Predlog diplomskog/master rada

1. **Definicija problema/cilja projekta.** Pravljenje chatbota koji kao ulaz prima pitanje, traži najsličnije pitanje u bazi i na njega vraća odgovor (koji se isto nalazi u bazi), sentiment tog odgovora i bolduje ključne reči u odgovoru.
2. **Motivacija problema rešavanog u projektu.** Ovaj chatbot bi mogao da se koristi na raznim sajtovima i pomogao bi ljudima da brže dođu do odgovora na pitanje, odmah bi znali da li je odgovor pozitivan ili negativan (ili možda neutralan) i dobili bi označene ključne reči tog odgovora kako im neki bitan deo ne bi promakao.
3. **Relevantna literatura (minimum 3 rada)**

**Sentiment Analysis: A Comparative Study on Different Approaches** <https://www.sciencedirect.com/science/article/pii/S187705091630463X>

* Ovaj rad predlaže različite metode za analizu sentimenta. Analiza se može odraditi na više različitih nivoa – na nivou dokumenta, rečenice i na nivou aspekta odnosno samog mišljenja.
* U ovom radu su isprobane razne metode mašinskog učenja (SVM, Naivni Bayes, KNN), metode zasnovane na određivanju pravila i leksičke metode.
* Nema informacija o skupu podataka koji je korišćen u ovom radu.
* Za evaluaciju rešenja je korišćen accuracy.
* Zaključak ovog rada je da metode mašinskog učenja daju mnogo bolje rezultate u odnosu na ostale metode koje su isprobane.

**Extracting Keywords from Texts based on Word Frequency and Association Features**

<https://www.sciencedirect.com/science/article/pii/S1877050921008139>

* Količina podataka koja se nalazi na internetu naglo raste i kako bi se sortirali, a kasnije i preporučivali svi ti podaci potrebno je izdvojiti ključne reči u njima. Ovaj rad predlaže da se za rešavanje ovog problema koristi TF-IDF.
* U radu se korisi nenadgledano učenje jer je labeliranje ključnih reči naporno, vremenski zahtevno i rezultati mogu biti subjektivni.
* Za ovaj eksperiment su korišćena 2 skupa podataka, News i Computer dataset.
* Za evaluaciju rezultata su koristili precision, recall i F meru
* Pokazalo se da rezultati dosta zavise od parametra koji određuje koliko ključnih reči je odabrano i ta razlika ide i do 10%.

**Research on Text Classification Techniques Based on Improved TF-IDF Algorithm and LSTM Inputs**

<https://www.sciencedirect.com/science/article/pii/S1877050922015058>

* Cilj ovog istraživanja je da se poboljša klasifikacija teksta tako što će se koristiti unapređeni TF-IDF algoritam, vektor reči i LSTM.
* Kao ulazi u LSTM se koristi kombinacija TF-IDF vrednosti i vektora reči, ona se dobija tako što se pomnoži argmax od TF-IDFa i vektor reči koji je dobijen pomoću word2vec metode.
* U ovom radu se koristi Sohu news skup podataka koji je podeljen na trening, test i validacioni skup u razmeri 8:1:1.
* Za evaluaciju rezultata se koristi accuracy, precision, recall i F mera. Korišćene su različite kombinacije parametara, a accuracy za najbolju kombinaciju je preko 94%.

**Research paper classification systems based on TF-IDF and LDA schemes**

<https://hcis-journal.springeropen.com/articles/10.1186/s13673-019-0192-7>

* Cilj ovog rada je da klasifikuje naučne radove po temama na osnovu ključnih reči.
* Ovi istraživači predlažu da se ključne reči traže samo u abstract-u jer se u njemu nalazi suština tog rada. Za pronalaženje kljlučnih reči koriste TF-IDF i LDA i izdvajaju se samo imenice, a za grupisanje sličnih radova koriste K-means klastering. Postoji 1394 jedinstvene ključne reči tako da je više sličnih reči predstavljeno jednom rečju.
* Kao skup podataka korišćeni su naučni radovi koji su objavljeni u časopisu FGCS u periodu od 1984 do 2017 godine. Ukupno je korišćeno 3246 radova.
* Za evaluaciju modela je korišćena F mera i primećeno je da model daje zadovoljavajuće rezultate.

**Siamese Multiplicative LSTM for Semantic Text Similarity**

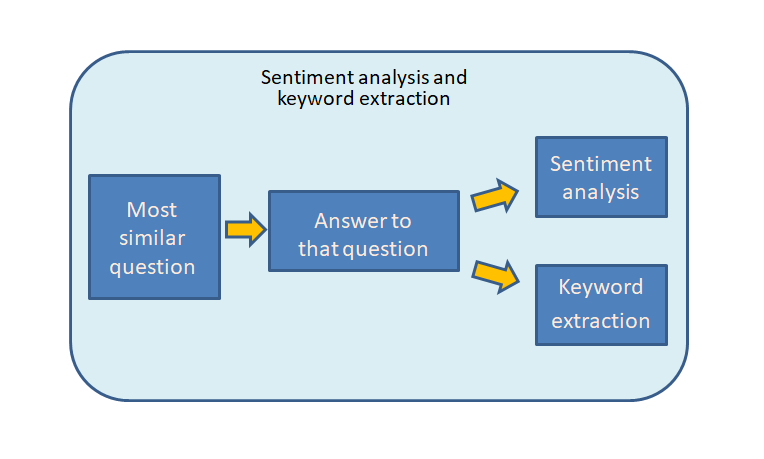
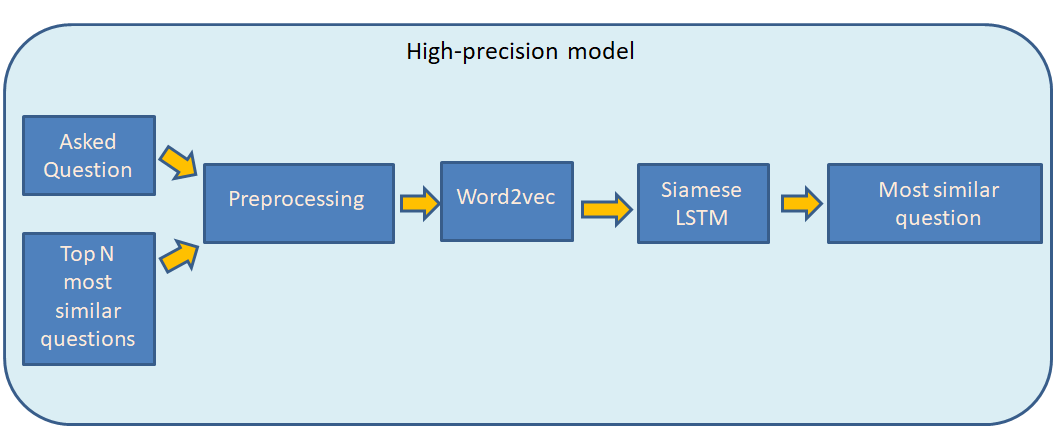
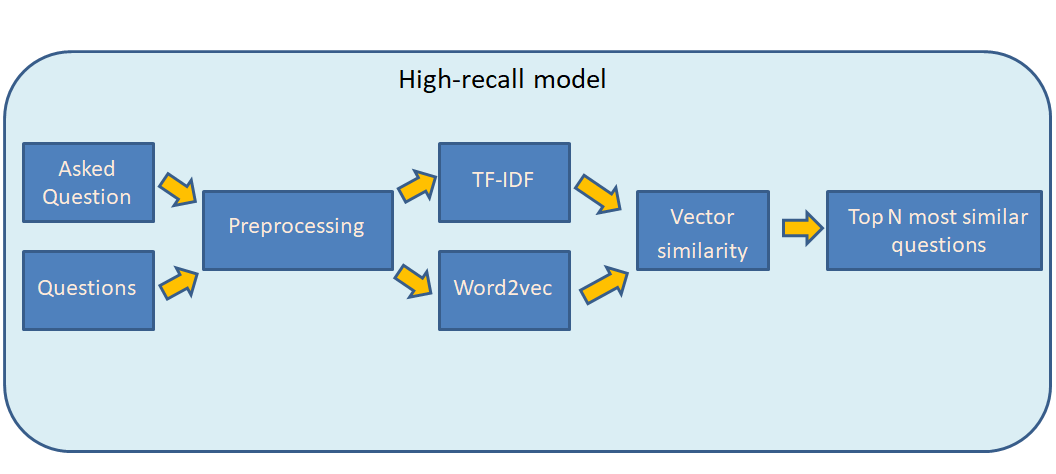
<https://dl.acm.org/doi/10.1145/3446132.3446160>

* Zadatak ovog rada je da se odredi sličnost između tekstova pomoću sijamskog LSTM-a.
* U radu se koriste 3 pristupa: sličnost između stringova, rečenica i knowledge-based (ne znam kako da prevedem)
* Korišćen je SICK skup podataka koji se sastoji od 10 000 parova rečenica na engleskom jeziku. Svaki par rečenica ima ocenu sličnosti od 1 do 5 koje je dobijena usrednjavanjem procene 10 različitih osoba.
* Za evaluaciju modela je korišćena Persanova i Spearmenova korelacija kao i srednja kvadratna greška.
* Rezultati pokazuju da predloženi pristup daje bolje rezultate.

1. **Skup podataka -** Opišite skup/skupove podataka koje planirate da koristite u projektu.

Postojeći (javno dostupan) skup podataka:

* Glavni skup podataka se sastoji od često postavljenih pitanja i odgovora iz oblasti osiguranja. Sadrži 16896 jedinstvenih pitanja i nema nedostajućih vrednosti. <https://github.com/shuzi/insuranceQA>
* Za obučavanje modela za klasifikaciju sentimenta ću koristiti IMDB Dataset of 50K Movie Reviews <https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews> koji ima dve kolone. U prvoj je tekst, a u drugoj je sentiment (pozitivan ili negativan) tog teksta. Ovaj skup ima 50 hiljada recenzija, nema nedostajućih vrednosti i odnos pozitivnih i negativnih recenzija je balansiran.

1. **Metodologija -** Chatbot će imati 4 modela: high-recall, high-precision, model za klasifikaciju sentimenta i model za ključne reči. Nakon preprocesiranja upotrebiću TF-IDF ili word2vec za dobijanje vektora reči. Za pronalaženje N najsličnijih pitanja koristiću kosinusno, menhetn ili euklidsko rastojanje, zavisi koje se najbolje pokaže, a za pronalaženje najsličnijeg pitanja koristiću LSTM. Što se tiče određivanja sentimenta koristiću NN, RF, SVM i KNN i izabrati najbolji model. Za određivanje ključnih reči ću koristiti TF-IDF i izvući imenice koje se najčešće ponavljaju. 
2. **Metod evaluacije** 
   1. **Mere performanse**Za **high-recall model** ću koristiti recall metriku jer je potrebno da se postavljeno pitanje nađe u skupu od N pitanja i nije bitno da li je prvo ili poslednje.  
      **High-precision model** će koristiti metrike kao što su tačnost i preciznost jer je potrebno da nađe tačno to pitanje iz baze koje je postavljeno.  
      **Model za klasifikaciju sentimenta** će koristiti više metrika uključujući tačnost, preciznost, f-meru...
   2. **Postupak evaluacije (eksperiment)** –   
      **High-recall model** će koristiti samo trening i test skup jer ne postoji dovoljan broj pitanja za pravljenje validacionog skupa. Odnos koji će se koristiti je 70-30.  
      **High-precision model** će koristiti će koristiti iste skupove kao high-recall model.  
      **Model za klasifikaciju sentimenta** će koristiti trening, validacioni i test skup jer postoji dovoljna količina podataka. Odnos koji će se koristiti je 70-15-15.
3. **Softver** – Za rešavanje ovog problema biće korišćen Python programski jezik i njegove biblioteke kao što su pandas, numpy, seaborn, scikit-learn...
4. **Github** - <https://github.com/stefan9988/Master_rad_chatbot>