Универзитет у Београду

Електротехнички факултет



Пројекат на предмету обрада природних језика

|  |  |
| --- | --- |
| Ментор: | Кандидати: |
| Др Вук Батановић | Алекса Вучковић  Милан Боићевић  Милица Јевтовић |

Београд, Септембар 2025.

Садржај

[Садржај i](#_Toc209644335)

[1. Увод 1](#_Toc209644336)

[2. Прикупљање података 2](#_Toc209644337)

[2.1. Новински текстови 2](#_Toc209644338)

[2.2. Књижевни текстови 2](#_Toc209644339)

[2.3. Правно-административни текстови 3](#_Toc209644340)

[2.4. Текстови са твитера/X-a 3](#_Toc209644341)

[3. Анотација података 4](#_Toc209644342)

[3.1. Правила анотације 4](#_Toc209644343)

[4. Евалуација статистичких модела 5](#_Toc209644344)

[4.1. Baseline приступ – мултиномијални наивни Бајесов класификатор 5](#_Toc209644345)

[4.2. CLASSLО 5](#_Toc209644346)

[4.2.1. Стандардни језик 5](#_Toc209644347)

[4.2.2. Нестандардни језик 5](#_Toc209644348)

[4.3. BERTић 5](#_Toc209644349)

[4.4. COMtext.SR 5](#_Toc209644350)

1. Увод

Пројекат из предмета **Обрада природних језика** за школску 2024/2025. годину реализован је од стране групе студената:

* Алекса Вучковић, 2024/3040
* Милан Богићевић, 2024/3056
* Милица Јевтовић, 2024/3113

Тема пројекта јесте израда **baseline приступа и евалуација напреднијих модела за препознавање именованих ентитета (Named Entity Recognition – NER)** на српском језику. Фокус је на три основна типа ентитета – **особе (PER), локације (LOC) и организације (ORG)**, при чему је коришћен IOB2 систем означавања. Пројекат је имплементиран у програмском језику **Python**.

Израда пројекта одвија се кроз три главне фазе:

1. **Прикупљање података** – формирање скупа текстова из различитих тематских домена (новински, књижевни, правно-административни, твитер), њихово чишћење и припрема за даље кораке.
2. **Анотација података** – ручно означавање именованих ентитета уз формулисање јасних упутстава и спровођење калибрације и анализе квалитета анотације.
3. **Евалуација статистичких модела** – поређење перформанси baseline приступа са више постојећих NER модела (CLASSLA, BERTић-NER, COMtext.SR) и детаљна анализа добијених резултата.
4. Прикупљање података

У оквиру пројекта било је неопходно формирати довољно велики и разноврстан корпус текстова на српском језику, који ће након анотације послужити за евалуацију различитих NER модела. Да би се обезбедила доменска разноликост и покривање различитих стилова писања, подаци су прикупљани из четири тематска домена:

1. **Новински текстови**
2. **Књижевни текстови**
3. **Правно-административни текстови**
4. **Поруке са друштвених мрежа (Твитер/X)**

Сваки извор је обрађен посебно, при чему је извршено аутоматско преузимање садржаја, чишћење текста и складиштење у унапред дефинисаној структури директоријума. Након тога сви текстови су пргледани мануелно и избачени су подаци који нису релевантни за даљу анализу. Ови текстови су даље коришћени за евалуацију различитих модела, који ће бити описани у четвртом поглављу, тако да се пазило да подаци који су коришћени за евалуацију не буду исти као они који су се користили за тренирање модела.

* 1. Новински текстови

Први извор података били су новински портали. За потребе пројекта одабран је сајт **nova.rs**, из секције „Вести“. Коришћен је Python програм заснован на библиотекама requests и BeautifulSoup, који је имао две основне функције:

* **Сакупљање линкова** – програм је прво слао HTTP GET захтев ка страници са вестима и прикупљао све хипервезе које воде ка појединачним чланцима. Линкови су филтрирани тако да се задрже само они који припадају домену https://nova.rs/vesti/. Дупликати су уклоњени, а јединствени линкови су снимљени у посебан фајл.
* **Преузимање чланака** – затим је сваки од прикупљених линкова обрађен тако што су извучени текстуални садржаји пасуса (<p> тагови). Текст је очишћен од вишеструких празнина и специјалних знакова, а чланци краћи од 100 карактера су прескочени како би се избегли нерелевантни уноси. Сваки чланак је сачуван као појединачан .txt фајл у UTF-8 енкодингу.

Уз сваки текст уметнут је и запис у метаподатке, који садржи име фајла и оригинални URL. На тај начин обезбеђена је трајна веза између извора и локално сачуваног текста.

* 1. Књижевни текстови

За књижевни домен коришћен је роман **„Ана Карењина” Лава Толстоја**, који је у целини јавно доступан у дигиталном формату. Текст је директно преузет и није вршена додатна модификација осим нормализације празних места и нових редова.

При обради је вођено рачуна да корпус садржи приближно **5000 токена.**

* 1. Правно-административни текстови

За домен правно-административних докумената прикупљени су подаци са портала **Paragraf** (<https://www.paragraf.rs>), који садржи законе, правилнике и друге прописе Републике Србије. Поред аутоматског преузимања већег броја докумената, као репрезентативан пример укључен је и **„Устав Републике Србије”**, који представља један од најзначајнијих правних текстова.

Скрапинг је реализован помоћу библиотека **requests** и **BeautifulSoup**, при чему су издвајани пасуси из <p> елемената HTML страница. Сваки документ је сачуван као посебан .txt фајл у директоријуму, док су метаподацима (име фајла и изворни линк) смештени у датотеци metadata.txt.

* 1. Текстови са твитера/X-a

Подаци су прикупљени претрагом на страници x.com/search, користећи филтере lang, since, until, min\_retweets, а подаци добављени у json формату су затим парсирани како би се издвојиле основне информације – текстуални садржај објаве и одговарајућа URL адреса. Садржај поруке је филтриран тако да су најпре избачени сувишни карактери (са уникод вредношћу већом или једнаком 0x10000), а затим одбачене објаве са мање од четрнаест речи, на основу размака. Подаци су смештени у data/twitter/metadata.json.

1. Анотација података
   1. Правила анотације
2. Евалуација статистичких модела
   1. Baseline приступ – мултиномијални наивни Бајесов класификатор
   2. CLASSLО
      1. Стандардни језик
      2. Нестандардни језик
   3. BERTић
   4. COMtext.SR