

Већу за студије при Универзитету
Универзитета у Београду

Пријава теме за израду мастер рада

Обраћам се Већу за студије при Универзитету Универзитета у Београду са молбом да ми се, као студенту мастер академских студија *Рачунарство у друштвеним наукама*, одобри тема за израду мастер рада.

Наслов рада:

МОДЕЛОВАЊЕ ТЕМА РОМАНА О ХАРИ ПОТЕРУ ПРИМЕНОМ LDA И BERTOPIC МЕТОДА

Предмет рада

У овом раду се применом метода тематског моделовања *Latent Dirichlet Allocation – LDA* и *BERTopic* анализира тематска структура романа *Harry Potter and the Philosopher's Stone* (Rowling, 1997). У питању је један од најпознатијих романа савремене фантастике, чија популарност и комплексна књижевна композиција чине његов текст погодним за рачунарско испитивање скривених семантичких и тематских образаца. Основна идеја тематског моделовања јесте да се квантитативним техникама обраде текста открију латентне тематске целине, које се затим могу упоредити са постојећим књижевно-теоријским тумачењима. Кључни концепти који ће бити обрађени у овом истраживању укључују интерсекцију теоријске тематске анализе, структурне наратологије, као и методе обраде природног језика (*LDA* и *BERTopic*). Ови концепти су значајни јер омогућавају вишеслојно разумевање књижевног текста — како кроз његове унутрашње семантичке и симболичке слојеве, тако и кроз формалне обрасце и језичке регуларности. Посебан значај ове анализе огледа се у могућности да се упореде два различита приступа моделовању — класични *LDA*, који претпоставља дистрибуцију тема као скуп статистички издвојених целина, и савременији *BERTopic*, који омогућава финије семантичке разлике и богатију интерпретативну структуру (Grootendorst, 2022). На овај начин, рад има за циљ да сагледа у којој мери рачунарска анализа може да допринесе дубљем разумевању наратива, митских мотива и симболике у роману, као и да покаже ограничења и предности различитих техника тематског моделовања. Истраживање се наслања на класичне и савремене теорије наративне структуре и архетипова у књижевности, од Фрајевог *оквира жанрова* (Frye, 1957) до Кампбеловог *пута хероја* (Campbell, 2004), а уједно уважава и критичка читања попут Вајтедовог указивања на *клишеизирани мотиве* (Whited, 2004) и Хејлмановог разматрања идеолошких матрица *рода, класе и расе* (Heilman, 2008). Такав теоријски контекст отвара простор да се испита колико тематско моделовање може потврдити или преиспитати постојећа тумачења у роману о Хари Потеру. Према тренутним сазнањима, претходна истраживања иста као ово нису публикована.

Предмет анализе у овом раду биће:

1. Тематске целине издвојене применом *LDA* модела, са нагласком на њихову кохерентност и упоредивост са књижевно-теоријском литературом;
2. Тематске целине издвојене применом *BERTopic* модела, које би требало да понуде интерпретативно прецизније и богатије резултате, а такође и упоредиве са

књижевно-теоријском литературом;

3. Компаративна анализа оба модела у погледу броја, садржаја и интерпретативне вредности тема, али и преклапања екстрахованих тема са онима из теоријско-књижевне анализе.

Оваква поставка омогућава да се роман *Harry Potter and the Philosopher's Stone* посматра као корпус подложен квантитативној анализи, али и као текст у чијем тумачењу је неопходно комбиновати дигитално-хуманистичке методе и класичну књижевну херменевтику. На тај начин, рад настоји да покаже како примена рачунарских метода у области књижевности може осветлити наративне обрасце из једне нове, интердисциплинарне перспективе.

Циљеви рада и полазне хипотезе

Сврха овог истраживања јесте допринос интердисциплинарном повезивању рачунарских метода обраде текста са теоријским оквиром књижевне анализе кроз примену тематског моделовања на један од најпознатијих романа савремене фантастике — *Harry Potter and the Philosopher's Stone* (Rowling, 1997). Главни циљ је да се квантитативним приступом открију и прикажу латентне семантичке и тематске структуре присутне у тексту горепоменутог романа, те да се ови резултати упореде и валидирају у односу на постојећа књижевно-теоријска тумачења истог. На овај начин, истраживање има за свој примарни циљ да пружи увид у степен до којег рачунарска анализа књижевног дела може да допринесе дубљем разумевању наратива, митских мотива и наративних образаца.

Конкретнији циљеви истраживања су следећи:

1. Испитати које тематске целине се могу идентификовати применом *LDA*, представника традиционалних модела тематског моделирања, и *BERTopic*, представника савремених модела тематског моделирања, на роману *Harry Potter and the Philosopher's Stone*?
2. Испитати у којој мери се издвојене теме структурно поклапају са већ утврђеним књижевним, наративним и митским обрасцима?
3. Испитати који од ова два модела пружа тематске резултате који су интерпретативно релевантнији и погоднији за књижевну анализу?

У складу са циљевима истраживањима, постављене су и хипотезе самог истраживања:

Прва хипотеза (X1): Очекује се да ће оба модела и *LDA* и *BERTopic* понудити сличан број екстрахованих тематских целина текста романа *Harry Potter and the Philosopher's Stone*. Водећи се резултатима претходних истраживања, претпоставља се да ће број димензија бити око 7 (Rafferty, 2018), као и то да ће семантичке тематске целине бити сличне онима које препозанје књижевно-теоријска литература (Heilman, 2008; Frye, 1957). Такође, следећи резултате претходних истраживања (Whited, 2004; Nikolajeva, 2003), очекује се да ће сам садржај екстрахованих тематских целина бити семантички сличан.

Друга хипотеза (X2): Очекује се и то да ће *BERTopic* модел екстраховати тематске целине које имају финију семантичку разлику у односу на оне екстраховане уз помоћ *LDA*, као и то

да ће *BERTopic* модел предложити и интерпретативно богатије теме у поређењу са класичним *LDA* моделом (Grootendorst, 2022).

Методe

За анализу ће бити коришћене методе *Latent Dirichlet Allocation (LDA)* и *BERTopic*, које представљају савремене приступе тематског моделовања. Целокупан процес тематског моделовања биће реализован у *Jupyter Notebook* окружењу (Kluuyver et al., 2016) користећи *Python3* (Van Rossum & Drake, 2009). Припрема текста за моделовање (токенизација, лематизација, уклањање знакова интерпункције, личних имена и слично) вршиће се помоћу библиотеке *NLTK* (Bird, Klein, & Loper, 2009) и *spaCy* (Honnibal et al., 2020). За издвајање латентних тема прво ће бити примењена метода *LDA* из *Python* библиотеке *gensim*, док ће за савременији приступ бити примењен *BERTopic*, који комбинује *transformers* моделе (Reimers & Gurevych, 2019) са методама кластеризације заснованим на *HDBSCAN* алгоритму. Визуелизација резултата биће реализована уз помоћ библиотеке *pyLDAvis* (за интерактивни приказ *LDA* тема) (Sievert & Shirley, 2014), *BERTopic* уграђених метода за редуkcију димензионалности и приказ тема, а по потреби и *Matplotlib* (Hunter, 2007) и *Plotly* (Inc., 2015). Сем рачунарских метода, биће извршена и књижевна интерпретација резултата, кроз поређење са филолошком (књижевном) критиком.

Оквирна структура рада

1. Увод

1.1 Предмет рада

1.2 Циљеви рада и истраживачке хипотезе

2. Преглед релевантних области

2.1 Приказ романа о Хари Потеру

2.2 *LDA* и *BERTopic*: дескрипција метода и њихова употреба за анализу текста

2.3 Теоријско-књижевна приступ роману о Хари Потеру: претходна истраживања

2.4 Рачунарско тематско моделовање романа о Хари Потеру: претходна истраживања

3. Материјал и методе

3.1 Дескрипција сета података (текст романа)

3.2 Дескрипција коришћених пакета у *Python*-у

3.3 Дескрипција процедуре моделовања

4. Резултати

4.1 Моделовање текста уз помоћ *LDA*

4.2 Моделовање текста уз помоћ *BERTopic*

5. Дискусија

5.1 Поређење резултата добијених *LDA* и *BERTopic*

5.2 Поређење резултата моделовања и теоријско-књижевних резултата

5.3 Ограничења реализованог истраживања и предлози за даља истраживања

6. Литература

Оквирна селектована литература

- Al Mamun, M. H., Keikhosrokiani, P., Asl, M. P., Anuar, N. A., Hadi, N. H., & Humida, T. (2022). Sentiment Analysis of the Harry Potter Series Using a Lexicon-Based Approach. In P. Keikhosrokiani & M. Pourya Asl (Eds.), *Handbook of Research on Opinion Mining and Text Analytics on Literary Works and Social Media* (pp. 263-291). IGI Global Scientific

Publishing. <https://doi.org/10.4018/978-1-7998-9594-7.ch011>

- Bettelheim, B. (1976). *The uses of enchantment: The meaning and importance of fairy tales*. New York: Vintage Books.
- Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. Sebastopol, CA: O'Reilly Media.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Booker, C. (2004). *The seven basic plots: Why we tell stories*. London: Continuum.
- Campbell, J. (1949). *The hero with a thousand faces*. New Jersey: Princeton University Press.
- Danesi, M. (2017). *The semiotics of emoji: The rise of visual language in the age of the internet*. London: Bloomsbury Academic.
- Frye, N. (1957). *Anatomy of criticism: Four essays*. New Jersey: Princeton University Press.
- Grootendorst, M. (2022). *BERTopic: Neural topic modeling with class-based TF-IDF*. arXiv preprint arXiv:2203.05794.
- Heilman, E. E. (Ed.). (2008). *Critical perspectives on Harry Potter*. Oxford: Routledge.
- Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python*. Retrieved from <https://spacy.io/>
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90–95. <https://doi.org/10.1109/MCSE.2007.55>
- Inc., P. (2015). *Plotly*. Retrieved from <https://plotly.com/python/>
- Iser, W. (1976). *The act of reading: A theory of aesthetic response*. Baltimore: Johns Hopkins University Press.
- Jockers, M. L., & Mimno, D. (2013). Significant themes in 19th-century literature. *Poetics*, 41(6), 750–769. <https://doi.org/10.1016/j.poetic.2013.08.007>
- Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., ... & Willing, C. (2016). Jupyter Notebooks—a publishing format for reproducible computational workflows. In F. Loizides & B. Schmidt (Eds.), *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (pp. 87–90). IOS Press. <https://doi.org/10.3233/978-1-61499-649-1-87>
- Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55–60). <https://doi.org/10.3115/v1/P14-5010>
- Moretti, F. (2005). *Graphs, maps, trees: Abstract models for a literary history*. London: Verso.
- Nikolajeva, M. (2003). Harry Potter: A return to the romantic hero. In E. E. Heilman (Ed.), *Harry Potter's world: Multidisciplinary critical perspectives* (pp. 125–140). Routledge Falmer.
- Rafferty, G. (2018, December 7). *LDA on the texts of Harry Potter: Topic modeling with Latent Dirichlet Allocation*. Towards Data Science. <https://medium.com/data-science/basic-nlp-on-the-texts-of-harry-potter-topic-modeling-with-latent-dirichlet-allocation-f3c00f77b0f5>
- Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). European Language Resources Association (ELRA).
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (pp. 3982–3992). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1410>
- Rowling, J. K. (1997). *Harry Potter and the Philosopher's Stone*. London: Bloomsbury.
- Sculley, D., & Pasanek, B. (2008). Meaning and mining: The impact of implicit assumptions in data mining for the humanities. *Literary and Linguistic Computing*, 23(4), 409–424.

- Shostack, A. (2024). *The boy who survived: Removing Harry Potter from an LLM is harder than reported* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2403.12082>
- Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces* (pp. 63–70). Association for Computational Linguistics. <https://doi.org/10.3115/v1/W14-3110>
- Underwood, T. (2019). *Distant Horizons: Digital Evidence and Literary Change*. Chicago: University of Chicago Press.
- Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. Charleston: CreateSpace.
- Vogler, C. (2007). *The writer's journey: Mythic structure for writers (3rd ed.)*. StudioCity, CA: Michael Wiese Productions.
- Whited, L. (2004). Harry Potter and the mythic hero. In *The Ivory Tower and Harry Potter: Perspectives on a literary phenomenon* (pp. 85–98). University of Missouri Press.

Радна биографија

Исидора Гатарић рођена је 1992. године у Новом Саду, где је 2015. године завршила основне академске студије психологије на Филозофском факултету Универзитета у Новом Саду, са просечном оценом 9,53. Мастер академске студије психологије на истом факултету завршила је 2016. године са просечном оценом 9,83, одбранивши мастер рад из истраживачке/квантитативне психологије на тему „Когнитивна обрада именица са вишезначним деривационим суфиксима: студија очних покрета“. Овај мастер рад ушао је у најужи избор за награду за најбољи мастер рад одбрањен те године у Републици Србији, коју додељује фонд „Катарина Марић“. Тренутно на истом факултету завршава докторске академске студије, под менторством проф. др Петра Чоловића. Током академске каријере била је ангажована као истраживач на пројектима Министарства науке, при Лабораторији за експерименталну психологију (од 2012. до 2016. године), као и предавач у Истраживачкој станици Петница (од 2015. до данас). Након академске каријере, од 2017. године ради у индустрији на позицији аналитичара података од банке Интезе, преко Инфостуда, Цегера, а данас води тим који се бави подацима и вештачком интелигенцијом у ВМГ Диџитл Трансформејшн компанији. Добитник је великог броја стипендија и награда, од којих се издвајају: 30 испод 30 (2022. године), Сакура стипендија (2017. године), Награда за најбоље дипломе Универзитета у Новом Саду (2015. године), Изузетна награда за научни рад Универзитета у Новом Саду (2015. године), Доситејева стипендија (2014. године), Стипендија Министарства науке за најбоље студенте (од 2012. до 2016. године), Награда Универзитета у Новом Саду за успех у току студија (од 2013. до 2016. године).

Објављени радови

Gatarić, I., Srdanović, S., & Kovač, A. (2021). The processing of process and result deverbal nominals in Serbian and English. *Psihologija*, 54(4), 381-396.
<https://doi.org/10.2298/PSI190928031G>

Gatarić, I., Srdanović, S., & Kovač, A. (2019). Do morphological differences affect the cognitive processing of deverbal nominals in Serbian? *Primenjena psihologija*, 12(2), 139-156. <https://doi.org/10.19090/pp.2019.2.139-156>

Gatarić, I. (2019). The cognitive processing of derived nouns with ambiguous suffixes: behavioral and eye-movement study. *Primenjena psihologija*, 12(1), 85-104.
<https://doi.org/10.19090/pp.2019.1.85-104>

Filipović-Đurđević, D., & Gatarić, I. (2018). Simultaneous effects of inflectional paradigms and classes in processing of Serbian verbs. *Psihologija*, 51(3), 259-288.
<https://doi.org/10.2298/PSI170811015F>

Менторка:

Проф. др Јелена Јовановић



Студенткиња:

Исидора Гатарић



Већу за студије при Универзитету
Универзитета у Београду

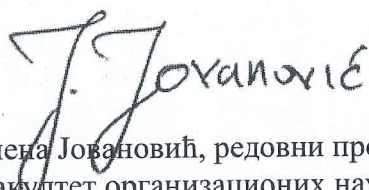
Студијски програм:

РАЧУНАРСТВО У ДРУШТВЕНИМ НАУКАМА

Наслов тезе:

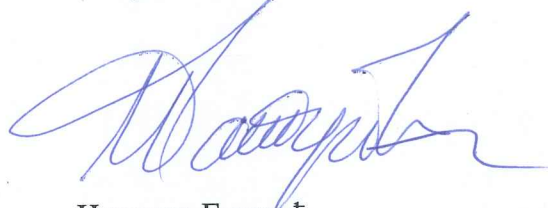
**МОДЕЛОВАЊЕ ТЕМА РОМАНА О ХАРИ ПОТЕРУ ПРИМЕНОМ LDA И BERTOPIC
МЕТОДА**

Менторка



Др Јелена Јовановић, редовни професор
Факултет организационих наука
Универзитет у Београду

Студенткиња



Исидора Гатарић
Бр. индекса:
19/2024

Будући да је кандидаткиња полагањем испита, обављањем стручне праске и одбраном приступног рада остварила обавезних 42 ЕСПБ (и додатних 6 ЕСПБ на факултативном предмету, што у суми даје 48 ЕСПБ), Комисија предлаже да јој се одобри израда предложеног мастер рада.

Чланови Комисије:

Проф. др Ранка Станковић, Катедра за примењену математику и информатику, Рударско-геолошки факултет, Универзитета у Београду

Др Нина Илић, Катедра за англистику, Филозофски факултет, Универзитета у Новом Саду

Изјава о академској честитости

Исидора Гатарић

Број индекса: 19/2024

Својим потписом гарантујем да ћу се у току студија, а нарочито у току израде мастер рада, у потпуности уздржати од плагирања, односно копирања туђих идеја и њиховог презентовања као сопствених, без одговарајућег признавања ауторства.

Разумем да копирање туђих идеја може бити извршено из књига и часописа и података који могу бити представљени у табелама, дијаграмима, дизајну, плановима, фотографијама, филму, музици, формулама, вебсајтовима, компјутерским програмима и на други начин. Јасно ми је да плагијат такође укључује и представљање, употребу и дистрибуирање рада предавачица или других студената као сопствених.

Свестна сам да намерним присвајањем идеја других и приказивањем као сопствених чиним повреде Закона о ауторском и сродним правима, као и других закона и одговарајућих аката Универзитета у Београду.

У Београду, 3/9/2025

Потпис студента

