УВОД У ДИГИТАЛНУ ХУМАНИСТИКУ: *радионице за имплементацију „удаљеног читања“ у истраживачкој пракси* (2024)

Семинар УВОД У ДИГИТАЛНУ ХУМАНИСТИКУ: *радионице за имплементацију „удаљеног читања“ у истраживачкој пракси* - ће омогућити едукацију студената основних, мастер и докторских студија у контексту „удаљеног читања“ (Distant Reading), која је ургентна и у најбољем националном интересу. Наиме, више није довољно само истраживати културно наслеђе, већ то треба чинити користећи савремене, технолошки напредне алате, како би само културно наслеђе, као и истраживање по себи имало већу видљивост и глобални карактер. Парадигма „удаљеног читања“, коју је у студије књижевности увео Франко Морети, изузетан је оквир за унапређење истраживачког рада путем софистицираних техничких могућности без преседана у прошлости. Самим коришћењем ових могућности, повећава се видљивост српске књижевности и културе у међународној арени и обезбеђује савремени стручни интегритет истраживања. Како би процес био сасвим заокружен, неопходно је у њега укључити и библиотекаре, као највеће савезнике истраживача, као и одабрано особље које ради у просвети (Филолошка гимназија, одабране гимназије...), како би адекватан процес обуке започео већ у току средњошколског периода.

Посебан легитимитет овог семинара представља чињеница да се он наслања на активности које су успешно спроведене 2023. године у НКОЦ „Вук Караџић“, са изузетним одзивом студената, као и у оквиру пројеката *Удаљено читање* (2019) и *Читање издалека* (2022), али и кровне европске акције у оквиру програма COST CA16204 - *Distant Reading for European Literary History*, која је започета 2018. године, а трајала до 2022. Међу основним циљевима ове акције је било да се изгради вишејезичка европска колекција књижевних текстова (European Literary Text Collection (ELTeC)), која би садржавала око 2.500 комплетних текстова романа на бар 10 различитих европских језика што омогућава, између осталог, да се пореде резултати анализе овако добијених корпуса кроз националне културе. Српски језик су у акцији представљали Универзитет у Београду и то његове чланице: Универзитетска библиотека „Светозар Марковић“ и Филолошки факултет, а посебно битна улога је припала члановима Друштва за језичке ресурсе и технологије (JePTex). Захваљујући пројекту *Удаљено читање* из 2019. године, српски језик је постао један од десет европских језика који су до окончања пројекта остварили постављени циљ: 100 дигитализованих романа обрађених према заједнички договореним смерницама (немачки, енглески, француски, шпански, италијански, норвешки, португалски, румунски, словеначки и српски). Од тог тренутка, српски језик (српска књижевност) је постала једна од најрепрезентативнијих колекција читаве акције, те је посебно похваљена од стране конзорцијума. Српски део ELTeC корпуса представља основу за даљу изградњу дигиталне колекције српске књижевности различитих жанрова (ELTeC-plus) која сада садржи преко 120 романа, новела и путописа.

За формирање репрезентативне колекције текстова утврђени су, између осталог и следећи принципи, који су у потпуности поштовани приликом обраде:

- ☐ У корпус су се уносили интегрални текстови романа чије прво издање датира из 1840-1920. (како би се осигурало да не постоји ограничење ауторским правима) дужине најмање 10.000 речи. Када год је то могуће сканирала су се прва издања одабраних дела.

- Како би се постигла репрезентативност корпуса, романи који улазе у састав корпуса покривали су цео одабрани временски период, дела су различитих аутора (води се рачуна о заступљености женског писма) и различитог степена каноничности (обухватају су и заборављени романи који нису у језичком канону).
- Сви текстови су у дигиталној верзији опремљени мета-подацима, као и XML-етикетама које описују логички и графички изглед текста, као и поједине структуриране елементе текста у складу са препорукама пројекта TEI. Ово значи да се експлицитно означавају наслови, пасуси, фусноте и делови текста који су истакнути на посебан начин (нпр. делови на страном језику), те пагинација изворног издања текста.

Да би се све ово остварило било је потребно, пре свега, прибавити одабрана дела, сканирати их коришћењем опреме Универзитетске библиотеке, трансформисати тако добијену слику у текст (OCR), обавити аутоматску корекцију текста, извршити додатну коректуру текста и његово снабдевање анотацијама у виду XML-етикета и метаподацима (појединачних текстова и корпуса у целини). Такође, текстови корпуса су успешно аутоматски обележени специфичним језичким објектима: именима људи, локација, организација, догађаја и сл. Ово обележавање је такође у потпуности усклађено са препорукама COST акције *D-Reading*. На тај начин су дигиталне верзије обрађених дела претраживе семантичким кључевима, што је значајно за романе из колекције, а још више за путописе.

Семинаром *УВОД У ДИГИТАЛНУ ХУМАНИСТИКУ* предвиђа се организовање предавања и радионица намењених како онима који ће се ангажовати на припреми текстова за корпус/дигиталну колекцију тако и будућим корисницима. Предвиђено је да радионице буду радног карактера. Полазници ће се кроз практичан рад упознати са целом линијом дигитализације: корекција сканираних и очитаних текстова, анотирање њихове структуре и изгледа, припремање мета-података. Напредним радионицама предвиђа се обучавање полазника за проверу и корекцију аутоматски обележених језичких објеката: имена људи, њихових улога, локација и друго. За ове потребе користиће се алати доступни на: http://nerbeyond.jerteh.rs/. Такође, радионице су намењене пре свега студентима (основне, мастер и докторске студије) хуманистичке оријентације, али и наставном особљу специјализованих средњих школа и младим истраживачима.

Организација радионице: комбиновање теоријског и практичног рада. Осим теоријских излагања тема наведених у претходној секцији, полазници би имали прилику да припреме један текст, од корекција и обележавања TEI етикета до аутоматске анотације именованих ентитета. Припремљене текстове би полазници користили за повезивање са отвореним подацима, а потом би уследили практични задаци удаљеног читања. Ресурси за практичан рад: srpELTeC корпус.

Дневни распоред: предавања се организују пре подне, у периоду од 10-13 часова са краћим паузама, а радионице после подне од 16-18 часова.

ТЕМЕ СЕМИНАРА:

Распоред за 5 дана:

Дан 1 (понедељак 9. 12.)

Теоријски део (пре подне):

- Важност дигиталне хуманистике данас, резултати досадашњих пројеката, генералне информације о COST акцији и циљевима дигиталне хуманистике на овом примеру, дигитална хуманистика и отворена наука, питање канона и дигиталне хуманистике..
- процес дигитализације (скенирање, рашчитавање), алатке које се користе. На примеру https://udaljenocitanje.unilib.rs/ демонстрација разлика слике и рачитаног текста. Карактеристични проблеми за рашчитавање ћириличног и латиничног писма, проблеми старих и ретких књига.

Практичан рад (после подне): Сваки полазник креира сопствени мини пројекат у ком креће од рашчитавања скенираног документа, након чега следе корекције текста. Примери ће обухватити оба писма.

Дан 2 (уторак 10. 12.)

Теоријски део (пре подне):

- Обрада дигиталних текстова
- Сегментација текста на реченице и токенизација. Врсте и значај анотација. Ручна анотација структуре документа.
- Значај стандардизације (TEI, пример Вукових пословица, ELTeC-а). Аутоматстка анотација токена врстама речи и лемама, аутоматска анотација делова текста именованим ентитетима (места, особе, организације, професије,...).
- Напомене о напреднијим врстама анотације: синтакса, директан говор, семантика (значења), стилске фигуре. Неопходност контроле и евалуације.

Практичан рад (после подне): Ручна анотација структуре документа сагласно TEI препорукама, контрола ручне анотације (добро формирани и валидни XML документи); аутоматска анотација врстама речи и лемама (http://obrada.jerteh.rs/), именованим ентитетима (https://ners.jerteh.rs/). Контрола и евалуација аутоматске анотације (http://inception.jerteh.rs/).

Дан 3 (среда 11. 12.)

Теоријски део (пре подне):

- Отворени подаци, упознавање са википодацима о српским романима: основни подаци о роману и његовим издањима, подаци о аутору, о местима радње, главним ликовима и њиховим везама. Ручни и аутоматски унос википодатака коришћењем алата OpenRefine и QuickStatements. Демонстрација рада са Википодацима из пајтон свески.
- Претраживање података језиком SPARQL и визуелизација резултата табеларно, у виду мапе, графа, стабла. Интеграција резултата упита у презентације

Практичан рад: сваки полазник ће унети по један роман у Википодатке, креирати упите са различитим излазима и интегрисати их са документом (HTML).

Дан 4 (четвртак 12. 12.)

Теоријски део (пре подне):

- Улога библиотека у дигиталној хуманистици: пружање приступа дигиталним збиркама и базама података, обука и подршка у коришћењу дигиталних технологија за анализу и обраду хуманистичких података; очување дигиталне културне баштине (дигиталних копија старих рукописа и других важних историјских докумената). Улога библиотекара у дигиталној хуманистици на примеру конкретног прикупљања материјала за припрему корпуса.

Практичан рад (после подне): рад у програму *Transkribus*, напредно претраживање у COBISS-у.

Дан 5 (петак 13. 12.)

Теоријски део (пре подне):

- Читање у дигиталном окружењу - е-књиге. Упознавање са корпусима доступним на https://noske.jerteh.rs/. Основе CQL језика и демонстрација упита над корпусом srpELTeC. Анализа резултата коришћењем конкорданци, фреквенција, извоз резултата. Текстометријска анализа текста коришћењем алата TXM: креирање корпуса, генерисање речника текста, конкорданце и фреквенције CQL образаца, креирање и анализа партиција и подкорпуса, анализа специфичности.
- Шири контекст српског језика у доба револуције вештачке интелигенције (ВИ), могућности и проблеми у вези са обрадом српског језика коришћењем ВИ, посебно у контексту великих језичких модела (попут OpenAI GPT), који тренутно привлаче глобалну пажњу. Статички и динамички језички модели. Иницијатива за развој јавно доступних квалитетних ресурса и алата за обраду српског коришћењем модела ВИ, како би се сачувала позиција српског језика у доба АИ револуције.

Практичан рад:

- Постављање CQL упита на https://noske.jerteh.rs/, почевши од једноставних упита, преко сложенијих грматичких образаца до упита који укључују етикете именованих ентитета. Инсталација и рад са алатом TXM: креирање сопственог малог корпуса текстова и његова текстометријска анализа.
- Коришћење модела ГПТ за српски https://plma.jerteh.rs, поређење модела фамилије BERT за српски.
- Инструкције (промптни инжењеринг) на примеру ChatGPT и других модела.

ПРЕДАВАЧИ:

Др Василије Милновић – координатор пројекта, Руководилац Центра за науку Универзитетске библиотеке „Светозар Марковић“ и доктор књижевних наука.

Проф. др Цветана Крстев – професор Филолошког факултета у пензији, руководилац Српског учешћа у COST акцији, суоснивач Друштва за језичке ресурсе и технологије.

Проф. др Ранка Станковић – професор Рударско-геолошког факултета, Катедра за примењену математику и информатику и суоснивач Друштва за језичке ресурсе и технологије.

Проф. др Душко Витас – професор Математичког факултета у пензији, оснивач Семинара за рачунарство и информатику и носилац највиших одликовања Републике Француске за допринос науци (посебно рачунарској лингвистици). Оснивач платформе АУРОРА и оснивач Друштва за језичке ресурсе и технологије.

Др Александра Тртовац – доктор библиотекарства, главни редактор каталога у оквиру COBISS.Net система.