Dokumentacija s primeri uporabe skozi razvoj projekta

Povezava: https://github.com/biolab/text-semantics/tree/main/examples

Branje dokumentov s strežnika

Primer uporabe API-ja za prenos korpusov s strežnika ter branje in izpisovanje dokumentov z meta podatki.

https://github.com/biolab/text-semantics/blob/main/examples/01-01-loading-documents.ipynb

Branje ontologij s strežnika

Primer uporabe API-ja za prenos ontologije iz strežnika, branje in izpisovanje ontologije.

https://github.com/biolab/text-semantics/blob/main/examples/01-02-ontologies.ipynb

Branje člankov Contributions to Contemporary History

Primer uporabe API-ja za prenos člankov s strežnika ter branje in izpisovanje dokumentov z meta podatki.

https://github.com/biolab/text-semantics/blob/main/examples/01-03-CTCH-exploration.ipynb

Branje člankov Elektrotehniškega vestnika

Primer uporabe API-ja za prenos člankov s strežnika ter branje in izpisovanje dokumentov z meta podatki.

https://github.com/biolab/text-semantics/blob/main/examples/01-04-el.vestnik-exploration.ipynb

Predobdelava dokumenta

Primer predobdelave dokumenta z delitvijo na pojavnice, filtriranjem odvečnih besed in prikazom v oblakov besed.

https://github.com/biolab/text-semantics/blob/main/examples/02-01-document-exploration.ipynb

Predobdelava dokumenta

Prikaz razlike med standardno predobdelavo besedila ter predobdelavo z odstranjevanjem strukturnih delov zakonskih aktov.

https://github.com/biolab/text-semantics/blob/main/examples/02-02-preprocessing-results.ipynb

Pridobitev vektorskih predstavitev besedil

V tem zvezku predstavimo, kako lahko pridobimo vektorske predstavitve (vložitve) besed in dokumentov za analizo besedil.

https://github.com/biolab/text-semantics/blob/main/examples/03_01_vector_representation_of_documents.ipynb

Uporaba razdalj in podobnosti

V tem zvezku predstavimo, kako med prej dobljenimi vložitvami dokumentov računamo in uporabljamo razdalje in podobnosti.

https://github.com/biolab/text-semantics/blob/main/examples/03_02_distances_and_similarities.ipynb

Odkrivanje skupin in izris kart dokumentov

Dimenzionalnost vektorskih predstavitev dokumentov lahko zmanjšamo na 2, kar nam omogoča prikaz dvodimenzionalne karte dokumentov, na kateri vsaka točka predstavlja dokument. Poleg tega lahko v dvodimenzionalnem prostoru odkrijemo skupine podobnih dokumentov in vsako skupino na karti obarvamo z različno barvo. To nam omogoča dober vpogled v celotno množico dokumentov.

https://github.com/biolab/text-semantics/blob/main/examples/03_03_document_maps.ipynb

Vektorske predstavitve besed

V tem zvezku predstavimo, kako lahko pridobimo vektorske predstavitve (vložitve) besed in kako izrišemo karto izrazov.

https://github.com/biolab/text-semantics/blob/main/examples/03_04_vector_representation_of_words.ipynb

Odkrivanje skupin in izris kart dokumentov na podlagi besed specifičnih za skupine

Dimenzionalnost vektorskih predstavitev dokumentov lahko zmanjšamo na 2, kar nam omogoča prikaz dvodimenzionalne karte dokumentov, na kateri vsaka točka predstavlja dokument. Poleg tega lahko v dvodimenzionalnem prostoru odkrijemo skupine podobnih dokumentov in vsako skupino na karti obarvamo z različno barvo. To nam omogoča dober vpogled v celotno množico dokumentov.

https://github.com/biolab/text-semantics/blob/main/examples/03_05_document_maps_specific.ipynb

Iskanje besed specifičnih za dokumente z uporabo vložitev fastText

https://github.com/biolab/text-semantics/blob/main/examples/04_01_specific_words_with_embeddings.ipynb

Iskanje besed specifičnih za dokumente z uporabo obogatitve besed

V prejšnjem primeru smo iskali specifične besede za dokumente z vložitvami dokumentov. V tem primeru, pa bomo v ta namen uporabili metodo imenovano obgatitev besed (ang. Word enrichment).

https://github.com/biolab/text-semantics/blob/main/examples/04_02_specific_words_with_enrichment.ipynb

Iskanje besed, specifičnih za dokumente z uporabo transformacije TF-IDF

Tokrat bomo poskusili specifične besede v dokumentu določiti z uporabo transformacije TF-IDF, ki uteži besede glede na njihovo pogostost v besedilu. Besede, ki močno zaznamujejo manjšo množico dokumentov, bodo tako imele večjo težo kot take, ki so vseprisotne v korpusu.

https://github.com/biolab/text-semantics/blob/main/examples/04_03_specific_words_with_tfidf.ipynb

Iskanje besed, specifičnih za dokumente z uporabo metod na grafih besed

Tokrat bomo poskusili specifične besede v dokumentu določiti z uporabo metod na grafih. Uporabili bomo metodi TextRank in RAKE. Metodi zgradita graf sopojavitev besed ter na podlagi grafa točkuje besede in fraze.

https://github.com/biolab/text-semantics/blob/main/examples/04_04_specific_words_graph_base.ipynb

Primerjava pristopov za specifične besede

V tej skripti primerjamo pristope za izbor specifičnih (ključnih) besed v besedilih.

https://github.com/biolab/text-semantics/blob/main/examples/04_05_specific_words_comparison.ipynb

Primerjava pristopov za luščenje ključnih besed na anotiranih besedilih iz revije Prispevki za novejšo zgodovino

V tem zvezku predstavljamo primerjavo pristopov za luščenje ključnih besed iz besedil iz revije Prispevki za novejšo zgodovino, ki imajo označene ključne besede. Ključne besede so označene s strani avtorjev člankov.

https://github.com/biolab/text-semantics/blob/main/examples/04_06b_specific_words_comparison_ctch_with_max_similarity.ipynb

Primerjava pristopov za luščenje ključnih besed na anotiranih besedilih iz revije Elektrotehniški vestnik

V tem zvezku predstavljamo primerjavo pristopov za luščenje ključnih besed iz besedil iz revije Elektrotehniški vestnik, ki imajo označene ključne besede. Ključne besede so označene s strani avtorjev člankov.

https://github.com/biolab/text-semantics/blob/main/examples/04_07_specific_words_comparison_el_vestnik.ipynb

Razlaga izraznih kart dokumentov

V tem zvezku bomo skupine označili s ključnimi besedami, kar razloži skupno tematiko dokumentov v skupini. Za pripis ključnih besed smo pripravili več metod, zato bomo pokazali razlago izraznih kar s štirimi različnimi metodami.

https://github.com/biolab/text-semantics/blob/main/examples/04_08_document_maps_explanation.ipynb

Primerjava pristopov za luščenje ključnih besed na izhodiščnih podatkih Schutz 2008

V tem zvezku predstavljamo primerjavo pristopov za luščenje ključnih besed iz nabora besedil Schutz 2008. Korpus sestavlja 1.231 člankov s področja medicine (PubMed Central), pri čemer ključne besede podajo avtorji člankov.

https://github.com/biolab/text-semantics/blob/main/examples/04_08_specific_words_comparison_schutz2008.ipynb

Razlaga izraznih kart dokumentov

V tem zvezku bomo skupine označili s ključnimi besedami, kar razloži skupno tematiko dokumentov v skupini. Za pripis ključnih besed smo pripravili več metod, zato bomo pokazali razlago izraznih kar s štirimi različnimi metodami.

https://github.com/biolab/text-semantics/blob/main/examples/04_08b_document_maps_explanation-laws.ipynb

Primerjava pristopov za luščenje ključnih besed na izhodiščnih podatkih SemEval

V tem zvezku predstavljamo primerjavo pristopov za luščenje ključnih besed iz nabora besedil SemEval, ki vsebuje 244 polnih člankov o računalništvu s portala ACM, pri čemer ključne besede podajo avtorji člankov.

https://github.com/biolab/text-semantics/blob/main/examples/04_09_specific_words_comparison_semeval.ipynb

Primerjava pristopov za luščenje ključnih besed na izhodiščnih podatkih SemEval

V tem zvezku predstavljamo primerjavo pristopov za luščenje ključnih besed iz nabora besedil SemEval, ki vsebuje 244 polnih člankov o računalništvu s portala ACM, pri čemer ključne besede podajo avtorji člankov.

https://github.com/biolab/text-semantics/blob/main/examples/04_09b_specific_words_comparison_semeval_including_transformers.ipynb

Primerjava pristopov za luščenje ključnih besed na povzetkih člankov s ključno besedo "Longevity"

V tem zvezku predstavljamo primerjavo pristopov za luščenje ključnih besed iz nabora povzetkov člankov s ključno besedo "Longevity" v zbirki PubMed.

https://github.com/biolab/text-semantics/blob/main/examples/04_10_specific_words_comparison_longevity.ipynb

Primerjava pristopov za luščenje ključnih besed na povzetkih člankov s ključno besedo "Longevity"

V tem zvezku predstavljamo primerjavo pristopov za luščenje ključnih besed iz nabora povzetkov člankov s ključno besedo "Longevity" v zbirki PubMed.

https://github.com/biolab/text-semantics/blob/main/examples/04_10_specific_words_comparison_longevity_lematizer.ipynb

Primerjava pristopov za luščenje ključnih fraz na povzetkih člankov s ključno besedo "Longevity" - not lemmatized

V tem zvezku predstavljamo primerjavo pristopov za luščenje ključnih besed iz nabora povzetkov člankov s ključno besedo "Longevity" v zbirki PubMed.

https://github.com/biolab/text-semantics/blob/main/examples/04_11_keyphrases_comparison_longevity.ipynb

Primerjava pristopov za luščenje ključnih fraz na povzetkih člankov s ključno besedo "Covid-19"

V tem zvezku predstavljamo primerjavo pristopov za luščenje ključnih besed iz nabora povzetkov člankov s ključno besedo "Longevity" v zbirki PubMed.

https://github.com/biolab/text-semantics/blob/main/examples/04_11_keyphrases_comparison_longevity_lematizer-covid19.ipynb

Primerjava pristopov za luščenje ključnih fraz na povzetkih člankov s ključno besedo "Longevity"

V tem zvezku predstavljamo primerjavo pristopov za luščenje ključnih besed iz nabora povzetkov člankov s ključno besedo "Longevity" v zbirki PubMed.

https://github.com/biolab/text-semantics/blob/main/examples/04_11_keyphrases_comparison_longevity_lematizer.ipynb

Annotated visualisations of longevity abstracts for AIIM publication – keyphrases

https://github.com/biolab/text-semantics/blob/main/examples/04_11_longevity_visualizations_aiim_covid.ipynb

Annotated visualisations of longevity abstracts for AIIM publication – keyphrases

https://github.com/biolab/text-semantics/blob/main/examples/04_11_longevity_visualizations_aiim_phrases.ipynb

Annotated visualisations of longevity abstracts for AIIM publication

https://github.com/biolab/text-semantics/blob/main/examples/04_11_longevity_visualizations_aiim_words.ipynb