Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.9
DOI: https://doi.org/10.24867/28BE33Tomcic

EKSTRAKTIVNA SUMARIZACIJA RECENZIJA HOTELA UPOTREBOM WORD2VEC, DOC2VEC I GPT-3.5 MODELA

EXTRACTIVE TEXT SUMMARIZATION OF HOTEL REVIEWS USING WORD2VEC, DOC2VEC AND GPT-3.5 MODELS

Sonja Tomčić, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIKA I RAČUNARSTVO

Kratak sadržaj – *U radu je prikazana ekstraktivna sumarizacija recenzija hotela upotrebom Word2Vec, Doc2Vec i GPT-3.5 modela. Objašnjene su teorijske osnove NLP-a, pomenutih modela i TextRank algoritma. Prikazan je proces sumarizacije teksta korišćen u implementaciji ovog rešenja. Skup podataka za treniranje i evaluaciju modela je preuzet sa sajta booking.com. Podaci su javno dostupni i nalaze se u vlasništvu sajta. U radu su prikazani eksperimenti izvršeni upotrebom različitih varijanti modela, kao i upotrebom različitih skupova podataka. Rezultati eksperimenata su poređeni međusobno, kao i sa rezultatima jednog od srodnih radova.*

Ključne reči: *Ekstraktivna sumarizacija, NLP, Word2Vec, Doc2Vec, GPT-3.5, TextRank*

Abstract – *This paper shows the extractive text summarization of hotel reviews using Word2Vec, Doc2Vec and GPT-3.5 models. The theoretical foundations of the NLP, the models and the TextRank algorithm are explained. The process of text summarization used in the implementation of this solution is presented. The dataset used for model training and evaluation was downloaded from the booking.com. The data is publicly available, and it is owned by the website. The paper shows experiments performed using different variations of the models, as well as using different datasets. The results of the experiments are compared with each other, as well as with the results of one of the related works.*

Keywords: *Extractive text summarization, NLP, Word2Vec, Doc2Vec, GPT-3.5, TextRank*

1. UVOD

Razvoj tehnologije i platformi za pronalaženje i rezervaciju smeštaja omogućio je njihovim korisnicima da brzo i lako sami pronalaze smeštaj bez dodatnih agencijskih troškova. Pomenute platforme pružaju mogućnost ostavljanja recenzija, što u značajnoj meri utiče na izbor potencijalnih korisnika. Komentari iz recenzija umeju da budu dugački i teški za čitanje i tumačenje, a njihova sumarizacija može da skrati vreme potrebno za odabir smeštaja i da na taj način korisnicima olakša donošenje odluke. Iako veoma korisna, ova opcija nije česta na platformama za pronalažanje i rezervaciju smeštaja.

Motivacija za pisanje ovog rada jeste to što bi uvođenje funkcionalnosti za sumarizaciju recenzija hotela omogućilo korisnicima tih platformi da brže i lakše donesu odluku prilkom odabira i rezervacije smeštaja. U radu je prikazan sistem za ekstraktivnu sumarizaciju recenzija hotela. Objašnjene su teorijske osnove NLP-a, Word2Vec, Doc2Vec i GPT-3.5 modela i TextRank algoritma, koji su korišćeni u implementaciji rešenja. Skup podataka je preuzet sa sajta booking.com, a podaci su javno dostupni i nalaze se u vlasništvu sajta.

Prikazan je celokupan proces sumarizacije, kao i eksperimenti izvršeni upotrebom različitih varijanti pomenutih modela i upotrebom različitih skupova podataka. Rezultati eksperimenata su poređeni međusobno, kao i sa rezultatima jednog od srodnih radova. Problem sumarizacije teksta, rešavan u ovom radu, pripada NLP (Natural Language Processing) oblasti, grani veštačke intelgincije koja se fokusira na interakciju između kompjutera i ljudskog jezika.

2. PREGLED STANJA U OBLASTI

U radu [1] iz 2019. godine autori su prikazali upotrebu Word2Vec i Doc2Vec algoritama u kombinaciji sa TextRank algoritmom, za pronalaženje ključnih reči u kratkim tekstovima sa društvenih mreža. Word2Vec algoritam je upotrebljen za pronalaženje semantičkih veza između reči, dok je Doc2Vec algoritam korišćen za pronalaženje vektora paragrafa i pronalaženje tačnosti. Rešenje je evaluirano pomoću accuracy, recall i F-measure metrika. Opisivana metoda se pokazala kao najbolja kada je u pitanju ekstrakcija ključnih reči kod kratkih tekstova, ali radi pouzdano i kod dugih tekstova. U radu [1] Word2Vec i Doc2Vec algoritmi su kombinovani kako bi se dobila semantička veza reči u tekstu, dok su u ovom radu te dve metode razdvojene.

Rad [2] iz 2020. godine opisuje rešenje problema sumarizacije biomedicinskh informacija u naučnim člancima, medicinskim zapisima, web dokumentima i kliničkim slikama. Za pronalazak lingvističke, semantičke i kontekstualne veze između rečenica, u radu [2] koriste se Word2Vec SkipGram i CBOW, context-free jezički modeli, kao i BioBERT, context-sensitive model. Za rangiranje koriste se PageRank, HITS i PPF tehnike. Rešenje je evaluirano upotrebom ROUGE-1 i ROUGE-2 metrike, a korpus za evaluaciju sadrži 2000 članaka iz

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

PubMed Central, gde je apstrakt članaka korišćen kao sumarizacija. Najbolji rezultat generalno dala je kombinacija BioBERT + GloVe + PageRank, a context-sensitive model se pokazao kao bolji u odnosu na context-free modele. U radu [2] prikazana je kombinacija različitih metoda za rangiranje i pronalazak veze između rečenica, dok su u ovom radu upotrebljene samo dve metode za traženje vektora kao dva odvojena rešenja u kombinaciji sa TextRank algoritmom.

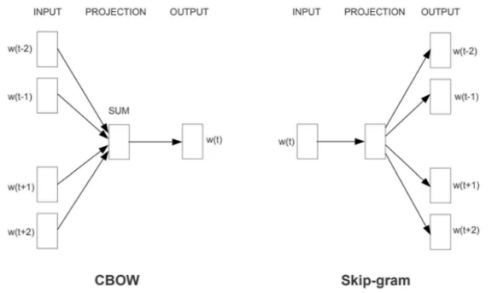
3. TEORIJSKI POJMOVI I DEFINICIJE

3.1. NLP (Natural Language Processing)

Natural Language Processing (NLP) je grana veštačke inteligencije koja omogućava računarima da procesiraju ljudski jezik u formi teksta ili govora na isti način kao što to rade ljudska bića. NLP kombinuje računarsku lingvistiku, koja predstavlja modelovanje ljudskog jezika zasnovano na pravilima, zajedno sa statističkim modelima i modelima mašinskog i dubokog učenja. Sumarizacija teksta, opisana u ovom radu, pripada NLP oblasti. Ekstraktivna sumarizacija podrazumeva identifikovanje i izdvajanje najvažnijih delova teksta, bez ikakvih modifikacija, što je tema ovog rada [3, 4].

3.2. Word2Vec

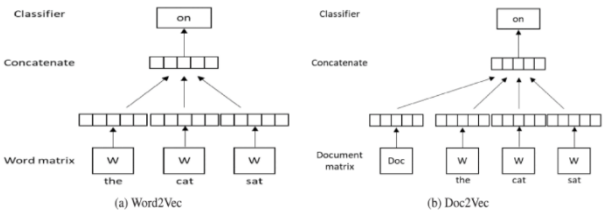
Word2Vec predstavlja metodu generisanja vektora reči pomoću neuronske mreže koja se sastoji od ulaznog, skrivenog i izlaznog sloja. Upotrebom ove metode, reči se predstavljaju kao vektori u kontinualnom vektorskom prostoru, kako bi se zabeležila semantička veza među njima. Reči sličnog značenja imaju sličnu vektorsku reprezentaciju, odnosno pozicionirane su blizu jedna drugoj u vektorskom prostoru, što je moguće predstaviti i izmeriti pomoću kosinusne sličnosti (eng. Cosine Similarity). Postoje dve različite arhitekture Word2Vec modela: Continuous Bag of Words (CBOW) i Skip-Gram. Na slici 1 prikazan je trening CBOW i Skip-Gram modela [5, 6].



Slika 1: *Trening CBOW i Skip-Gram modela [7]*

3.3. Doc2Vec

Doc2Vec je unsupervised algoritam za generisanje numeričkih vektora fiksne dužine koji reprezentuju dokument. Baziran je na Word2Vec modelu. Vektori se uče na sličan način kao kod Word2Vec modela - slični dokumenti se mapiraju na obližnje tačke u vektorskom prostoru, odnosno nalaze se blizu jedan drugom. Postoje dve varijante Doc2Vec modela: Distributed Memory (DM) i Distributed Bag of Words (DBOW). Na slici 2 prikazana je razlika između Word2Vec i Doc2Vec modela [8].



Slika 2: *Razlika između Word2Vec i Doc2Vec modela [9]*

3.4. TextRank algoritam

TextRank algoritam pripada algoritmima zasnovanim na graforvima, kod kojih se važnost temena unutar grafa određuje uzimajući u obzir globalne informacije rekurzivno izračunate iz celog grafa, umesto da se oslanjaju na lokalne informacije vezane za temena. TextRank algoritam je baziran na PageRank algoritmu koji se koristi za rangiranje web stranica prilikom online pretrage. Umesto web stranica koriste se rečenice iz teksta, a matrica sličnosti je popunjena brojevima koji označavaju sličnost između rečenica [10].

3.5. GPT-3.5 Turbo model

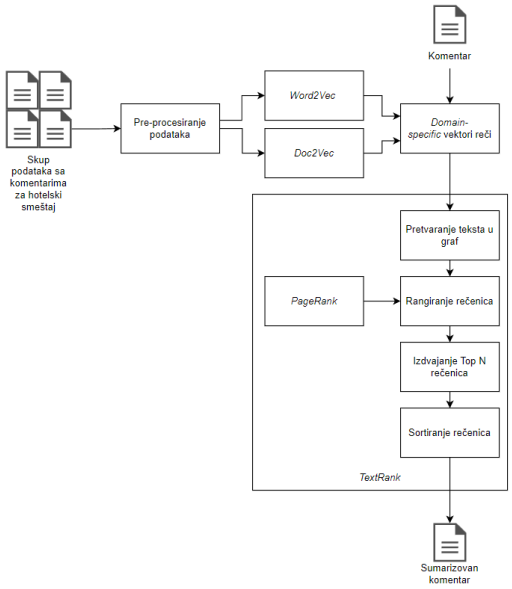
GPT-3.5 Turbo model pripada familiji GPT-3 modela i predstavlja jezički model baziran na neuronskim mrežama, treniran nad velikim skupom podataka. GPT-3.5 model je treniran nad preko 570GB podataka i sposoban je da uči kompleksne šablone u ljudskom jeziku, a njegova najbitnija osobina je duboko razumevanje ljudske komunikacije. Arhitektura GPT-3.5 modela je bazirana na transformer modelima.

Model je treniran po principu predikcije sledeće reči u rečenici, koristeći unsupervised učenje. Izgrađen je pomoću transformer dekoder blokova [11, 12].

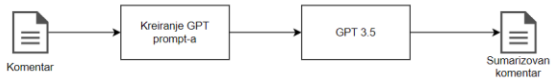
4. METODOLOGIJA

4.1. Arhitektura sistema

Za implementaciju rešenja kreirane su dve različite arhitekture. U prvoj arhitekturi upotrebljeni su Word2Vec i Doc2Vec modeli (slika 3), dok je u drugoj arhitekturi upotrebljen GPT-3.5 model (slika 4).



Slika 3: *Arhitektura sistema sa Word2Vec i Doc2Vec modelima*



Slika 4: *Arhitektura sistema sa GPT-3.5 modelom*

4.2. Implementacija rešenja

Kod arhitekture sa slike 3 pristuna su dva ulaza u sistem. Prvi ulaz je skup podataka sa komentarima iz recenzija za hotelski smeštaj. Komentare je potrebno pretprocesiraati kako bi bili pogodni za treniranje Word2Vec i Doc2Vec modela. Nakon pretprocesiranja, modeli se treniraju nad domain-specific skupom reči.

Drugi ulaz u sistem je komentar koji je potrebno sumarizovati. Na osnovu ulaznog komentara se prave domain-specific vektori reči pomoću istreniranih Word2Vec i Doc2Vec modela. U narednoj fazi rešenja, koja podrazumeva upotrebu TextRank algoritma, vektori reči iz prethodnog koraka se koriste za kreiranje grafa, a graf se dalje koristi za rangiranje rečenica pomoću PageRank algoritma, nakon čega se izdvaja top N rangiranih rečenica, koje se na kraju sortiraju. Izlaz iz sistema je sumarizovan ulazni komentar.

Na ulazu u sistem prikazan na slici 4 nalazi se originalni komentar. On se zatim koristi za kreiranje prompt-a, koji se prosleđuje GPT-3.5 modelu. Izlaz iz sistema predstavlja sumarizovan komentar.

5. EKSPERIMENTI

5.1. Skup podataka

Skup podataka za obučavanje modela korišćenih u arhitekturi sa slike 3 napravljen je scrape-ovanjem podataka sa sajta booking.com. Svi podaci su već javno dostupni i nalaze se u vlasništvu sajta.

Ovaj skup podataka kreiran je samo za potrebe rada i ne koristi se u druge svrhe, a sadrži komentare iz recenzija za smeštaj u hotelima u mestu Puket. Sa web sajta je preuzeto 27055 komentara. Obučavajući skup podataka se sastoji od 2188 komentara na engleskom jeziku, dužine od 30 ili više reči. Ukupan broj rečenica je 9906, dok je ukupan broj reči 72871, nakon što su izbačene engleske stop words. Za deo rešenja koje koristi arhitekturu sa slike 4 nije potreban obučavajući skup podataka.

5.2. Organizacija eksperimenata

Rešenje je evaluirano na osnovu četiri eksperimenta. U sva četiri eksperimenta koriste se Word2Vec (CBOW i Skip-gram), Doc2Vec (DBOW i DM) modeli i njihove vrste, kao i GPT-3.5 model. Za svaku od navedenih vrsta modela, podaci su pretprocesirani na tri načina: bez upotrebe Stemming i Lemmatization procesa, upotrebom Stemming procesa i upotrebom Lemmatization procesa. Sistem je testiran tako da vraća sumarizovane komentare dužine od tri i pet rečenica. U sva četiri eksperimenta postavka je ista, a razlikuje se jedino skup podataka za evaluaciju.

5.3. Eksperiment 1

U Eksperimentu 1 isti skup podataka se koristi za treniranje modela i za evaluaciju. Evaluacioni skup sadrži 20 kratkih komentara. Dužina originalnih komentara je 8 rečenica, a svaki komentar ima manje od 500 karaktera.

5.4. Eksperiment 2

U Eksperimentu 2 isti skup podataka se koristi za treniranje modela i za evaluaciju. Evaluacioni skup sadrži 20 dugačkih komentara. Dužina originalnih komentara je 10-12 rečenica, a svaki komentar ima više od 800, a manje od 1200 karaktera.

5.5. Eksperiment 3

U Eksperimentu 3 se skup podataka za treniranje modela i za evaluaciju razlikuju. Skup podataka za evaluaciju napravljen je ručno kopiranjem komentara za smeštaj na drugoj lokaciji. Za evaluaciju je izdvojeno 20 kratkih komentara, dužine kao u Eksperimentu 1.

5.6. Eksperiment 4

U Eksperimentu 4 se skup podataka za treniranje modela i za evaluaciju razlikuju. Skup podataka za evaluaciju napravljen je ručno kopiranjem komentara za smeštaj na drugoj lokaciji. Za evaluaciju je izdvojeno 20 dugačkih komentara, dužine kao u Eksperimentu 2.

5.7. Evaluacija

Za evaluaciju rezultata upotrebljen je ROUGE set metrika za evaluaciju automatske sumarizacije teksta na osnovu manuelno kreiranih referntnih sumarizacija, a dodatno rezultati su evaluirani i manuelno, na osnovu subjektivnog mišljenja autora. Korišćene su ROUGE-1, ROUGE-2 i ROUGE-L metrike. Rezultati ovih metrika prikazuju se pomoću recall, precision i F1 Score metrika. U navedenim eksperimentima prikazane su vrednosti F1 Score metrike, koja predstavlja harmonijsku sredinu precision i recall metrika.

6. REZULTATI

U tabeli 1 prikazani su najbolji rezultati iz sva četiri eksperimenta. Najbolji rezultat imao je Eksperiment 3 u kombinaciji Word2Vec CBOW + Stemming za sumarizaciju od 5 rečenica, što je delom iznenađujuće jer CBOW model u poređenju sa SkipGram modelom nije efikasan u beleženju nijansiranih veza između reči. Ovaj eksperiment vršen je nad novim komentarima koji su nepoznati sistemu, dok su rezultati Eksperimenta 1 i Eksperimenta 2, koja su vršena nad komentarima koji su već poznati sistemu, veoma slični mogu se smatrati jednako uspešnim. S obzirom na to da ovaj eksperiment nije mogao da se izvrši za pojedine Doc2Vec varijante, njegov rezultat se može smatrati nedovoljno relevantnim.

Tabela 1: *Ukupni rezultati eksperimenata*

Eksperiment	**Model**	**Pre-processing**	**N**	**ROUGE-1 F1**	**ROUGE-2 F1**	**ROUGE-L F1**
Eksp. 1	Doc2Vec DBOW	None	5	0.7119	0.6417	0.7099
Eksp. 2	Doc2Vec DM	None	3	0.7091	0.6389	0.7073
Eksp. 3	Word2Vec CBOW	Stemming	5	**0.7507**	**0.6848**	**0.7470**
Eksp. 4	Word2Vec SG	Lemmatization	5	0.6502	0.5509	0.6437

7. DISKUSIJA

U radu [2], gde su sumarizovani biomedicnki članci, korišćene su različite metode za rangiranje rečenica i kreiranje vektora reči. Rezultati su se pokazali boljim, ako se u obzir uzmu samo rezultati za kombinaciju Word2Vec + PageRank, s obzirom na to da su njeni rezultati relevantni za ovaj rad. Uzimajući u obzir veličinu skupova podataka za razvoj i evaluaciju, kao i to da su za referentne sumarizacije uzeti apstrakti članaka koji predstavljaju objektivnu sumarizaciju, za razliku od subjektivnih referentnih sumarizacija korišćenih u ovom radu, očekivano je da su vrednosti rezultata u radu [2] bolji u odnosu na rezultate u ovom radu.

Subjektivno mišljenje autora ovog rada upotrebljeno je kao dodatna metoda evaluacije pored ROUGE metode. Manuelno su evaluirani komentari za koje su vrednosti rezultata najbolje (tabela 1), kao i oni koji imaju najlošije rezultate. Može se zaključiti da za primere gde je rezultat dobar, sumarizacija daje smislene komentare. Najbolji rezultati su prisutni kod sumarizacija od pet rečenica, što ima simsla jer algoritam u tim slučajevima mora da izabere veći broj rečenica i to daje manje šanse za grešku. S obzirom na nemogućnost modela da konvergiraju u eksperimentima 3 i 4, ovaj sistem nije spreman za upotrebu u produkciji.

Probleme sa nemogućnošću sistema da konvergira je moguće rešiti promenom parametara modela i ponovnim testiranjem. Pored promene parametara modela, moguće unapređenje bi bilo i treniranje modela nad većim i raznovrsnijim skupom podataka. Evaluaciju ovog sistema je moguće poboljšati kreiranjem većeg evaluacionog skupa, tako da je on kreiran od strane više ljudi, kako bi se isključila subjektivnost jedne osobe, ili pronalaskom skupa podataka na internetu. Promenom načina kreiranja evaluacionog skupa bi i rezultati ekperimenata bili drugačiji, a najverovatnije i tačniji.

8. ZAKLJUČAK

U ovom radu predstavljen je sistem za ekstraktivnu sumarizaciju komentara iz recenzija za hotelski smeštaj. Dva od tri rešenja su implementirana u vidu dva odvojena dela: modul za treniranje modela za kreiranje vektora reči i modul za rangiranje rečenica. Treće rešenje ne koristi vektore reči niti algoritam za rangiranje rečenica i veoma je jednostavno za implementaciju.

Word2Vec CBOW model u kombinaciji sa Stemming pretprocesiranjem za sumarizacije od pet rečenica se pokazao kao najbolji uzimajući u obzir varijante modela i pretprocesiranja u kom je sistem uspešno izvršio sumarizaciju. Ovaj eksperiment je vršen nad kratkim nepoznatim komentarima. Rezultati dva eksperimenta koja su vršena nad komentarima koji su već poznati sistemu imaju veoma slične vrednosti i mogu se smatrati jednako uspešnim. Doc2Vec model je generalno imao bolje rezultate, ali za pojedine modele sistem nije uspeo da izvrši sumarizaciju do kraja, pa zbog toga određeni rezultati fale, a možda bi doprineli drugačijem ishodu i zaključku o uspešnosti sistema. GPT-3.5 model je bio srednje uspešan i sva četiri eksperimenta se slično ponašao.

Uspešnost ovog sistema mogla bi se povećati upotrebom većeg i raznovrsnijeg skupa podataka. Metode evaluacije je takođe moguće poboljšati kreiranjem većeg i manje subjektivnog evaluacionog skupa koji bi bio kreiran od strane više ljudi, što bi doprinelo drugačijim i tačnijim rezultatima. Kao dodatno poboljšanje sistema za ekstraktivnu sumarizaciju komentara, drugi algoritmi za rangiranje i kreiranje vektora reči bi mogli da se testiraju. Sistem bi dodatno mogao da se proširi mogućnošću da radi i za ostale jezike, a ne samo za enegleski jezik.

9. LITERATURA

[1] LI, JUN; HUANG, GUIMIN; FAN, CHUNLI; SUN, ZHENGLIN; and ZHU, HONGTAO (2019) "Key word extraction for short text via word2vec, doc2vec, and textrank," *Turkish Journal of Electrical Engineering and Computer Sciences*: Vol. 27: No. 3, Article 17.

[2] Milad Moradi, Maedeh Dashti, Matthias Samwald, "Summarization of biomediacal articles using domain-specific word embeddings and graph ranking", *Journal of Biomedical Informatics*: Vol. 107, July 2020, 103452.

[3] https://www.geeksforgeeks.org/natural-language-processing-overview/ (pristupljeno u januaru 2024.)

[4] https://medium.com/sciforce/towards-automatic-text-summarization-extractive-methods-e8439cd54715 (pristupljeno u februaru 2024.)

[5] https://www.geeksforgeeks.org/python-word-embedding-using-word2vec/ (pristupljeno u februaru 2024.)

[6]https://wiki.app.uib.no/info216/images/c/c4/IntroToWordEmbeddings.pdf (pristupljeno u februaru 2024.)

[7] https://medium.com/@manansuri/a-dummys-guide-to-word2vec-456444f3c673 (pristupljeno u februaru 2024.)

[8] https://www.geeksforgeeks.org/doc2vec-in-nlp/ (pristupljeno u februaru 2024.)

[9] A. Hernández-Castñeda, R. A. García-Hernández, Y. Ledeneva and C. E. Millán-Hernández, "Extractive automatic text summarization based on lexical-semantic keywords", *IEEE Access*, vol. 8, pp. 49896-49907, 2020.

[10] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text", *Proceedings of the 2004 conference on empirical methods in natural language processing*, pp. 404-411, July 2004.

[11] https://www.linkedin.com/pulse/chatgpts-guide-understanding-gpt-35-architecture-heena-koshti#:~:text=The%20GPT%2D3.5%20architecture%20is%20a%20neural%20network%2Dbased%20language,text%20completion%2C%20and%20question%20answering (pristupljeno u februaru 2024.)

[12] https://medium.com/nerd-for-tech/gpt3-and-chat-gpt-detailed-architecture-study-deep-nlp-horse-db3af9de8a5d (pristupljeno u februaru 2024.)

Kratka biografija:

Sonja Tomčić rođena je 04.02.1999. godine u Vrbasu. Osnovne akademske studije završila je 2022. godine na Fakultetu tehničkih nauka, nakon kojih upisuje master akademske studije na studijskom programu Primenjene računarske nauke i informatika, usmerenje Inteligentni sistemi. kontakt: sonjatomcic@gmail.com