja objekata, koja nudi vodeću skalabilnost, dostupnost podataka, sigurnost i performanse. Servis omogućava skladištenje i razmenu podataka u oblaku. S3 se koristi za skladištenje datoteka i objekata, kao i za arhiviranje, sigurnosne kopije i druge vrste podrške za aplikacije.

Python [6] je programski jezik koji se koristi za razvoj aplikacija i automatizaciju procesa. To je jednostavan i intuitivan jezik, koji je dobro poznat po svom dinamičnom tipiranju i širokom spektru biblioteka i alata za različite aplikacije, od web razvoja do naučne analize podataka. Ima jasnu sintaksu i jedan je od najpopularnijih programskih jezika u svetu i daje jedinstvenu kombinaciju jednostavnosti, snage i prilagodljivosti.

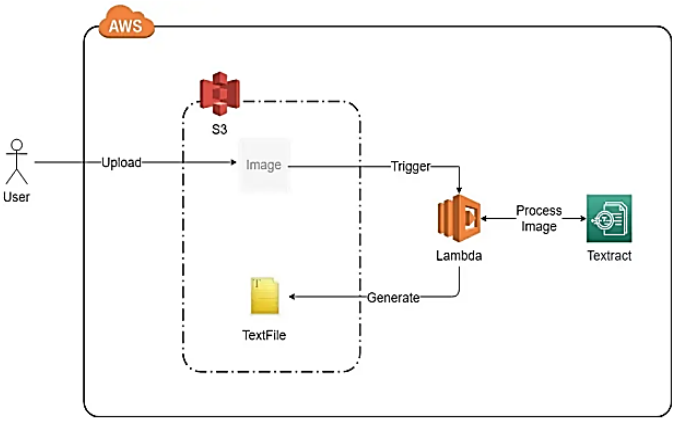
4. TEHNIČKI OPIS REŠENJA

Za tehničku implementaciju zadatka korišćeni su sledeći Amazon Web servisi:

- Simple Storage Service - S3,
- Lambda funkcija i
- Textract servis.

Slika 1 predstavlja prikaz arhitekture projekta [7]. Nju čine prethodno izlistani Amazon servisi.

Na slici 1 takođe može da se vidi proces rada sistema. Potrebno je da korisnik ubaci set ulaznih podataka u S3 skladište, što će aktivirati Lambda funkciju. Lambda funkcija preuzima podatke iz S3 skladišta i prosleđuje ih Textract servisu za obradu. Textract servis obrađuje sliku i ekstraktuje tekst iz nje. Lambda funkcija izrašunava vreme koje je proteklo za obradu slike koristeći in'build funkcije u Python-u. Vreme koje je proteklo za obradu slike je snimljeno u tekstualni fajl, koji se čuva u S3 skladištu.



Slika 1: *Prikaz arhitekture projekta*

Proizvod navedenog procesa jeste set tekstualnih fajlova, koji su korespodentni ulaznom setu slika u JPEG formatu. U svakom od tih fajlova nalazi se vreme koje je proteklo prilikom obrade odgovarajuće ulazne slike. Da bi se stekao realan uvid u performanse servisa i došlo do rezultata, neophodno je agregirati pojedinačne rezultate merenja i pronaći njihovu srednju vrednost. Srednje vrednosti svakog ulaznog skupa podataka će biti rezultati kompletnog procesa.

Za izračunavanje srednje vrednosti rezultata koristi se još jedna Lambda funkcija, koja prikuplja numeričke vrednosti za svaku obrađenu sliku, sabira ih, a potom deli sa ukupnim brojem obrađenih slika unutar jednog ulaznog seta. Dobijena vrednost se snima u tekstualni fajl, čuva u S3 kanti i kasnije koristi za analizu performansi Textract servisa.

5. METODOLOGIJA

5.1. Skupovi podataka

U svrhe testiranja performansi Textract servisa, korišćena su četiri skupa podataka:

- IAM Handwriting Database - skup ručno pisanih tekstova na engleskom jeziku,
- A Handwritten French Dataset for Word Spotting – CFRAMUZ - skup ručno pisanih tekstova na francuskom jeziku,
- My receipts (pdf scans) - skup skeniranih računa i
- Dataset of invoices and receipts including annotation of relevant fields - skup računa slikanih fotoaparatom.

"IAM Handwriting Database" [8] je skup biblioteka slika i tekstova koji su napisani rukom. Sadrži rukopise koje su napisali različiti autori u različitim stilovima pisanja, što čini ovaj skup podataka idealnim za testiranje i treniranje modela prepoznavanja rukopisa. Smatra se kvalitetnim data setom za istraživanje u oblasti prepoznavanja rukopisa, sa visokom kvalitetom slika i tekstova.

"A Handwritten French Dataset for Word Spotting – CFRAMUZ" [9] je prvi istorijski skup podataka na francuskom jeziku, koji se često koristi za pretragu reči bez segmentacije, predstavljen na radionici Historical

Image Processing 2017 u Kjotu. Skup podataka se sastoji od sedam romana sa annotacijom od 18,000 reči, na francuskom jeziku, od strane slavnog pisca iz Lozane, Charles-Ferdinand Ramuza.

“My receipts (pdf scans)” [10] predstavlja ličnu kolekciju skeniranih računa iz nekoliko različitih država sveta, koju je sakupio Jens Walter. Računi su skenirani i sačuvani u PDF formatu. U svrhe ovog projekta korišćen je podskup od 100 ulaznih fajlova, kog sačinjavaju računi iz Hrvatske, Nemačke i Sjedinjenih Američkih Država. Podaci iz ovog skupa podataka su javni i predstavljaju ličnu kolekciju skupljenu od strane Jensa Waltera.

“Dataset of invoices and receipts including annotation of relevant fields”[11] je skup koji sadrži račune i fakture, uključujući annotaciju relevantnih polja. Obuhvata 813 slika računa i potvrda privatne kompanije na portugalskom jeziku.

5.2. Integracija programskog rešenja

Za realizaciju programskog rešenja napisane su dve Lambda funkcije *getTextFromS3Documents* i *aggregateElapsedTime,* korišćenjem Python programskog jezika.

Prva lambda funkcija koja se izvršava jeste *getTextFromS3Documents*. Funkcija koristi modul boto3 za rad sa Amazon Textract servisom i S3 uslugom. Glavna metoda "lambda_handler" se poziva kada se dogodi neki događaj - konkretno, kada se JPEG ili PDF dokument otpremi u S3 skladište. Metoda najpre dobavlja dokument iz S3 skladišta i prosleđuje ga metodi get_textract_data. Ukoliko dođe do greške tokom obrade, ona će prijaviti grešku.

Metoda get_textract_data koristi Amazon Textract API da izdvoji tekstualne informacije iz dokumenta koji se nalazi u određenoj AWS S3 kanti. Metoda takođe meri vreme koje je potrebno da bi se dobili podaci iz dokumenta, počevši od trenutka kada se metoda pozove (start_time) i završava kada se odgovor vrati iz Textract servisa (end_time). Na kraju, metoda vraća start_time i end_time kao rezultat.

Metoda write_elapsed_time meri proteklo vreme između start_time i end_time vremena i zapisuje ga u tekstualnu datoteku koja je sačuvana na AWS S3 skladište. Ime datoteke se dobija iz created_s3_document argumenta tako što se ekstenzija dokumenta zameni sa ".elapsed_time.txt" ekstenzijom. Metoda koristi s3.put_object metodu iz boto3 modula kako bi se napisao tekstualni sadržaj (elapsed_time) u izgenerisanu datoteku.

Druga Lambda funkcija koju je potrebno izvršiti da bi se dobili konačni rezultati merenja jeste aggregateElapsedTime. Ova funkcija je handler funkcija koja se koristi u AWS Lambda servisu za obradu događaja koji se pokreću kada se dokumenti sa tekstualnim sadržajem skladište u AWS S3 skladište. Funkcija prvo izvlači ime skladišta iz kojeg će čitati dokumente. Zatim, uzima direktorijum iz event argumenta i koristi taj direktorijum da bi pronašla sve dokumente u skladištu unutar tog direktorijuma. To se radi pozivanjem metode s3.list_objects iz boto3 modula, koja vraća listu svih objekata (dokumenata) iz skladišta, sa zadatim prefiksom (direktorijumom). Funkcija zatim prolazi kroz sve dokumente koji su pronađeni i proverava da li se u imenu dokumenta nalazi reč "elapsed_time". Ako se ta reč nalazi u imenu, uzima sadržaj datoteke (koji sadrži vreme obrade dokumenta), a to radi pozivanjem metode s3.get_object i dodaje ga u listu elapsed_times. Nakon što je vreme obrade svih dokumenta u listi, funkcija računa prosečno vreme obrade tako što uzima sumu vremena u listi i deli ga sa brojem elemenata u listi. Rezultujuće prosečno vreme obrade se zapisuje u tekstualnu datoteku koja se sačuva u skladište na AWS S3 servisu.

Naziv datoteke dobija se kombinovanjem imena direktorijuma i "average_elapsed_time.txt" ekstenzijom. Na kraju, funkcija vraća status kod 200 i event argument u obliku odgovora (response body).

6. REZULTATI I DISKUSIJA

Nakon što su prethodno opisane dve lambda funkcije izvršene nad četiri seta podataka koji su korišćeni u svrhe testiranja performanski Textract servisa, dobijena su četiri tekstualna dokumenta u kojima su zapisani rezultati merenja.

Krajnji rezultat izvršavanja dveju Lambdi jeste prosečno vreme izvršavanja koje je izračunato i sačuvano u tektualni dokument pod nazivom average_elapsed_time. Dakle, za četiri skupa podataka dobijeno je četiri rezultata. U sledećoj listi su navedene vrednosti prosečnog vremena izvršavanja za obradu skupova podataka IAM Handwriting Database, A Handwritten French Dataset for Word Spotting – CFRAMUZ, My receipts (pdf scans) i Dataset of invoices and receipts including annotation of relevant fields, respektivno, u sekundama.

- 2.29,
- 2.47,
- 1.78,
- 2.24.

Iz priloženih rezultata, može se doći do sledećih zaključaka - najveće prosečno vreme za obradu podataka je potrošeno je na obradu skupa podataka CFRAMUZ, a najkraće je potrošeno na My receipts (pdf scans). Uzimajući u obzir da je kvalitet slika iz CFRAMUZ skupa podataka bio znatno lošiji i sa mnogo više šuma i zatamljenosti u odnosu na slike iz drugih skupova podataka, način na koji je napisan, pisana slova na francuskom jeziku, kao i sve izazove koje potencijalno zadaje Textract servisu pri obradi podataka, očekivano je da će performanse prilikom obrade biti najsporije za ovaj skup podataka. Sa druge strane, My receipts (pdf scans), skup podataka sa nakraćim prosečnim vremenom obrade, sadrži skenirane račune sa štampanim tekstom ili tabelarnim prikazom, na različitim jezicima. Uzimajući u obzir da su računi skenirani, šum na dokumentima je minimalan, a izazovi za Textract znatno manji. Ove činjenice objašnjavaju zbog čega je najmanje vremena utrošeno na obradu ovog skupa podataka.

Zanimljivo je i napraviti paralelu između obrade štampanih i ručno pisanih rečenica u podacima. Iz dobijenih rezultata, zaključuje se da su performanse Textract servisa znatno bolje pri obradi štampanih, nego ručno pisanih podataka.

Iz priloženih rezultata, takođe, vidimo da je skup podataka sa ručno pisanim tekstom na francuskom jeziku iziskivao više vremena za obradu od skupa sa ručno pisanim tekstom na engleskom jeziku. Ali, kvalitet slike je bio znatno bolji u

engleskom skupu podataka i on nije sadržao dodatne izazove poput pocepanog papira, mutnih fotografija, mrlja na papiru i slično.

Ukoliko se pravi paralela između rezultata skupa podataka sa skeniranim računima i skupa podataka sa kamerom uslikanim računima, zaključuje se da je su uslikani računi zahtevali više vremena za obradu. Samim tim što su slikani kamerom, u tom skupu se nalazi više mutnih fotografija sa slabo čitljivim tekstom, slabijom osvetljenošću i drugim izazovima sa Textract servis.

7. ZAKLJUČAK

Na osnovu dobijenih rezultata izvršavanja Textract servisa na četiri različita skupa podataka, može se zaključiti da performanse ovog servisa značajno variraju u zavisnosti od karakteristika podataka koji se obrađuju. Rezultati pokazuju da je prosečno vreme obrade bilo najduže za skup podataka CFRAMUZ, koji je sadržao lošije kvalitete slika, veći šum i druge izazove koji su negativno uticali na performanse Textract servisa. Sa druge strane, skup podataka My receipts (pdf scans) sa najkraćim prosečnim vremenom obrade, sadržao je skenirane račune sa minimalnim šumom, štampanim tekstom i tabelarnim prikazima, što je znatno olakšalo rad Textract servisu.

Takođe, zaključeno je da Textract servis ima bolje performanse u obradi štampanih, u odnosu na ručno pisanih podataka. Kada se uporede dva skupa podataka sa ručno pisanim tekstom, pokazalo se da je skup podataka na francuskom jeziku zahtevao više vremena za obradu od skupa podataka na engleskom jeziku, zbog lošijeg kvaliteta slike i dodatnih izazova.

Uzimajući u obzir sve navedene činjenice, može se zaključiti da je performansa Textract servisa značajno uslovljena kvalitetom ulaznih podataka, kao i vrstom teksta koji se obrađuje. Stoga, prilikom korišćenja ovog servisa u realnom okruženju, potrebno je voditi računa o kvalitetu ulaznih podataka, kako bi se dobili što bolji rezultati.

8. LITERATURA

[1] About AWS, preuzeto sa https://aws.amazon.com/about-aws/

[2] Artificial intelligence services AWS, preuzeto sa https://aws.amazon.com/machine-learning/ai-services/

[3] Amazon Textract, preuzeto sa https://aws.amazon.com/textract/

[4] AWS Lambda Function, preuzeto sa https://docs.aws.amazon.com/lambda/latest/dg/welcome.html

[5] Amazon Simple Storage Service, preuzeto sa https://aws.amazon.com/s3/

[6] Python, preuzeto sa https://www.python.org/

[7] Suminda Niroshan, “AWS Textract with Lambda Walkthrough” (28. Jun 2019), preuzeto sa https://medium.com/@sumindaniro/aws-textract-with-lambda-walkthrough-ed4473aedd9d

[8] Dr Urs Marti, Institut za računarske nauke i poslovnu matematiku, ETH Zurich, Švajcarska, “IAM Handwriting Database”, (1999) preuzeto sa https://fki.tic.heia-fr.ch/databases/iam-handwriting-database

[9] Nikolaos Arvanitopoulos, Gaspard Chevassus, Daniele Maggetti, Sabine Süsstrunk, “A Handwritten French Dataset for Word Spotting: CFRAMUZ” (Novembar 2017), preuzeto sa https://dl.acm.org/doi/10.1145/3151509.3151523

[10] Jens Walter, “my receipts (pdf scans)”, preuzeto sa https://www.kaggle.com/datasets/jenswalter/receipts

[11] Francisco Cruz, Mauro Castelli “Dataset of invoices and receipts including annotation of relevant fields” (21. Mart 2022), preuzeto sa https://zenodo.org/record/6371710#.ZAHGW3aZOUl

Kratka biografija:

Mia Knežević, rođena 19.02.1996. u Novom Sadu. Završila Gimnaziju “20. oktobar” u Bačkoj Palanci i osnovne akademske studije na Fakultetu Tehničkih nauka u Novom Sadu, smer Elektrotehnika i računarstvo. Ispunila je sve obaveze i položila je sve ispite predviđene studijskim programom.