Република Србија
Министарство културе

УВОД У ДИГИТАЛНУ ХУМАНИСТИКУ:
радионица за имплементацију „удаљеног читања“ у истраживачкој пракси

Семинар УВОД У ДИГИТАЛНУ ХУМАНИСТИКУ: радионице за имплементацију „удаљеног читања“ у истраживачкој пракси - ће омогућити едукацију истраживачког и библиотекарског кадра у контексту „удаљеног читања“ (Distant Reading), која је неодложна 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 само по себи имало већу видљивост и глобални карактер. Парадигма „удаљеног читања“, коју је у студије књижевности увео Франко Морати, 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 на међународном пољу и обезбеђује савремени стручни интегритет истраживања. Како би процес био сасвим заокружен, неопходно је у њега укључити и библиотекаре, као највеће савезнике истраживача.

Посебну аутентичност овог семинара представља чињеница да се он наслања на активности које су успешно спроведене 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[1] (2018-2022). Један од основних циљева ове акције била је изградња вишејезичке европске колекције књижевних текстова (European Literary Text Collection - ELTeC), која би садржала 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 је у акцији представљао 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 тренутка завршетка пројекта, српски језик и српска књижевност су постали једна од најрепрезентативнијих колекција читаве акције, те су посебно похваљени од стране конзорцијума. Српски део корпуса ELTeC представља основу за даљу изградњу

[1] https://www.cost.eu/actions/CA16204/

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 се уносе 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 покривају цео одабрани временски период, дела су различитих аутора (води се рачуна о заступљености женских аутора) и различитог степена каноничности (обухватају се и заборављени романи који нису у језичком канону).
- Сви текстови су у дигиталној верзији опремљени мета-подацима, као и XML анотацијом која описују логички и графички изглед текста, као и поједине структуриране елементе текста у складу са препорукама пројекта TEI[2]. То 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 XML анотацијом 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 по семантичким кључевима, што је значајно за романе из колекције, а још више за путописе.
Семинаром УВОД У ДИГИТАЛНУ ХУМАНИСТИКУ предвиђа се организовање радионице намењене онима који ће се ангажовати на припреми текстова за корпус/дигиталну колекцију тако и будућим корисницима. Предвиђено је да радионице буду радног карактера. Радионица је намењена будућим волонтерима који кроз практичан рад треба да се упознају са целом линијом дигитализације: корекција сканираних и очитаних текстова, анотирање њихове структуре и изгледа, припремање мета-података. Напредне активности предвиђају обучавање полазника за проверу и корекцију аутоматски обележених језичких објеката: имена људи, њихових улога, локација и друго. Радионица је намењена стручњацима хуманистичке оријентације (наука о књижевности, лексикографија итд), пре свега студентима мастер и докторских студила, али су добродошли и други полазници.

Реализатори радионице: др Василије Милновић, др Александра Тртовац, проф. др Душко Витас, проф. др Цветана Крстев, проф. др Ранка Станковић, Слободан Марковић

Циљ радионице: Упознавање студената филолошких факултета са методама дигиталне хуманистике, припреме и обраде текста за примену техника „удаљеног читања“

[2]http://www.tei-c.org/

Време извођења: 4-9 децембар 2023
Трајање извођења: 5 дана
Теме предавања:

1. Увод у дигиталну хуманистику: циљеви и методе
   Теоријски део: Резултати досадашњих пројеката, генералне информације о COST акцији и важности дигиталне хуманистике. Циљ дигиталне хуманистике: коришћење информационих технологија за приступ, претрагу и преглед хуманистичких извора; истраживање културног наслеђа и најбитнијих хуманистичких питања. Методе: дигитализација и транскрипција рукописних и штампаних докумената; креирање и коришћење база података и онлајн архива, проналажење и екстракција информација, машинско учење и остале гране вештачке интелигенције, визуелизација података и графички приказ резултата истраживања.

2. Библиотеке и дигитална хуманистка
   Теоријски део: 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). Методе: чување, описивање и проналажење информација. Улога библиотекара у дигиталној хуманистици на примеру конкретног прикупљања материјала за припрему корпуса, укључујући и COBISS.
   Практичан рад: Напредно претраживање система COBISS.

3. Дигитализација: од слике до дигиталног текста
   Теоријски део: 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. Онлине алати, слободни и комерцијални алати.
   Практичан рад: Сваки полазник креира сопствени мини пројекат у ком креће од рашчитавања скенираног документа, након чега следе корекције текста. Примери ће обухватити оба писма. Конкретан рад sa различитим алатима (https://texase.jerteh.rs/, https://www.ocrconvert.com/, ...)

4. Обрада дигиталних текстова.
   Теоријски део: Сегментација текста на реченице и токенизација. Врсте и значај анотација. Ручна анотација структуре документа. Значај стандардизације (TEI, пример Вукових пословица, ELTeC-a). 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: 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/).

5. Отворени подаци: пут до отворених података и како доћи до њих.
   Теоријски део: 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 и веб стране. Упознавање са осталим видовима отворених података о романима.
   Практичан рад: сваки полазник ће унети по један роман у Википодатке, креирати упите са различитим излазима и интегрисати их са документом (HTML).

6. Читање, читање изблиза (корпуси, конкорданце), читање из далека (статистичка обрада, визуализација, TXM)
   Теоријски део: 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.
   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.

7. Дигитална хуманистика и дигитални подаци у функцији великих језичких података
   Теоријски део: Изазови отворених података: прикупљање, публиковање, доступни и недостајући скупови. 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: Коришћење модела ГПТ за српски https://palma.jerteh.rs, поређење модела фамилије BERT за српски. Инструкције (промптни инжењеринг) на примеру ChatGPT и других модела.

Врста 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.

Организација радионице: Препоручује се да сваки полазник понесе сопствени лаптоп, а приступом интернету ће бити обезбеђен. Полазници плаћају само трошкове превоза, смештај и храна су обезбеђени, а сама радионица бесплатна.

Дневни распоред: предавања се обично организују пре и после подне: углавном теоријски део 10-13 часова пре подне са неким краћим паузама и после подне од 16-18/19 часова практични.

Пријавите се попуњавањем следећег кратог формулара
https://forms.gle/4T7H6d2chUkz3aWh7 .

Пријава за радионицу ће бити отворена до понедељка 27. новембра 2023. године.