Kaj je na dnevnem redu?

Tematsko modeliranje parlamentarnih razprav pred in med epidemijo covida-19

# Uvod

V demokratičnih državah parlament deluje kot osrednje predstavniško in zakonodajno telo. To pomeni, da ga sestavljajo izvoljeni predstavniki državljanov[[1]](#footnote-1), ki v njihovem imenu sodelujejo pri oblikovanju in sprejemanju zakonov ter tako urejajo vsa področja življenja in delovanja v družbi. Parlament pogosto nadzoruje tudi delo izvršilne oblasti (Norton, 2002). Zaradi ključne vloge, ki jo ima parlament pri razvoju in življenju družbe, je delovanje parlamenta vedno znova pomemben in aktualen predmet humanističnih in družboslovnih raziskav.

V zadnjih dveh desetletjih so zaradi napredka tehnologije, povečanega zanimanja medijev in državljanov za delo parlamenta ter želje po večji transparentnosti podatki o dogajanju v parlamentu, vključno z zapisi parlamentarnih sej, postali bolj dostopni (Norton, 2002). Zapisi sej so edinstven raziskovalni vir, saj parlamentarne razprave odsevajo politično, družbeno in kulturno vzdušje določenega časa (Ilie, 2010). Zaradi ustaljenosti parlamentarnega diskurza in digitalnega formata pa so zapisi parlamentarnih sej tudi priročen vir za izdelavo parlamentarnih korpusov. Pri tem gre za časovno zamejene in strukturirane nabore zapisov sej, ki so dodatno opremljeni z metapodatki o govorcih in govorih ter jezikoslovnimi oznakami (Truan in Romary, 2021).

Parlamentarni korpusi običajno vsebujejo velike količine podatkov, ki jih je v razumnem času nemogoče obdelati ročno. Eno od priljubljenih orodij za analizo korpusov so konkordančniki, s katerimi se lahko spoznate v [sorodnem učnem gradivu](https://sidih.github.io/voices/index-sl.html) (gl. Fišer in Pahor de Maiti, 2021), druga orodja, npr. [Orange](https://orangedatamining.com/) (Demšar idr., 2013), ki ga bomo uporabljali v tem učnem gradivu, pa omogočajo uporabo naprednih avtomatskih analitičnih tehnik oz. t. i. rudarjenje besedil, na podlagi katerega je iz velikih količin podatkov mogoče izluščiti vzorce in informacije, ki na prvi pogled niso razvidni iz besedila (Wiedemann, 2016).

Tehnike rudarjenja besedil so raznovrstne in se jih med drugim uporablja za analizo čustvene zaznamovanosti parlamentarnih razprav (Rheault idr., 2016; Rudkowsky idr., n.d.), prepoznavo politične opredelitve (Bergmann idr., 2018) ali stališča poslancev (Abercrombie in Batista-Navarro, 2020), modeliranje argumentacije v parlamentarnih razpravah (Petukhova idr., 2015) in podobno. Ena od njih je tudi tematsko modeliranje (Meeks in Weingart, 2012), ki sodi med najbolj razširjene tehnike s področja besedilnega rudarjenja v digitalni humanistiki in ki se ji bomo posvetili v tem učnem gradivu.

Namen učnega gradiva je raziskovalce s področja humanistike in družboslovja vpeljati v svet rudarjenja besedil ter prikazati vrednost tovrstnih pristopov za družboslovne in humanistične raziskave. Učno gradivo razlaga posebnosti parlamentarnega diskurza in uporabo tematskega modeliranja za reševanje konkretnih raziskovalnih vprašanj. Analiza temelji na prosto dostopnem korpusu slovenskih parlamentarnih razprav [ParlaMint](http://hdl.handle.net/11356/1432) (Erjavec idr., 2021) ter na orodju [Orange](https://orangedatamining.com/) (Demšar idr., 2013), ki uporabniku brez znanja programiranja omogoča uporabo naprednih metod rudarjenja besedil.

# Pregled učnega gradiva in navodila za uporabo

Učno gradivo je razdeljeno na teoretični in praktični del. V teoretičnem delu predstavimo značilnosti parlamentarnih razprav in korpus ParlaMint (3. poglavje) ter metodo tematskega modeliranja (4. poglavje). Praktični del se začne s poglavjem, ki nas vodi skozi pripravo na analizo in v katerem je razloženo, kako namestiti program Orange, uvoziti in pregledati podatke ter pripraviti vzorec podatkov za analizo (5. poglavje). V naslednjem poglavju preidemo na osrednji praktični del učnega gradiva (6. poglavje), ki ga sestavljajo tri povezane naloge. V teh nalogah z metodo tematskega modeliranja in različnimi vizualizacijami raziščemo, o čem so govorili poslanci, kateri temi so posvetili največ pozornosti in katere teme so še posebej izstopale med epidemijo v primerjavi s predepidemičnim obdobjem.

Vsi viri in orodja, uporabljeni v tem učnem gradivu, so prosto dostopni na spletu. Podrobna navodila glede prenosa podatkov [ParlaMint](http://hdl.handle.net/11356/1432) in namestitve programa [Orange](https://orangedatamining.com/) so navedena v poglavju 5.1. Če vas zanima predvsem analiza besedil v programu Orange, lahko začnete s 5. poglavjem, vendar priporočamo, da si preberete tudi začetna teoretična poglavja, saj vsebujejo ključne informacije za razumevanje podatkov in metode tematskega modeliranja, kar zmanjša možnost za nekritično uporabo metode in neustrezno interpretacijo rezultatov.

Poleg opisa postopkov učno gradivo vsebuje tudi številne posnetke zaslonov, ki prikazujejo nastavitve gradnikov[[2]](#footnote-2) in dobljene rezultate. Ob strani vsakega praktičnega poglavja je prikazan delotok, ki prikazuje sosledje uporabljenih gradnikov. Celoten delotok je na voljo za prenos v poglavju 5.1, vendar priporočamo, da delotok sestavite sami ob sledenju navodilom v učnem gradivu, saj boste tako najbolje razumeli posamezne korake analize. Imena **gradnikov** so v besedilu zapisana odebeljeno, *nastavitve gradnikov*, *imena spremenljivk, iskalni ukazi, obravnavane besede* ter *angleške ustreznice izrazov* pa so označene ležeče.

Oranžni okvirčki *Poskusite sami* vsebujejo navodila za samostojno raziskovanje podatkov in utrjevanje pridobljenega znanja.

# Parlamentarne razprave

V učnem gradivu analiziramo govore poslancev na parlamentarnih sejah. V tem poglavju se zato najprej posvetimo nekaterim splošnim značilnostim, ki veljajo za parlamentarne razprave, saj je dobro poznavanje gradiva ključnega pomena tako pri oblikovanju raziskovalnih vprašanj kot pri interpretaciji rezultatov.

## Parlamentarne razprave

Zaradi svoje institucionalne narave ima parlament jasno določeno strukturo in kompleksna pravila delovanja[[3]](#footnote-3), skozi zgodovino pa so se oblikovali tudi številni neformalni dogovori obnašanja (Norton, 2002). Ta pravila se med posameznimi parlamenti razlikujejo, spreminjajo pa se tudi skozi čas (Sieberer idr., 2011), zato jih je za ustrezno zasnovo analize in interpretacijo podatkov nujno treba poznati. Poleg tega se je pri analizi treba zavedati tudi vpliva, ki ga imajo na parlamentarne razprave lokalni in globalni politični kontekst, različne javne in zasebne vloge parlamentarcev in razmerja moči med njimi ter različna občinstva, ki lahko spremljajo isto razpravo (npr. ostali parlamentarci, gosti, javnost) (Ilie, 2010).

Parlamentarne seje so vedno natančno strukturirane: imajo torej opredeljen dnevni red, seje vodi za to določena oseba, predajanje besede pa poteka po jasnih pravilih (prim. Proksch in Slapin, 2010). Posebna pravila veljajo tudi za specifične točke dnevnega reda oziroma vrste razprav, npr. za poslanska vprašanja in pobude ali za interpelacije. Vse to pomembno oblikuje in omejuje parlamentarni diskurz, tj. dejanje sporazumevanja v specifičnem okolju parlamenta.

Parlamentarni diskurz je sicer le del širšega pojma političnega diskurza, znotraj katerega predstavlja najbolj institucionalizirani, s pravili zamejeni in uradni podtip (Bayley, 2004). Parlamentarni diskurz je tako ključna značilnost parlamenta, ki je osrednji prostor skupnosti za politično razpravo. Pri tem pa ni pomembna zgolj vsebina razprave, ampak v veliki meri tudi način govora oziroma diskurzivne strategije, ki jih govorci uporabljajo v svojih govorih, ter ostale, nejezikovne okoliščine. Trend raziskav parlamentarnega dogajanja gre zato vse bolj v smeri interdisciplinarnega pristopa h gradivu, ki omogoča široko interpretacijo dogodkov in procesov, vzrokov in posledic (Bayley, 2004; Ilie, 2010).

## Korpusi parlamentarnih razprav

Temeljni vir za raziskovanje parlamentarnega diskurza so zapisi parlamentarnih sej, ki so za večino parlamentov že javno dostopni v digitalni obliki. Obstoj zapisov, predvsem pa njihova dostopnost v digitalni obliki je izjemnega pomena tako za javnost, saj lajša nadzor nad delom parlamenta, kot tudi za raziskovalno skupnost, vendar je dejanska uporabnost zapisov odvisna od raziskovalnega problema (prim. Mollin, 2007). Tako kot parlamentarni diskurz imajo namreč tudi zapisi sej svoje posebnosti, ki izhajajo iz tradicije zapisovanja, ki velja v posameznem parlamentu (smernice za zapisovanje se med parlamenti razlikujejo in običajno niso prosto dostopne), in iz dejstva, da so zapisi pisna različica sicer govorjenih nastopov poslancev. Zapisi parlamentarnih sej kot uradni pisni viri so nedvomno verodostojni na ravni vsebine, ne pa nujno tudi na ravni izgovorjenega besedila. Zapisi namreč [niso natančne transkripcije govorov](https://sidih.github.io/voices/ch4-sl.html#ch4.2-sl) in zato običajno ne vsebujejo, vsaj ne v celoti, prvin govorjenega jezika (npr. mašil, popravkov) in številnih informacij o neverbalni komunikaciji v dvorani (npr. prekinitev, gest), ki jih je sicer mogoče zaznati ob spremljanju dejanskega nastopa govorcev (Bayley, 2004). Pogosto pa zapisi vsebujejo tudi različne dodatne informacije oziroma metapodatke, npr. seznam govorcev, rezultate glasovanja, obravnavano gradivo ipd.

Parlamentarni zapisi v digitalni obliki so priročen vir za izdelavo [parlamentarnih korpusov](https://sidih.github.io/voices/ch3-sl.html), torej strukturiranih zbirk besedil, ki so obogatena z različnimi informacijami. Glede na dostopno gradivo so lahko parlamentarni korpusi tudi večmodalni, če poleg besedila vsebujejo tudi zvočne in/ali video posnetke, npr. [korpus razprav iz češkega parlamenta](https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0005-CF9C-4). Parlamentarni korpusi običajno vsebujejo bogate metapodatke, ki vključujejo raznovrstne podatke o sejah (npr. datum, vrsta seje, dnevni red), govorih in govorcih (npr. ime, datum rojstva, strankarska pripadnost), običajno pa so tudi bogato jezikoslovno označeni (npr. besedne vrste, osnovna oblika besede, imenske entitete). Vse te oznake raziskovalcem omogočajo, da pri oblikovanju raziskovalnih vprašanj upoštevajo tudi različne kombinacije spremenljivk (glejte Pančur in Šorn, 2016).

Ravno zaradi bogatih informacij in kontinuiranosti so [parlamentarni korpusi](https://www.clarin.eu/resource-families/parliamentary-corpora) vir informacij za vrsto področij, ki se pri raziskovanju vse bolj prepletajo, mdr. za jezikoslovje (Bayley, 2004), zgodovino (Piersma idr., 2014), politične vede (Rheault in Cochrane, 2020), demografijo (Kilroy, 2021) ipd. Raziskovalcem so parlamentarni korpusi [dostopni prek konkordančnikov](https://sidih.github.io/voices/ch3-sl.html) (tj. spletnih orodij za raziskovanje in analizo besedil v korpusih) ali pa v [repozitorijih jezikovnih virov](https://www.clarin.si/repository/xmlui/discover?query=parl*&submit=I%C5%A1%C4%8Di&filtertype_2=title&filter_relational_operator_2=contains&filter_2=&query=parl*), kjer je mogoče dostopati do celotnega korpusa v različnih formatih in ga nato obdelati z različnimi orodji. To zadnjo možnost bomo izkoristili tudi v tem učnem gradivu (glejte poglavje 5.2).

## Korpus ParlaMint

V tem učnem gradivu bomo uporabili podatke iz družine korpusov [ParlaMint](http://hdl.handle.net/11356/1432) (Erjavec idr., 2021), ki vsebuje zapise parlamentarnih sej 17 držav: Belgije, Bolgarije, Češke, Danske, Francije, Hrvaške, Islandije, Italije, Latvije, Litve, Madžarske, Nizozemske, Poljske, Romunije, Slovenije, Turčije in Velike Britanije. Večina korpusov ParlaMint pokriva obdobje od 2015 do sredine 2020. Družina korpusov je nastala pod okriljem raziskovalne infrastrukture za jezikovne vire in tehnologije [CLARIN ERIC](https://www.clarin.eu) in vsebuje pol milijarde besed oziroma pet milijonov govorov, ki jih je izreklo približno 11 tisoč govorcev. Korpusi ParlaMint so razdeljeni na dva podkorpusa, in sicer *Reference* (referenčno obdobje) in *COVID,* ki zajemata čas pred epidemijo covida-19 in med njo (tj. pred novembrom 2019 oziroma od novembra 2019 naprej).

Vsak od korpusov za posamezno državo je označen po enotni shemi, ki temelji na smernicah za označevanje [Parla-CLARIN](https://clarin-eric.github.io/parla-clarin/) (Erjavec in Pančur, 2019). To zagotavlja maksimalno primerljivost nacionalnih delov ParlaMinta. Ravno zaradi te značilnosti je ParlaMint dragocen vir za politične, sociološke, zgodovinske, jezikoslovne in druge raziskave, saj omogoča primerjalne analize, ki so bile do zdaj zaradi različne strukture parlamentarnih korpusov težko izvedljive. Poleg tega ParlaMint pokriva širok nabor evropskih držav, kar širi območje vedenja o parlamentarizmu, ki je še vedno zgoščeno predvsem okoli zahodnih držav. Primerjalne raziskave so tako ključnega pomena za napredek pri razumevanju pozitivnih in negativnih parlamentarnih praks in posledično za razvoj parlamentarnih sistemov (Norton, 2002).

V učnem gradivu bomo uporabili slovenski del korpusa ParlaMint. ParlaMint-SI vključuje razprave iz Državnega zbora, ki je osrednja in najvišja predstavniška in zakonodajna institucija v državi. Državni zbor sestavlja 90 poslancev (88 predstavnikov ljudstva in dva predstavnika narodnih manjšin), ki so izvoljeni na podlagi splošne volilne pravice (Brezovšek idr., 2012). ParlaMint-SI pokriva dva parlamentarna mandata v šestletnem obdobju (avgust 2014–julij 2020) in obsega približno 20 milijonov besed (Erjavec idr., 2022). Uporabili bomo [jezikoslovno označeno različico korpusa 2.1](https://www.clarin.si/repository/xmlui/handle/11356/1431) (Erjavec idr., 2021), ki je tudi stavčno segmentirana (s tem so določene meje povedi) in tokenizirana (s tem je pojavnica – beseda, številka, ločilo – določena kot osnovna enota za analizo).

# Tematsko modeliranje besedil

Za analizo parlamentarnih razprav bomo uporabili tematsko modeliranje, eno od tehnik rudarjenja besedil, ki se uporablja pri raziskovanju velikih količin podatkov. Za ustrezno izbiro metode raziskovanja, ki bo dala kakovostne rezultate, in za kritično interpretacijo rezultatov je nujno poznati delovanje ter prednosti in slabosti uporabljenih metod (prim. Shadrova, 2021). Ker bomo v naši analizi kot tehniko tematskega modeliranja uporabili latentno Dirichletovo dodelitev, sledi predstavitev delovanja tehnike in izsledkov nekaterih raziskav, pri katerih je bila le-ta uporabljena na parlamentarnem diskurzu.

Tematsko modeliranje je priljubljena metoda za avtomatsko analizo besedil, na podlagi katere lahko iz korpusa izluščimo prevladujoče teme, ki se pojavljajo v preučevanih besedilih. Rezultat tematskega modeliranja je tematski model, ki ga sestavljajo *teme*, ki pa ne zaznamujejo dejanske tematike, o kateri govori besedilo, ampak so zgolj skupine besed, ki glede na matematično verjetnost sodijo skupaj in naj bi predstavljale isto tematiko. Raziskovalec pa mora nato ročno opredeliti oziroma poimenovati temo, ki naj bi jo te besede predstavljale.

Za tematsko modeliranje korpusa je mogoče uporabiti različne tehnike oziroma algoritme (Vayansky in Kumar, 2020). Med najpogosteje uporabljenimi je tehnika LDA oziroma *latentna Dirichletova dodelitev*, ki jo bomo uporabili tudi v naši analizi. Tehniko so razvili Pritchard idr. (2000), za analizo besedil pa so jo predlagali Blei, Ng in Jordan (2003). Tehnika je izrazito primerna za obdelavo obsežnih zbirk besedilnih podatkov, ki jih zaradi velikosti ne moremo obdelati ročno.

## Tehnika LDA

LDA vključuje naslednje korake, izvedene v več ponovitvah:

1. Algoritem najprej vsem besedam v besedilu naključno dodeli oznako teme.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Beseda**  **Dokument** | EPIDEMIJA | KRIZA | DAVEK | GOSPODARSTVO |
| **dok1** | tema 1 | tema 2 | tema 2 | tema 1 |
| **dok2** | tema 1 | tema 1 | tema 2 | tema 1 |
| **dok3** | tema 2 | tema 1 | tema 1 | tema 2 |
| **dok4** | tema 1 | tema 2 | tema 2 | tema 2 |

1. Algoritem prešteje, kolikokrat se določena tema pojavi v posameznem besedilu (leva tabela) in kolikokrat je določena tema pripisana posamezni besedi (desna tabela).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | |  | **tema 1** | **tema 2** | | **dok1** | 2 | 2 | | **dok2** | 3 | 1 | | **dok3** | 2 | 2 | | **dok4** | 1 | 3 | | |  |  |  | | --- | --- | --- | |  | **tema 1** | **tema 2** | | **epidemija** | 3 | 1 | | **kriza** | 2 | 2 | | **davek** | 1 | 3 | | **gospodarstvo** | 2 | 2 | |

1. Nato algoritem za posamezno besedo privzame, da ne pozna več njene teme.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **epidemija** | **kriza** | **davek** | **gospodarstvo** |
| **dok1** | **?** | tema 2 | tema 2 | tema 1 |
| **dok2** | tema 1 | tema 1 | tema 2 | tema 1 |
| **dok3** | tema 2 | tema 1 | tema 1 | tema 2 |
| **dok4** | tema 1 | tema 2 | tema 2 | tema 2 |

1. Za tem posodobi obe tabeli iz 2. koraka, tako da znova določi frekvenco določene teme v posameznih besedilih (leva tabela) in frekvenco določene besede v posameznih temah (desna tabela).

|  |  |  |
| --- | --- | --- |
|  | **tema 1** | **tema 2** |
| **dok1** | **1** | 2 |
| **dok2** | 3 | 1 |
| **dok3** | 2 | 2 |
| **dok4** | 1 | 3 |

|  |  |  |
| --- | --- | --- |
|  | **tema 1** | **tema 2** |
| **epidemija** | **2** | 1 |
| **kriza** | 2 | 2 |
| **davek** | 1 | 3 |
| **gospodarstvo** | 2 | 2 |

1. Izračuna, kako močna je **povezava med besedilom in temo** (verjetnost teme v dokumentu: moder pravokotnik) ter **temo in izbrano besedo** (verjetnost besede v temi: rdeč pravokotnik).

[Slika0.1\_LDA]

1. Vijolični pravokotnik je zmnožek rdečega in modrega pravokotnika in predstavlja verjetnost, da beseda pripada določeni temi. Na podlagi slednje verjetnosti (oba vijolična kvadrata) nato tehnika določi, kateri temi bo pripadal dokument (zelena zvezda, ki, glede na izračunano verjetnost kombinacije besede in teme, predstavlja naključno dodelitev teme izbranemu dokumentu).

[slika 0.2\_LDA-prob]

Na kratko, algoritem na podlagi verjetnosti (vijolična kvadrata) določi novo temo (zelena zvezda). Verjetnostna porazdelitev tem v besedilu se izračuna na podlagi Dirichletove porazdelitve, ki določa, da verjetnost nikoli ni ničelna. To pomeni, da ima vsaka beseda vsaj majhno možnost, da pripada tudi manj pogosti temi, oziroma da je v dokumentu prisotna tudi manj pogosta tema.  
Ko besedi algoritem določi temo, posodobi tudi tabelo besedila-besede.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **epidemija** | **kriza** | **davek** | **gospodarstvo** |
| **dok1** | **tema 2** | tema 2 | tema 2 | tema 1 |
| **dok2** | tema 1 | tema 1 | tema 2 | tema 1 |
| **dok3** | tema 2 | tema 1 | tema 1 | tema 2 |
| **dok4** | tema 1 | tema 2 | tema 2 | tema 2 |

1. Glede na novo vrednost iz tabele v 6. koraku algoritem znova posodobi obe tabeli, torej tabeli teme-besedila in besede-teme.

|  |  |  |
| --- | --- | --- |
|  | **tema 1** | **tema 2** |
| **dok1** | 1 | **3** |
| **dok2** | 3 | 1 |
| **dok3** | 2 | 2 |
| **dok4** | 1 | 3 |

|  |  |  |
| --- | --- | --- |
|  | **tema 1** | **tema 2** |
| **epidemija** | 2 | **2** |
| **kriza** | 2 | 2 |
| **davek** | 1 | 3 |
| **gospodarstvo** | 2 | 2 |

1. Postopek se ponavlja, dokler se dodelitve tem ne nehajo spreminjati. Končni rezultat je tematski model. Temo predstavlja skupina besed, ki se pogosto pojavljajo skupaj v besedilih.

Zgoraj smo opisali verzijo LDA z Gibbsovim vzorčenjem. Orange namesto Gibbsovega vzorčenja uporablja metodo variacijske inference. Gibbsovo vzorčenje je veliko bolj natančno, medtem ko je variacijska inferenca hitrejša za večje podatke.

## Značilnosti tehnike LDA

LDA temelji na več predpostavkah. Prva je, da temo zaznamujejo besede, ki se v besedilih pogosto pojavljajo skupaj. To pomeni, da je LDA jezikovno neodvisna tehnika, saj besede združuje v skupine na podlagi njihove pojavitve v besedilu in ne na podlagi njihovega pomena. Isto metodo je torej mogoče uporabiti na podatkih v različnih jezikih. Obenem velja, da je lahko posamezna beseda uvrščena v več tem, vendar običajno z različno verjetnostjo.

Druga predpostavka je, da se teme v korpusu pojavljajo različno pogosto in da so medsebojno neodvisne. LDA ne opredeli morebitne povezave med temami, pokaže pa verjetnost razporeditve tem v besedilu, pri čemer opredeli njihovo pomembnost v določenem besedilu. Posamezno besedilo običajno vsebuje več tem, vendar določena tema najpogosteje izstopa (tj. v glavno skupino besed algoritem razvrsti več besed kot v druge skupine besed oziroma teme).

Tretja predpostavka je, da je število tem vnaprej določeno. To pomeni, da mora raziskovalec vnaprej določiti število tem, v katere naj algoritem razvrsti dokumente. Optimalno število tem za določeno besedilo je tisto, pri katerem je za dobljene skupine besed najlažje opredeliti smiselne teme, ki so še vedno informativne glede na zastavljeni raziskovalni problem. Zato raziskovalci običajno postopek tematskega modeliranja za vsako raziskavo posebej večkrat ponovijo z različnim številom tem, pri čemer ocenjujejo povednost skupin besed, ki jih je iz besedila izluščil algoritem. Nekateri si pomagajo tudi z dodatnimi statističnimi izračuni (prim. Smith in Graham, 2019). Čeprav se torej ustrezno število tem razlikuje od primera do primera, običajno število tem sega od 5 do 50 (Arun idr., 2010), veliko raziskav pa se odloči za 20 tem (Zhao idr., 2015; Gkoumas idr., 2018; Rosa idr., 2021).

Četrta predpostavka je, da vrstni red besed v besedilu ni pomemben. LDA namreč deluje na podlagi t. i. vreče besed, ki ne upošteva jezikovne strukture in povezav med besedami. Ta predpostavka je problematična, saj je vrstni red besed ena ključnih značilnosti realnih besedil.

Na rezultate pa ključno vpliva vrstni red besedil, saj tehnika v začetku besedilom naključno dodeli teme. Če se spremeni vrstni red, se bo tudi začetna dodelitev spremenila. Vrstni red besedil v smislu njihovega časovnega zaporedja prav tako lahko pomembno vpliva na značilnosti besedil (Vayansky in Kumar, 2020).

## Predobdelava podatkov

Za potrebe tematskega modeliranja morajo biti podatki ustrezno pripravljeni. To največkrat vključuje tokenizacijo (členitev na pojavnice, običajno besede, števila, ločila), lematizacijo (pripis osnovne oblike vsaki pojavnici) in oblikoskladenjsko označevanje (pripis besedne vrste vsaki pojavnici, npr. glagol). Slednje nam omogoča, da tematsko modeliranje izvedemo zgolj na določeni besedni vrsti. Raziskave namreč kažejo, da lematizacija in omejitev korpusa zgolj na samostalnike izboljšata tako hitrost algoritma kot njegove rezultate, in sicer z vidika koherence oziroma smiselne povezanosti besed v posameznih skupinah, zaradi česar jim je lažje pripisati temo (Martin in Johnson, 2015). Razlikovanje po oblikoskladnji lahko uporabimo tudi za iskanje odgovorov na različna raziskovalna vprašanja. Tako so na primer Van der Zwaan idr. (2016) tematsko modeliranje najprej izvedli na samostalnikih in tako pridobili teme, nato pa algoritem pognali še na glagolih, pridevnikih in prislovih, na podlagi česar so nato izluščili stališča parlamentarcev. Pred uporabo metode LDA iz korpusa običajno odstranimo tudi zelo pogoste besede, ki so lahko preveč splošne (npr. zaimki, predlogi) ali preveč specifične za preučevani žanr (npr. pozdravne besede, kot je *spoštovani*, v parlamentarnem korpusu), glede na izbrani raziskovalni problem pa lahko odstranimo tudi zelo redke besede, ločila, veliko začetnico ipd. (Smith in Graham, 2019).

## Omejitve metode LDA

Kot prvo omejitev lahko izpostavimo dejstvo, da LDA za uspešno delovanje potrebuje daljša besedila, saj temelji na porazdelitvah besed, ki za kratka besedila niso izrazite. Tako na primer LDA ni primerna tehnika za tematsko modeliranje tvitov, uporabniških komentarjev ali poezije. Čeprav jo je mogoče uporabiti tudi za analizo npr. objav na Facebooku (Serrano idr., 2019), se za tovrstna besedila priporočajo druge metode tematskega modeliranja (Albalawi idr., 2020; Morstatter idr., 2018). Čeprav so parlamentarni govori običajno dovolj dolgi za učinkovito uporabo metode LDA, pa lahko nekatere seje vključujejo tudi zelo kratke govore, ki jih je pred uporabo metode LDA dobro odstraniti (glejte poglavje 5.4).

Druga omejitev je povezana s predpostavko, da so besede medsebojno neodvisne, prav tako pa tudi besedila in teme, kar je po eni strani jezikoslovno netočno, po drugi strani pa raziskovalcem onemogoča pregled nad korelacijami med besedami oziroma med besedili. Ta informacija pri mnogih raziskovalnih problemih sicer ni potrebna, zato tehnika LDA zadostuje. Če pa je ta informacija pomembna (npr. če nas zanima razvoj teme skozi daljše časovno obdobje), so na voljo druge metode tematskega modeliranja, npr. dinamično tematsko modeliranje (Müller-Hansen idr., 2021). Poleg tega LDA ne deluje dobro, če besedilo o neki temi ne govori koherentno, ampak jo omeni zgolj s par besedami. Po drugi strani pa tehnika dobro deluje za daljša, tematsko opredeljena besedila, na primer za novice, akademske članke, politične govore in določena prozna dela.

Naslednja omejitev je vezana na število tem, ki ga mora raziskovalec določiti po lastni presoji in je običajno rezultat preizkušanja. Kot opozarjata Allen in Murdock (2020), bodo pri zelo velikem številu nekatere teme predstavljale zgolj zelo specifičen del enega od besedil v korpusu, zaradi česar je težje iskati tematske povezave med besedili. Kadar pa je število tem postavljeno zelo nizko, jih je običajno lažje interpretirati, vendar so najpogosteje teme zelo splošne in tako precej neinformativne za analizo.

Kot zadnjo omejitev lahko omenimo težavnost interpretacije rezultatov tematskega modeliranja, vključno z načinom navajanja rezultatov v publikacijah. Ker so rezultat izolirane skupine besed, obstaja nevarnost, da raziskovalci v njih vidijo vzorec, ki ga pravzaprav ni, oziroma prepoznajo teme, ki so jih že v izhodišču pričakovali (Shadrova, 2021). Pri interpretaciji je zato pomembno, kakšno število besed upošteva raziskovalec, ko opredeljuje temo. Ta je lahko namreč precej drugačna, če raziskovalec sklepa o temi na podlagi prvih 10 ali prvih 30 besed, ki jih ponudi algoritem (Allen in Murdock, 2020). Pri interpretaciji skupin besed in opredeljevanju ter interpretaciji tem je zato nujno potrebno kvalitativno branje oziroma poznavanje izvirnih delov besedila, v katerih se izbrane besede pojavljajo, še posebej za skupinsko delo pa je priporočljivo oblikovati tudi smernice za odločanje.

Tematsko modeliranje po metodi LDA torej ni hitra rešitev, ki bi raziskovalca lahko brez kritične presoje rezultatov privedla do kakovostnih zaključkov. Prav razumevanje omejitev je bistvenega pomena za uspešno uporabo tematskega modeliranja v raziskovalne namene. Kvantitativne, avtomatske metode analize torej lahko pomembno dopolnjujejo raziskovalčeve sposobnosti za analizo, vendar ne morejo nadomestiti človeške interpretacije (Grimmer in Stewart, 2013). Tematsko modeliranje omogoča pomembno prednost pred kvalitativnim branjem, in sicer možnost obdelave velike količine podatkov, kar raziskovalcem omogoča robustnejše posploševanje na podlagi rezultatov kot pri analizi, izvedeni na majhnem vzorcu (Jacobs in Tschötschel, 2019). Poleg tega postopek tematskega modeliranja omogoča večjo objektivnost rezultatov (Müller-Hansen idr., 2021), čeprav zaradi prej omenjenih omejitev, kot je opredelitev tem na podlagi skupine besed, tehnika ni v celoti objektivna. To je obenem tudi prednost te tehnike, saj algoritem ne poda neposrednih odgovorov, ampak raziskovalce sili v upoštevanje konteksta pri oblikovanju končnih rezultatov (Schmidt, 2012). Tematsko modeliranje vnaša tudi večjo sistematizacijo analize in omogoča sorazmerno večjo ponovljivost rezultatov (Jacobs in Tschötschel, 2019). Priljubljenost in aktualnost te tehnike za humanistične in družboslovne raziskave se kažeta tudi v velikem številu raziskav, ki tematsko modeliranje uporabljajo v svoji metodologiji (gl. poglavje 4.5).

## Tematsko modeliranje parlamentarnih razprav

V politologiji in drugih humanističnih in družboslovnih vedah, ki analizirajo velike količine realnih podatkov, se tematsko modeliranje vse bolj uveljavlja kot pomembna analitična tehnika, ki dopolnjuje ustaljene, kvalitativne pristope za analizo. Rezultati tematskega modeliranja namreč lahko služijo kot usmeritev za kvalitativno analizo (npr. identifikacijo relevantnih besedil o najbolj izstopajoči temi) ali pa že predstavljajo del končnih rezultatov. V tem poglavju predstavljamo nekaj primerov uporabe tematskega modeliranja parlamentarnih podatkov.

S tematskim modeliranjem lahko **izluščimo vsebinske tematike**, o katerih se razpravlja v parlamentu. Schuler (2020) je na primer z metodo LDA analiziral razprave v vietnamskem parlamentu in rezultate povezal s temami iz novic, tudi pridobljenimi z LDA, ter seznamom področij pod neposrednim upravljanjem Komunistične partije (KP). S tem je raziskal, ali poslanci v avtoritarnih sistemih, kot je vietnamski, izražajo svoja stališča in aktivno razpravljajo glede pomembnih tem. Ugotovil je, da do dejanske razprave pride zgolj pri temah, ki ne pokrivajo področij, ki so v neposrednem upravljanju KP, pri čemer slednja spodbuja takšno razpravo zato, da lahko pritiska na vlado in ji obenem naprti odgovornost za posledice sprejetih ukrepov. Kadar pa tema razprave zadeva eno od področij, ki jo s svojimi odbori neposredno upravlja KP, slednja ne spodbuja javne razprave. Obenem je zanimiva njegova ugotovitev, da se tudi pri temah, ki so odprte za izmenjavo stališč, v razpravo ne vključujejo vsi poslanci, ampak predvsem tisti, ki niso člani KP in so bili za polni delovni čas kot predstavniki ljudstva izvoljeni na lokalni ravni.

Za **odkrivanje vsebinskih tematik** v parlamentarnih razpravah sta tematsko modeliranje uporabila tudi Chizhik in Sergeyev (2021), ki sta analizirala približno tri desetletja govorov ruskih poslancev, da bi ugotovila, ali obstajajo povezave med dejavnostjo strank v parlamentu in skepticizmom javnosti glede večstrankarskega sistema kot temelja za demokracijo. Ugotovila sta, da stranke ruskega parlamenta svojo pozornost predvsem posvečajo zunanjim zadevam, gospodarstvu in razmerju moči med vejami oblasti, medtem ko se precej manj posvečajo drugim družbenim problemom. Pri tem njihove govore, še posebej to velja za uveljavljene stranke, v veliki meri zaznamuje ideološki in propagandistični diskurz.

Tematsko modeliranje torej omogoča, da pridobimo splošen pregled nad gradivom, ki lahko zadošča, če je naš namen, na primer, opazovati izpostavljenost tematik, o katerih se razpravlja v parlamentu. Tematsko modeliranje pa lahko uporabimo tudi za pridobivanje bolj specifičnih rezultatov. Ker so parlamentarni korpusi običajno bogato označeni z metapodatki, lahko **teme raziskujemo v razmerju do drugih spremenljivk** (npr. spol, starost, pripadnost politični stranki, mandat itd.) in tako izluščimo tiste teme, ki najbolj razlikujejo med vrednostmi izbrane spremenljivke. Tako lahko na primer opazujemo, kako aktualna je bila določena tema skozi čas ali pri določeni stranki. Curran idr. (2018) so na primer z metodo LDA in analizo kompleksnih mrež razkrili teme razprav v novozelandskem parlamentu in jih povezali s poslanci ter njihovimi strankami, ki so te teme naslavljali v svojih govorih. Na ta način so izluščili prevladujoče teme razprav v posameznih obdobjih in jih interpretirali v povezavi z zunanjimi dogodki (npr. potres leta 2011). Poleg tega so na podlagi rezultatov opredelili, koliko zanimanja je določeni temi namenila posamezna stranka. Izkazalo se je, da je Laburistična stranka veliko bolj zavzeto razpravljala o nepremičninski krizi kot takrat vladajoča Nacionalna stranka. Rezultati so pokazali tudi, da je Nacionalna stranka zavzela glavnino razprave, medtem ko se je prispevek manjših strank skozi čas zmanjševal. Zmanjšala pa se je tudi specializacija parlamentarcev za različne teme, kar se kaže v velikem številu tem, ki jih naslavlja večina poslancev.

Ena od možnosti **uporabe metapodatkov pri tematskem modeliranju** je predstavljena tudi v raziskavi, ki so jo izvedli de Campos idr. (2021). Metodo LDA so namreč uporabili za določanje tematskega profila španskih poslancev, ki odraža področja, ki jih ti pokrivajo v parlamentu. Metapodatke sta s pridom uporabila tudi Høyland in Søyland (2019), ki sta želela ugotoviti, ali je sprememba volilnega sistema leta 1919, ko je norveški politični sistem postal bolj odvisen od strankarske politike in poslanci niso imeli več toliko avtonomije kot pred reformo, vplivala na tematiko parlamentarnih razprav. Pri tem sta uporabila različico metode LDA, in sicer strukturno tematsko modeliranje (STM), ki poleg porazdelitve besed pri izračunu rezultatov sočasno upošteva tudi izbrane metapodatke. Ugotovila sta, da razporeditev tem jasno kaže, da institucionalna ureditev vpliva na delovanje parlamentarcev. Po reformi, ki je poudarila pomen strankarske politike, so postale teme, ki so izrazito kazale ideološke razlike med strankami, pogostejše, medtem ko so poslanci manj govorov namenjali neposredni kritiki drugih poslancev. Obenem so poslanci po reformi pogosteje razpravljali o temah splošnega zanimanja (npr. o izobraževalnem sistemu) in redkeje o temah, ki so neposredno obravnavale težave njihovih volilnih okrajev (npr. o izgradnji infrastrukture v nekem oddaljenem kraju).

S tematskim modeliranjem lahko tudi **raziščemo kontekst in nabor tem, s katerimi je povezan določen pojem.** Müller-Hansen idr. (2021) so uporabili različico tehnike LDA, imenovano dinamično tematsko modeliranje(DTM), ki omogoča analizo tem skozi čas. S to metodo so analizirali sedemdeset let nemških parlamentarnih razprav o premogu, in tako raziskali, kako so se razprave o premogu spreminjale skozi čas[[4]](#footnote-4). Razprave iz zgodnjih let korpusa kažejo, da so poslanci premog dojemali kot gonilo gospodarskega razvoja in jamstvo energetske varnosti, medtem ko v zadnjih letih razprave o premogu predvsem govorijo o energetskem prehodu, postopni opustitvi premoga in razmahu obnovljivih virov energije. Obenem so raziskovalci ugotovili, da manjše in mlajše stranke (npr. Zeleni) v nemškem parlamentu pogosteje kot druge stranke govorijo o premogu v kontekstu energetskega prehoda in varovanja okolja.

Namesto izbranega pojma lahko na podlagi rezultatov tematskega modeliranja **opazujemo kontekst določene teme oziroma prepletanje tem.** Blätte idr. (2020) so, na primer, želeli izvedeti, kako pogosto se o migracijah razpravlja v kontekstu skupne evropske politike. Z metodo LDA so ustvarili tematski model za parlamentarne razprave Avstrije, Francije, Nemčije in Nizozemske, izbrali tri teme, ki so najbolje predstavljale migracije oziroma evropske zadeve, nato pa izluščili vse govore, kjer sta se ti dve krovni temi križali. Raziskava je pokazala, da so bile razprave o migrantih v dveh večjih državah iz korpusa (Francija, Nemčija) bolj usmerjene navznoter. Še posebej v Nemčiji je evropski vidik skoraj izginil, medtem ko se je v manjših državah iz vzorca (Avstrija, Nizozemska) delež govorov, ki o migracijah razpravljajo z vidika evropskega odziva, povečal.

V tem učnem gradivu se delno naslanjamo na metodologijo, ki so jo uporabili Curran in sodelavci (2018) v zgoraj omenjeni raziskavi novozelandskih govorov, pri čemer v naši analizi obravnavamo krajše časovno obdobje, zato podrobnejša delitev po letih ni smiselna. Kot je razvidno iz zgoraj opisanih raziskav, bi lahko analizo nadgradili s strukturnim tematskim modeliranjem, pri čemer bi upoštevali, na primer, strankarsko pripadnost govorcev in opazovali razlike med njimi. Lahko pa bi analizirali tudi celoten korpus ParlaMint-SI in z dinamičnim tematskim modeliranjem opazovali razlike tem skozi čas. Vendar je za primerjavo tem v dveh obdobji, pred in med epidemijo, tehnika LDA povsem ustrezna.

# Priprava na analizo

S tem poglavjem se začne praktični del učnega gradiva. Poglavje nas bo vodilo vse od namestitve programa ter uvoza in kontrolnega pregleda osnovnih podatkov do izdelave vzorca za analizo in predprocesiranja podatkov.

Za prenos in namestitev programa Orange, vključno s programom Miniconda, ki je del namestitvenega paketa za Orange, boste potrebovali približno 1 GB prostora na disku. Za prenos datotek ParlaMint boste potrebovali 2,3 GB prostora. Upoštevajte, da je tematsko modeliranje računsko zahteven proces, zato bo morda delovanje računalnika upočasnjeno, sploh če uporabljate napravo z malo delovnega pomnilnika (RAM).

## Namestitev in uporaba programa Orange

Analizo bomo izvedli v odprtokodnem in prostodostopnem programu [Orange](https://orangedatamining.com/) v3.32.0, ki temelji na programskem jeziku Python (Demšar idr., 2013). Program je namenjen podatkovni analitiki, razširitev *Text*  (v1.10.0) pa ponuja poseben nabor orodij za rudarjenje besedil. Orange je osnovan na vizualnem programiranju, kar pomeni, da potek analize določimo tako, da korake analize oziroma tako imenovane gradnike sestavljamo v analitski delotok.

Program najprej s spletne strani [orangedatamining.com](https://orangedatamining.com/) prenesemo na računalnik. Preneseno programsko datoteko odpremo in sledimo navodilom za namestitev. Nato program odpremo in naložimo še razširitev *Text*. To storimo tako, da izberemo zavihek *Options* in v spustnem meniju kliknemo *Add-ons.* Odpre se okno, kjer obkljukamo polje *Text* in namestitev razširitve potrdimo s klikom gumba OK (Slika 1).

Slika 1: Namestitev razširitve Text.

Po namestitvi moramo znova zagnati program[[5]](#footnote-5). Ko se program odpre, se v levem meniju prikaže zavihek *Text Mining*, ki vsebuje različne gradnike (npr. *Corpus, Bag of Words*), ki so namenjeni analizi besedil (Slika 2). Na desni strani je belo polje, ki ga imenujemo platno (*canvas*). Nanj zlagamo gradnike (*widgets*), ki jih povezujemo v analitski delotok.

Slika 2: Začetni zaslon ob zagonu programa po namestitvi razširitve Text.

Gradnike na platno dodajamo tako, da jih povlečemo iz menija na levi in spustimo na platno, ali pa z desnim klikom na platno odpremo spustni meni, vtipkamo ime gradnika, npr. **Corpus**, in ga s tipko **Enter** dodamo na platno. Če gradnik dvokliknemo, se odpre okno z nastavitvami. Vsak gradnik ima vhod ali izhod oziroma oboje, kar je označeno s črtkano črto ob strani gradnika. Vhod v gradnik je na levi, izhod pa na desni strani. Analiza v Orangeu vedno poteka od leve proti desni, nikoli obratno.

Z enakim postopkom, kot smo dodali gradnik **Corpus**, dodajmo še gradnik **Corpus Viewer**. Če ga dvokliknemo, se odpre prazno okno. Gradnik namreč še ni prejel podatkov za obdelavo. Podatke pošljemo v gradnik tako, da z miško povežemo gradnika med sabo od desne črtkane črte gradnika **Corpus** do leve črtkane črte gradnika **Corpus Viewer** (Slika 3). Ta gradnik z dvoklikom znova odpremo in tokrat pojavno okno prikazuje podatke.[[6]](#footnote-6)

Slika 3: Povezava med gradnikoma omogoča prehod podatkov od izhoda prvega gradnika do vhoda naslednjega gradnika.

V učnem gradivu bomo sestavili delotok, s katerim bomo izvedli tematsko modeliranje parlamentarnih govorov in teme dodatno raziskali z vizualizacijami. Celoten delotok si lahko [prenesete](https://www2.sistory.si/publikacije/material/parlamint/tutorial-slo.ows), vendar priporočamo, da sledite postopnim korakom v gradivu in sosledje gradnikov sestavite sami, saj boste tako najbolje razumeli posamezne faze analize.

## Prenos in uvoz podatkov v Orange

Korpus ParlaMint-SI obsega parlamentarne govore med leti 2014 in 2020. V praktičnem delu nas bo zanimala primerjava govorov tik pred in med epidemijo COVID-19, zato podatke najprej zamejimo na primerljivo dolgi obdobji pred in med epidemijo. Ker epidemično obdobje, ki je zajeto v korpusu ParlaMint-SI, vključuje devet mesecev (november 2019–julij 2020, Erjavec idr. 2022), bomo uporabili tudi podobno dolgo obdobje tik pred začetkom pandemije, in sicer od januarja 2019 do oktobra 2019.[[7]](#footnote-7)

Pri analizi bomo uporabili slovenski del jezikoslovno označenega korpusa parlamentarnih podatkov ParlaMint 2.1 (gl. poglavje 3.3). Parlamentarni govori s pripisanimi jezikoslovnimi oznakami so zapisani v [formatu CoNLL-U](https://universaldependencies.org/format.html). Da bomo lažje začeli, so podatki za potrebe tega učnega gradiva že pripravljeni in si jih [prenesemo](https://www2.sistory.si/publikacije/material/parlamint/ParlaMint-SI-conllu.zip) na računalnik.[[8]](#footnote-8),[[9]](#footnote-9)

Na platno dodamo gradnik **Import Documents.** Z dvoklikom odpremo okno za nastavitve in v prvi vrstici določimo mapo, v katero smo shranili podatke (Slika 4). Uvoza ni treba posebej potrjevati, zgodi se samodejno, ko določimo mapo. Malo nižje obkljukamo tudi možnosti *Lemma* in *POS tags*. Tako bomo poleg govorov uvozili tudi leme in oblikoskladenjske oznake. Govor posameznega poslanca za posamezno sejo bo predstavljen kot posamezen dokument. V spodnjem delu okna nas program obvesti, da smo uvozili 18.476 dokumentov oziroma govorov.

Slika 4: Okno za uvoz podatkov.

Za boljše razumevanje strukture podatkov si na kratko oglejmo značilnosti formata CoNLL-U. Gre za obliko formata TSV, v katerem so vrednosti ločene s tabulatorjem in ki se na področju obdelave naravnega jezika uporablja za zapisovanje jezikoslovno označenih besedil, saj zaradi razporeditve besedila in oznak po stolpcih omogoča enostavno računalniško obdelavo. V tem formatu vsaka poved predstavlja svoj sklop, besedilo pa je tudi vertikalizirano, kar pomeni, da so besede zapisane ena pod drugo, kar omogoča jasen pregled nad pripisanimi jezikoslovnimi oznakami. Vsaka poved ima na začetku navedene tudi metapodatke (npr. ID govora, ID povedi in besedilo) (Slika 5).

Slika 5: Zapis podatkov v formatu CoNLL-U za prvi dve povedi iz dokumenta ParlaMint-SI\_2019-01-28-SDZ8-Redna-04.conllu.

## Pregled podatkov

[Workflow1]

Preden začnemo z analizo, se je dobro najprej prepričati, da so naloženi podatki res pravi. To bomo storili z gradnikom **Corpus Viewer**, ki ga dodamo na platno in ga od leve proti desni povežemo s prejšnjim gradnikom. **Corpus Viewer** odpremo z dvoklikom in prikaže se seznam dokumentov, ki v našem primeru predstavljajo posamezne govore (Slika 6). S klikom na seznam si lahko ogledamo tudi druge govore, če pa med klikanjem pridržimo tipko *Shift,* lahko prikažemo več govorov hkrati.

Slika 6: Pregled govorov v gradniku Corpus Viewer.

V levem zgornjem kotu vidimo osnovne informacije o korpusu: število pojavnic (*tokens*) in različnic (*types*)[[10]](#footnote-10) ter število govorov (*matching documents*), ki ustrezajo iskalnemu filtru (*regexp filter*), če bi ga uporabili. Ker je filter trenutno prazen, so prikazani vsi govori (18476/18476). Zadnji podatek (*matches*) je število zadetkov iskalnega izraza, ki bi ga lahko vnesli v polje *RegExp Filter*.[[11]](#footnote-11)

V pregledovalniku na desni strani vidimo številne metapodatke[[12]](#footnote-12) o govorih in govorcih, s katerimi je opisana vsebina korpusa.[[13]](#footnote-13) Vsak govor je označen z imenom seje, ki ji pripada (*name*), in unikatno oznako (*utterance*), pri čemer zadnja številka izkazuje zaporedno številko govora v dotični seji. Celoten govor si lahko preberemo pod spremenljivko *content*. Pomemben je tudi podatek o podkorpusu (*Subcorpus*), ki označuje časovno obdobje, v katerem je bil govor podan (*Reference* označuje govore pred novembrom 2019, *COVID* pa od novembra 2019).

Sledijo podatki o govorcu – njegova vloga (*Speaker role*), ki je lahko ali predsedujoči ali običajni govorec, funkcija (*Speaker type*), ki je lahko ali poslanec ali gost, pripadnost politični stranki (*Speaker party*), parlamentarni status stranke (*Party status*), ki je lahko opozicija ali koalicija, ime govorca (*Speaker name*), spol (*Speaker gender*) in leto rojstva (*Speaker birth*).

## Priprava vzorca in predprocesiranje

Kot smo videli v 3. poglavju, za parlamentarni diskurz velja jasna struktura, ki vključuje številne značilne fraze, npr. *besedo ima poslanec …, hvala za besedo, spoštovana gospa ministrica, potrjujem dnevni red* ipd.Nekatere fraze, ki urejajo potek razprave, so še posebej značilne za predsedujoče parlamentarnim sejam, druge so zgolj del vljudnostnega izražanja. Zaradi svoje narave so take fraze v parlamentarnih korpusih zelo pogoste, vendar za tematsko analizo niso zanimive, temveč bi predstavljale zgolj šum v rezultatih, zato jih želimo izločiti iz vzorca. V celoti jih je sicer nemogoče izločiti na enostaven, avtomatiziran način, vseeno pa lahko njihovo število pomembno omejimo. To bomo storili tako, da bomo pripravili vzorec podatkov, iz katerega bomo izločili vse govore predsedujočih in govore, ki so krajši od 50 besed (Slika 8).

»Hvala lepa. V skladu s poslovnikom bi želel v imenu poslanske skupine 30 minut odmora pred glasovanjem« (Brane Golubović, redna seja, 6. marec 2019).

To odločitev smo sprejeli po ročnem pregledu dela korpusa, v katerem se je izkazalo, da so taki govori večinoma zgolj zahvale ali medklici. Podobno filtriranje uporabljajo tudi sorodne raziskave (Curran idr., 2018). Čeprav taki govori seveda niso nujno le proceduralne narave, so, zato ker so kratki, v vsakem primeru manj primerni za tematsko modeliranje z metodo LDA, ki za dobre rezultate potrebuje daljša besedila (glejte 4. poglavje). Poleg tega bomo iz vzorca za analizo izločili tudi vse govore posameznikov in posameznic, ki v parlamentu lahko nastopajo kot gosti.

### **Odstranjevanje neželenih govorov**

[Workflow2]

Za oblikovanje vzorca potrebujemo gradnika **Statistics** in **Select Rows**. Na platno dodamo gradnik **Statistics** in ga povežemo z gradnikom **Import Documents**, pri čemer ohranimo prednastavitve, na podlagi katerih gradnik prešteje število besed (*word count)* in število znakov (*character count*) v govorih.Gradnik **Statistics** podatkom doda stolpca s številom besed in znakov, kar si lahko ogledamo tako, da dodamo gradnik **Data Table,** ga odpremo in se pomaknemo skrajno desno(Slika 7). Tako dobimo podatek o dolžini govorov, kar nam omogoča, da zdaj z gradnikom **Select Rows** izberemo zgolj tiste, ki ustrezajo našim merilom glede dolžine.

Slika 7: Stolpca skrajno desno prikazujeta podatke o številu besed in znakov, ki so dodani osnovnim podatkom.

Ko dodamo gradnik **Select Rows** (Slika 8), ga povežemo z gradnikom **Statistics**, odpremo in nastavimo tri pogoje (pogoje dodajamo z gumbom *Add condition*):

* s prvim pogojem določimo najmanjšo dolžino govora – spremenljivko *Word count* omejimo z možnostjo *is greater than* in vpišemo želeno minimalno dolžino, v našem primeru 50, s čimer bomo v vzorec vključili zgolj govore, ki imajo 51 besed ali več;
* z drugim pogojem iz vzorca izločimo predsedujoče sejam – spremenljivko *Speaker role* določimo s parametrom *is* in parametrom *Regular*;
* s tretjim pogojem v vzorcu ohranimo zgolj govore poslancev – spremenljivko *Speaker type* določimo s parametrom *is* in parametrom *MP*, s čimer izločimo govore gostov.

Slika 8: Izbor govorov z več kot petdeset besedami, ki so jih izrekli redni poslanci.

Na spodnjem robu gradnika **Select Rows** vidimo, da se je število govorov skrčilo na 6861 (s prejšnjih 18.476; podrobne podatke dobimo, če kliknemo številke na spodnjem robu).

### **Odstranjevanje neželenih besed**

[Workflow3]

Za dobre rezultate tematskega modeliranja je pomembna tudi predobdelava podatkov (glejte poglavje 4.3.). To storimo s filtriranjem v gradniku **Preprocess Text**, ki ga dodamo na platno, vendar ga še ne povežemo. Najprej bomo nastavili vse parametre, saj se bo tako postopek izvedel bolj gladko. Kljub temu upoštevajte, da je ta korak računsko zahteven in zato lahko tudi precej zamuden. Ko torej odpremo gradnik **Preprocess Text**, se prikažejo privzeti izbrani koraki za predobdelavo besedila, katerih vrstni red in nastavitve je mogoče poljubno spreminjati. Naši podatki so že tokenizirani in besede transformirane v zapis z malo začetnico (prim. poglavje 5.3). Zato koraka *Transformation* in *Tokenization* odstranimo s klikom križca levo zgoraj (Mac OS) oziroma desno zgoraj (Windows).[[14]](#footnote-14) Ostane korak *Filtering*, pri katerem bomo spremenili nekaj nastavitev (Slika 9):

* izključimo možnost *Stopwords* – te možnosti, ki je namenjena odstranitvi nepolnopomenskih besed, kot so zaimki, vezniki, predlogi ipd., ne potrebujemo, ker se bomo osredotočili zgolj na samostalnike;
* obkljukamo možnost *Document frequency* in kot mero izberemo absolutno vrednost (*Absolute)*, pri kateri nastavimo razpon od 10 do 6861 (število vseh govorov) – tako bomo pri analizi prezrli besede, ki se pojavijo v manj kot desetih govorih, kar pomeni, da bomo iz analize izločili zelo redke besede, ki ne vplivajo na oblikovanje specifičnih tem;
* obkljukamo tudi možnost *POS tags*, kar lahko storimo, ker so naši podatki že oblikoskladenjsko označeni (glejte poglavje 3.3) – ta možnost je privzeto nastavljena tako, da pri poznejših analizah upošteva zgolj samostalnike (*noun*) in glagole (*verb*), a ker so se samostalniki izkazali kot najkoristnejši pri tematskem modeliranju (Martin in Johnson, 2015), za analizo ohranimo zgolj to besedno vrsto, glagole (verb) pa izbrišemo.

Sedaj **Preprocess Text** povežemo z gradnikom **Select Rows**.

Slika 9: Nastavitve gradnika Preprocess Text.

Najpogostejše besede našega vzorca lahko preverimo v oblaku besed. Izdelamo ga z gradnikom **Word Cloud**, ki ga pripnemo na gradnik **Preprocess Text.** V oblaku najdemo zgolj samostalnike, pri čemer je velikost besede v oblaku sorazmerna z njeno frekvenco (Slika 10). Prikazane besede zelo jasno odražajo parlamentarni žanr besedil. Na seznamu na levi vidimo, da je najpogostejša beseda *zakon*, ki se v korpusu pojavi 19.362-krat. Ker je parlament glavno zakonodajno telo v državi, tak rezultat seveda ni presenetljiv.

Slika 10: Najpogostejše besede oblikovanega vzorca podatkov po predprocesiranju.

Predprocesiranje je pomemben del obdelave besedilnih podatkov, vendar je potrebno vsak korak jasno utemeljiti. Vsaka odločitev namreč vpliva na končne rezultate, kar je potrebno upoštevati tudi pri interpretaciji. Omenimo še, da v Orangeu filtriranje ne spreminja izhodiščnih podatkov. Po predprocesiranju imamo torej še vedno na voljo izhodiščni korpus iz gradnika **Select Rows** ter dodano informacijo o pojavnicah iz gradnika **Preprocess Text**.

# Analiza parlamentarnih govorov

To poglavje je razdeljeno na tri praktične naloge, v katerih s tematskim modeliranjem in nekaterimi vizualizacijami raziščemo vsebino parlamentarnih razprav pred in med epidemijo. Pri tem bomo odgovorili na naslednja vprašanja:

* v 1. nalogi: Katere teme zaznamujejo naše podatke?
* v 2. nalogi: O katerih temah so poslanci največ razpravljali?
* v 3. nalogi: Katere teme so izstopale med epidemijo v primerjavi s predepidemičnim obdobjem?

## Teme parlamentarnih razprav

V tem poglavju bomo za vzorec podatkov najprej pripravili številski opis, ki je potreben za izvedbo metode LDA, nato bomo teme izluščili in jih poimenovali, na koncu pa bomo preverili še, kako so teme razporejene po dokumentih in kako lahko poiščemo govore na izbrano temo.

### **Ustvarjanje besedilnih vektorjev**

[Workflow4]

Preden lahko izvedemo tematsko modeliranje, potrebujemo predprocesirane podatke in vektorsko reprezentacijo govorov. Predprocesiranje podatkov smo že opravili (glejte poglavje 5.4), vektorsko reprezentacijo pa izvedemo z gradnikom **Bag of Words**, ki naredi številski opis govorov, na podlagi katerega je nato mogoče izračunati porazdelitev besed po temah oziroma izvesti tematsko modeliranje govorov. Številski opis, ki ga pridobimo s tehniko vreče besed, predstavljajo besede v stolpcih, vrednosti pa izkazujejo število pojavitev posamezne besede v danem govoru. Vsak govor je tako predstavljen s svojim vektorjem, ki predstavlja njegovo vsebino. Pogostejša kot je neka beseda v govoru, bolj vektor tega govora kaže v smeri te besede.

Vendar pa niso vse besede enakovredne. Nekatere besede v korpusu so lahko izrazito proceduralne in žanrsko specifične (glejte poglavje 5.4), nepolnomenske (npr. zaimki, vezniki) ali pa preprosto nespecifične za izbrani govor. Beseda *hvala* se, na primer, pojavi v tematsko zelo heterogenih govorih, saj se večina poslancev zahvali za predajo besede, zato tematsko gledano ni informativna. Besede v našem vzorcu zato želimo obtežiti tako, da bodo imele večjo težo take, ki so izrazito specifične za posamezni govor, nižjo težo pa take, ki se pogosto pojavljajo v vseh govorih. Taka obtežitev se imenuje TF-IDF, kar je angleška kratica za *term frequency-inverse document frequency* oziroma *frekvenca izraza–inverzna frekvenca dokumenta* (Jones, 1972), in je na voljo kot izbirna možnost v gradniku **Bag of Words**.

Gradnik **Bag of Words** dodamo takoj za gradnikom **Preprocess Text** in ga nastavimo tako, da ohranimo možnost preprostega štetja besed (*Count*) v prvi vrstici (*Term Frequency*), v drugi vrstici (*Document Frequency*) pa izberemo IDF (Slika 11).

Slika : Nastavitve gradnika Bag of Words.

### **Luščenje tem**

[Workflow5]

Sedaj, ko smo pripravili vektorsko reprezentacijo govorov, lahko začnemo z luščenjem tematik. To bomo storili z gradnikom **Topic Modelling**, ki ga povežemo z **Bag of Words** in nastavimo tako, da izberemo metodo LDA (*Latent Dirichlet Allocation*)in določimo 20 tem (*Number of topics*). Izbira števila tem je sicer poljubna, vendar sorodne raziskave kažejo, da se pri večjih korpusih najbolje obnese 20 tem (glejte 4. poglavje). Na desni se izpiše dvajset skupin besed, ki zaznamujejo teme našega vzorca (Slika 12). Nekatere teme je enostavno poimenovati že na prvi pogled, medtem ko pri drugih krovna tematika ni povsem jasna. V naslednjem koraku bomo zato izvedli dodatno analizo, na podlagi katere bomo lažje opredelili teme.

Slika : Rezultat tematskega modeliranja s tehniko LDA za dvajset tem.

Pri tem koraku velja omeniti še, da je tehnika LDA stohastična, kar pomeni, da ob vsakem izračunu vrne drugačne rezultate, saj temelji na naključnem določanju izhodiščnih tem. To značilnost metode, ki omejuje ponovljivost raziskav, v programu Orange zaobidemo z uporabo enotne začetne točke, kar omogoča, da uporabniki za iste podatke dobijo enake rezultate.

Poskusite sami: uporabite različna števila tem, npr. 5 in 10. Ali z manjšim številom tem dobite boljše rezultate? Kaj pa se zgodi ob uporabi precej večjega števila tem, npr. 50?

### **Opredelitev tem**

[Workflow6]

LDA za vsako temo vrne 10 besed, ki so najpogosteje povezane z njo.[[15]](#footnote-15) A kot smo videli, te besede niso nujno dovolj informativne, da bi omogočale opredelitev tematike. Zato si pri opredeljevanju lahko pomagamo z gradnikom **LDAvis** (Sievert in Shirley, 2014). Bistvena prednost te vizualizacije je, da besede oceni na podlagi relevantnosti, ki predstavlja razmerje med *specifičnostjo besede v določeni temi* in *specifičnostjo besede v celotnem korpusu*. Vrednost parametra lambda, ki ga v gradniku nastavljamo z drsnikom *Relevance*, lahko izbiramo med 0 in 1, pri čemer 1 prikaže besede glede na specifičnost besede v temi (kot jih vidimo v gradniku **Topic Modelling**), 0 pa glede na njihovo pogostost v celotnem korpusu. Specifičnost besede v celotnem korpusu imenujemo tudi dvig (*lift*), ki predstavlja razmerje med verjetnostjo besede v temi in verjetnostjo besede v korpusu.

Gradnik **LDAvis** torej povežemo z gradnikom **Topic Modelling**. Pri tem moramo paziti, da v LDAvis podamo prave vhodne podatke. Gradnik namreč potrebuje tabelo zastopanosti besed v temah, ki je dostopna v signalu *All Topics*. Povezavo uredimo tako, da dvakrat kliknemo povezavo med gradnikoma, nato pa povežemo signala *All Topics* in *Topics* (Slika 13)*.*

Slika : Določitev ustreznih vhodnih podatkov za gradnik LDAvis.

**LDAvis** privzeto prikaže besede, razvrščene uravnoteženo, torej z enakim razmerjem med specifičnostjo besede v temi in specifičnostjo besede v korpusu, kar v praksi običajno da dobre rezultate.

Z nastavljeno relevantnostjo dobimo drugačen, bolj informativen nabor besed, kot smo ga lahko videli v osnovnem gradniku **Topic Modelling** (Slika 14). Zdaj je očitno, da tema 8 govori o upravljanju prostora, tema 16 o političnih sporih in medijskem poročanju, tema 18 je mešanica tem o kulturi, športu in znanosti.

Slika : Prikaz pogostosti besed v temi (rdeča) in v celotnem vzorcu (siva) za prvo temo.

Poskusite sami: poljubno premaknite drsnik relevantnosti v desno in levo. Pri kateri nastavitvi je za skupine besed najlažje opredeliti temo?

[Slika0.3]

Tako pregledamo vse teme. Za boljšo preglednost rezultatov lahko nadomestimo generična poimenovanja tem (npr. Topic 1) z ustreznimi pomenskimi oznakami (npr. T1: prevoz otrok). Za ta korak moramo na gradnik **Topic Modelling** najprej pripeti gradnik **Select Columns,** nanj pa še **Edit Domain.**

Najprej odpremo gradnik **Select Columns**, kjer vidimo celoten nabor spremenljivk, vključnos frekvencami besed, ki smo jih za potrebe tematskega modeliranja ustvarili z gradnikom **Bag of Words**. Frekvenc besed ne potrebujemo več, zato jih odstranimo tako, da v razdelku *Features* na desni s tipkama Ctrl+A (Cmd+A) izberemo vse spremenljivke in jih prenesemo na levo stran, ki je namenjena spremenljivkam, ki jih želimo prezreti. Na levi strani v filter vpišemo *Topic*, s čimer poiščemo spremenljivke, ki označujejo teme (*Topic 1*, *Topic 2* itd.), jih izberemo in prenesemo nazaj na desno stran, ki je namenjena spremenljivkam, ki jih želimo obdržati v tabeli (Slika 15).

Slika : Urejen izbor spremenljivk v gradniku Select Columns.

Nato odpremo gradnik **Edit Domain**, v katerem bomo izvedli dejansko preimenovanje. S seznama na levi izberemo prvo temo in na desni strani v polju *Name* določimo njeno ime, npr. *T1: prevoz otrok* (Slika 16). Poimenovanje tem nam pomaga pri interpretaciji vizualizacij, s katerimi bomo teme raziskali v nadaljevanju.

Slika : Preimenovanje prve teme v gradniku Edit Domain.

Ko poimenujemo vse teme, dobimo seznam tem, o katerih so razpravljali poslanci od januarja 2019 do julija 2020. Ni presenetljivo, da na seznamu tem najdemo *epidemiološke ukrepe* in *zdravstvo,* medtem ko je morda nekatere druge teme, npr. *prevoz otrok, davek na vozila* ali *šport in kultura*, težje umestiti v prostor in čas, če nam aktualno dogajanje tistega časa ni znano. V grobem vidimo, da teme pokrivajo področja večine ministrstev, med drugim pravosodje, notranje zadeve, finance, zdravje, socialne zadeve in infrastrukturo. Obenem pa je za razumevanje takratnih prednostnih nalog tak seznam zanimiv tudi z vidika manjkajočih področij, ki bi jih sicer lahko pričakovali na seznamu tem parlamentarnih razprav, saj jih pokrivajo specifična ministrstva (npr. zunanje zadeve). Dejstvo, da npr. zunanjih zadev ni na seznamu tem, sicer ne pomeni, da poslanci o tej temi sploh niso razpravljali, nakazuje pa, da o njej ni bilo toliko razprave, da bi se uvrstila med 20 najizrazitejših tem za naš vzorec.

Tak seznam tem nam torej omogoča hiter pregled nad temami, ki so zaznamovale parlamentarno razpravo v preiskovanem obdobju, vendar pa ne razkriva dodatnih informacij, npr. o kateri temi je bilo največ razprave, kako so teme med sabo povezane oziroma kako so razporejene po obdobjih. Za odgovore na ta vprašanja je treba rezultate tematskega modeliranja dodatno analizirati, kar bomo storili v naslednjih poglavjih. Pred tem pa si bomo ogledali še, kako so teme razporejene po govorih. Na ta način lahko preverimo kontekst govorov in po potrebi prilagodimo poimenovanja tem, obenem pa je to eden od načinov, kako izluščimo govore, v katerih izbrana tema prevladuje.

Poskusite sami: poimenujte vse teme z ustrezno nadpomenko. Nekatere teme bo težje opredeliti. Pri teh si lahko pomagate z gradnikom **Corpus Viewer**, s katerim poiščete ustrezno besedo in raziščete njen kontekst.

### **Zastopanost tem po govorih**

[Workflow7]

Zaradi zakonitosti tematskega modeliranja (glejte 4. poglavje) govore običajno zaznamuje več kot ena tema, vendar so teme različno močno zastopane. Zastopanost oziroma verjetnost tem po govorih je izražena v razponu od 0 do 1, pri čemer 1 pomeni, da je tema v največji meri prisotna v govoru, 0 pa, da tema sploh ni prisotna. V praksi najpogosteje srečamo vmesne vrednosti. Ker imamo torej opravka z vrednostmi v istem razponu in ker želimo primerjati zastopanost tem med sabo, je najprimernejša vizualizacija toplotni diagram. Ustvarimo ga tako, da gradnik **Heat Map** povežemo z gradnikom **Edit Domain**.

V diagramu je vrednost ponazorjena z barvo: visoke vrednosti so označene z rumeno in belo (oziroma drugo barvo na desnem robu lestvice), nizke vrednost pa z modro (oziroma drugo barvo na levem robu lestvice). Vsak stolpec v diagramu predstavlja posamezno temo, vsaka vrstica pa govor. Govori so v diagramu razvrščeni v takem vrstnem redu, kot smo jih naložili v program, zaradi česar je trenutno diagram precej nepregleden, vendar lahko to z nekaj nastavitvami popravimo.

Najprej bomo govore, ki imajo podobno izražene teme, združili med sabo. Opravka imamo namreč z velikim številom govorov (6861), zato se vizualizacija raztegne v višino. V tem primeru lahko podobne govore predstavimo z eno vrstico in tako poskrbimo za kompaktnejšo postavitev. To storimo z možnostjo *Merge by k-means*, ki s postopkom metode k-voditeljev združi podobne govore med sabo. Prednastavljena vrednost je 50, mi pa jo bomo povečali na 150, saj je naših govorov veliko in ne želimo izgubiti preveč podrobnosti.

Vizualizacija je tako že bolj pregledna, vendar bi bila še bolj informativna, če bi bile podobne vrstice blizu ena drugi. Pozor, sedaj vrstice niso več posamezni govori, temveč skupine podobnih govorov. Vrstice organiziramo s še enim postopkom odkrivanja gruč, in sicer s hierarhičnim razvrščanjem v skupine, ki ga nastavimo v razdelku *Clustering,* kjer pri možnosti *Rows* izberemo *Clustering (opt. ordering*).

Izriše se precej lažje berljiv diagram (Slika 17), ki ima na levi strani dendrogram oziroma drevesno strukturo podobnosti govorov, ki kaže povezave med gručami, obenem pa je priročen za natančno izbiro želene skupine govorov. V prejšnjem poglavju smo ugotovili, da krovna tematika za določeno temo ali pa interpretacija nekaterih tem, npr. *šport in kultura*, ni popolnoma jasna. S tem diagramom lahko enostavno izberemo želene govore in podrobneje analiziramo njihovo vsebino.

Slika : Toplotni diagram zastopanosti tematik v govorih z izbrano vejo dendrograma, ki združuje govore z močno izraženo temo T18: šport in kultura.

Kot primer bomo izbrali temo *T18: šport in kultura*, ki jo predstavlja izbrana gruča v diagramu. Govore označimo tako, da izberemo vejo dendrograma, kjer je tema najbolj izrazita (rumene ali zelene barve) . Tako bo ta podmnožica govorov na voljo na izhodu gradnika. Zdaj na gradnik **Heat Map** pripnemo gradnik **Corpus Viewer** in si v njem preberemo izbrane govore.

Ugotovimo lahko, da govori obravnavajo različne teme, npr. izdajo potrdil o opravljenih izobraževanjih, založniško dejavnost ter financiranje vrhunskih športnikov. Na tak način smo torej potrdili poimenovanje teme (ki je sicer nabor več podtem), obenem pa smo iz celotnega vzorca podatkov izluščili zgolj govore na izbrano temo, ki jih nato lahko uporabimo za nadaljnje analize.

Pri interpretaciji tematskega modeliranja je nujno upoštevati, da govora ne zaznamuje zgolj ena tema, temveč je v vsakem govoru prisotnih več tem z različno verjetnostjo, kar lahko opazimo tudi pri temi *T11: proračun*. Če govore izberemo in jih pozorno preberemo, opazimo, da govori obravnavajo različne pristope k rebalansu proračuna, od povečanja sredstev za Urad za Slovence v zamejstvu in po svetu do financiranja domov za ostarele. Pravzaprav je to razvidno že iz diagrama, kjer vidimo, da je v nekaterih govorih v izbrani gruči precej močno izražena tudi tema *T10: domovi za ostarele*. Toplotni diagram torej odlično pokaže tudi primere tematskih presekov, kar je pomembno upoštevati, če to metodo uporabljamo zgolj za izbiro govor na določeno temo, ki jo želimo podrobneje raziskati. Na presek tem lahko kliknemo v diagramu in preberemo govore, ki se dotikajo obeh tem hkrati. Presek govorov o proračunu in domovih za ostarele predstavljajo govori o znižanju oskrbnine ter o uveljavitvi Zakona o dolgotrajni oskrbi.

Poskusite sami: izberite govore, ki govorijo o epidemiji, in jih preglejte.

## Najizrazitejše teme in povezave med njimi

[Workflow8]

Sedaj poznamo razporeditev tem po govorih. Videli smo, da govore zaznamuje več tem hkrati, zato nas zanima, kako so te teme v posameznih govorih med seboj povezane. Poleg tega nas zanima, katere teme v našem vzorcu najbolj izstopajo. Na obe vprašanji najlažje odgovorimo s tematsko karto, na kateri so teme razporejene glede na njihovo zastopanost in razmerja do drugih tem.

Tematsko karto bomo izdelali z gradnikom **MDS**. MDS je kratica za *multidimensional scaling* oziroma večrazsežnostno lestvičenje. Ta vizualizacija na podlagi vhodnih podatkov poskuša najti tak prikaz v ravnini, da povezane teme ležijo skupaj, nepovezane pa narazen. Pri postopku MDS se povezanost med temami izračuna glede na pomembnost besed v temah. Visoka povezanost med temami tako odraža zelo podobno porazdelitev besed med temami, pri čemer so lahko nekatere besede med temami celo deljene.

Povezavo med gradnikoma **Topic Modelling** in **MDS** nastavimo tako, da povežemo *All Topics* v *Data*. Na začetku vidimo zgolj sive točke v ravnini, ki predstavljajo teme. Za lažjo interpretacijo bomo točkam dodali oznake. To storimo tako, da pri možnosti *Labels* izberemo spremenljivko *Topics*. Kot lahko vidimo, so točkam pripisane zgolj prvotne oznake tem, ne pa tudi poimenovanja, ki smo jih dodali, zato si moramo pri interpretaciji pomagati s seznamom tem, ki ga pripravimo ročno (Tabela 1).

Tabela : Seznam tem z izvirno in preimenovano oznako.

|  |  |
| --- | --- |
| **Izvirno** | **Preimenovano** |
| Topic 1 | T1: prevoz otrok |
| Topic 2 | T2: nasilje nad živalmi |
| Topic 3 | T3: stanovanjska politika |
| Topic 4 | T4: davek na vozila |
| Topic 5 | T5: trgovine ob nedeljah |
| Topic 6 | T6: zdravstvo |
| Topic 7 | T7: vojska |
| Topic 8 | T8: upravljanje prostora |
| Topic 9 | T9: epidemija in mediji |
| Topic 10 | T10: domovi za starejše |
| Topic 11 | T11: proračun |
| Topic 12 | T12: sodstvo in financiranje šolstva |
| Topic 13 | T13: promet |
| Topic 14 | T14: epidemija in zdravstvo |
| Topic 15 | T15: proceduralno |
| Topic 16 | T16: politika in mediji |
| Topic 17 | T17: banke |
| Topic 18 | T18: šport in kultura |
| Topic 19 | T19: pokojnine |
| Topic 20 | T20: obvladovanje epidemije |

Nastavili bomo tudi velikost točk, in sicer tako, da bo ustrezala zastopanosti teme (ki je vsota verjetnosti teme v govorih, utežena z dolžino govorov). Možnost *Size* nastavimo na *Marginal Topic Probability*. Za boljšo preglednost pa pri možnosti *Color* prav tako izberemo *Marginal Topic Probability.*

Slika : Prikaz podobnosti tem v vizualizaciji MDS.

Tematska karta s položajem točk kaže podobnost tem, velikost in barva točk pa njihovo zastopanost v našem vzorcu govorov (Slika 18). Kadar so si teme podobne, vendar točke zaradi omejitev dvodimenzionalne vizualizacije ležijo daleč narazen, je ta podobnost izražena s premico, ki povezuje ti dve točki. Vidimo, da sta najbolj zastopani obsežni temi 15 (*proceduralno*) in 11 (*proračun*), najmanj pa tematsko ozke teme 1 (*prevoz otrok*), 3 (*stanovanjska politika*) in 20 (*obvladovanje epidemije*).

Močna zastopanost teme 11 (*proračun*) ni presenetljiva, saj je posledica pogostih razprav o rebalansu proračuna kot odgovoru na epidemiološke razmere (Vlada Republike Slovenije, 2020). Blizu nje ležita tudi temi 10 (*domovi za starejše*) in 19 (*pokojnine*), kar kaže, da se je o domovih za starejše in pokojninah razpravljalo tudi v kontekstu razporejanja davkoplačevalskega denarja. Ti dve temi (*T10* in *T19*) sta si blizu tudi ena drugi, kar pa zopet ni presenetljivo, saj obe govorita o starejši populaciji.

[Slika0.4]

Blizu teme, ki zaznamuje razprave o proračunu, ležita tudi temi 9 (*epidemija in mediji*) in 16 (*politika in mediji*), ki sta obenem povezani med sabo. Te povezave (med temama 9 in 16)ni težko potrditi, če smo vsaj malo seznanjeni z dogajanjem v času epidemije, ko so mediji odigrali ključno vlogo pri obveščanju javnosti o najnovejših ukrepih, obenem pa so bili pod hudimi pritiski politike. Zgolj na podlagi vizualizacije pa je težje interpretirati povezavo med temama 16 (*politika in mediji*) in 11 (*proračun*). Da bi lažje razumeli, na kakšen način sta ti dve temi povezani, si lahko ogledamo govore, ki izkazujeta podobno zastopanost obeh tem.

Govore na preseku obeh tem si ogledamo tako, da najprej ustvarimo podmnožico govorov, ki imajo močno izraženi obe temi, torej *T16: politika in mediji* in *T11: proračun*. To storimo z gradnikom **Select Rows**, ki ga povežemo z gradnikom **Topic Modelling**. V gradniku nastavimo dva pogoja, in sicer *Topic 16 is greater than 0.3* in *Topic 11 is greater than 0.3* (Slika 19). Tako bomo izbrali zgolj govore, v katerih sta temi 16 in 11 izraženi z več kot 30-odstotno verjetnostjo.[[16]](#footnote-16)

Slika : Nastavitev mej za prisotnost tem v grandiku Select Rows.

**Select Rows** nato povežemo v gradnik **Corpus Viewer**, v katerem preberemo izbranih šest govorov, ki predstavljajo presek tem o proračunu ter politiki in medijih (Slika 20). Iz govorov je razvidno, da gre pri tem za strankarska prerekanja o rebalansu proračuna.

Slika : Pregled izbranih dokumentov v gradniku Corpus Viewer.

*Poskusite sami: Na enak način raziščite, kako sta povezani tema 9 in tema 11.*

## Teme pred in med epidemijo

[Workflow9]

Ugotovili smo, katere teme v našem vzorcu najbolj izstopajo, zdaj pa nas zanima, katere teme so najbolj značilne za obdobje pred epidemijo in med njo. Razliko med obdobjema (ki sta označena kot podkorpusa *Reference* in *COVID*) bomo raziskali v gradniku **Box Plot**. Ta vizualizacija, v slovenščini znana tudi kot škatla z brki, kaže porazdelitve spremenljivk ter omogoča njihovo enostavno primerjavo po [kategoričnih spremenljivkah](https://en.wikipedia.org/wiki/Categorical_variable) (npr. spol, datum, stranka).

**Box Plot** povežemo z gradnikom **Edit Domain**, saj bodo tako med rezultati že poimenovane teme. Ker želimo obdobji primerjati med sabo, v spodnjem razdelku na levi strani izberemo spremenljivko *Subcorpus*, v zgornjem razdelku pa kar temo *T1: prevoz otrok*. Na desni se prikažeta dve škatli z brki, zgornja za obdobje med epidemijo (*COVID*), spodnja pa za obdobje pred epidemijo (*Reference*) (Slika 21). Iz vizualizacije lahko razberemo, da so bile razprave na temo sodstva pogostejše pred epidemijo kot pa med njo. Obenem nam rezultat statističnega testa, ki je naveden pod vizualizacijo, potrjuje, da gre za statistično značilno razliko, njegova vrednost p je namreč nižja od 0,05. Na podlagi tega lahko zaključimo, da je bila tema o sodstvu bolj značilna za predepidemično obdobje.

Slika : Škatli z brki za temo T1: prevoz otrok za obdobje pred in med epidemijo.

[Slika0.5]

Na tak način bi lahko pregledali porazdelitev za vsako temo posebej, vendar bi bilo to precej časovno zamudno. Ker trenutno nismo osredotočeni na eno samo temo, ampak bi radi ugotovili, pri katerih temah se kažejo največje razlike med obdobjema, si bomo pomagali z možnostjo *Order by relevance to subgroups* pod razdelkom *Variable.* Tako bomo samodejno razvrstili spremenljivke glede na njihovo vrednost statističnega testa. Na vrhu bodo torej prikazane tiste spremenljivke, ki za izbrano razdelitev, ki smo jo določili v razdelku *Subgroups* (v našem primeru *Subcorpus*), izkazujejo največje razlike.

Povsem na vrh razdelka *Variable* so uvrščene spremenljivke, kot so *Title, From, To* in *Meeting*. Gre torej za spremenljivke, ki izkazujejo izrazito različno porazdelitev med podkorpusoma *Reference* in *COVID*, vendar to ni presenetljivo, saj so vse te spremenljivke časovno zamejene, kar je bil tudi kriterij za oblikovanje obeh podkorpusov (glejte poglavje 3.3). Za naše potrebe je zanimiv vrstni red tem, ki sledijo tem prvim spremenljivkam.

Najvišje na seznamu sta temi *T12: sodstvo in financiranje šolstva* (Slika 22) in *T20: obvladovanje epidemije*. Ko ju izberemo, vizualizacija na desni pokaže, da so poslanci o sodstvu govorili več pred epidemijo, o obvladovanju epidemije pa, nepresenetljivo, med epidemijo. Rezultat študentovega t-testa (13,824, p<0,05), naveden pod vizualizacijo, obenem potrjuje, da gre za statistično pomembno razliko med obdobjema in da sta torej temi pogostejši v določenem času. Glede na okoliščine tak rezultat ni presenetljiv, kaže pa na zanesljivost metode za podobne raziskave. Sledita temi *T14: epidemija in zdravstvo* in *T2: nasilje nad živalmi* (prva izkazuje večjo pogostost v epidemičnem, druga pa v predepidemičnem času). Tudi tukaj se kaže statistično značilna razlika pri zastopanosti posamezne teme med obdobjema.

Slika : Škatli z brki za temo T12: sodstvo in financiranje šolstva.

Dejstvo, da sta temi o epidemiji bolj izraziti v COVID podkorpusu, pravzaprav ni presenetljivo. Bolj kot kaj drugega to potrjuje uporabnost tematskega modeliranja za podobne raziskave.

Teme lahko podrobneje raziščemo s pomočjo gradnikov **Select Rows** in **Corpus Viewer**. **Select Rows** povežemo na **Edit Domain**. V **Select Rows** izberemo spremenljivko »T14: epidemija in zdravstvo« in nastavimo pogoj *is greater than* *0.7*, s čimer bomo izbrali tiste govore, ki imajo zastopanost teme višjo kot 0,7 (več kot 70 % verjetnost, da je tema prisotna v govoru). Izbrane govore nato lahko pogledamo v gradniku **Word Cloud**.

Slika : Oblak najpogostejših besed za temo 14.

Vidimo, da je govora zlasti o *pogodbah* za *nabavo* *opreme*, delu *vlade* in *ministrov* v času epidemije, vlogi *podjetij* za nakup opreme, na primer *ventilatorjev*, ter o *Zavodu* za blagovne rezerve (Slika 23). Veliko poudarka pri reševanju epidemije je zlasti na dobavi opreme, birokratski odgovornosti za nakup ter vlogi države pri zagotavljanju dobave. Ministrstvo za zdravje je že 3. 2. 2020 pričelo z ugotavljanjem zalog zaščitne in medicinske opreme, med drugim tudi ventilatorjev (RS RS, 2021), 2. 4. 2020 pa je Državni zbor za seji sprejel Zakon o interventnih ukrepih za zajezitev epidemije COVID-19 in omilitev njenih posledic za državljane in gospodarstvo (DZ RS, 2020).

Poskusite sami: Na enak način, kot smo primerjali podkorpusa Reference in COVID, primerjajte med sabo tematsko porazdelitev v govorih opozicije in koalicije.

# Zaključek

Parlamentarni korpusi, ki vsebujejo zapise parlamentarnih sej, so pomemben vir za raziskovanje političnega in družbenega dogajanja. Običajno vsebujejo številne metapodatke o govorcih in govorih, pogosto pa so tudi bogato jezikoslovno označeni, kar raziskovalcem omogoča oblikovanje najrazličnejših raziskovalnih vprašanj. Zaradi velikega obsega so za raziskovanje vsebine parlamentarnih korpusov izrazito primerne metode rudarjenja besedil, med katere uvrščamo tudi tematsko modeliranje, ki omogoča luščenje vsebinskih tematik iz korpusa. V tem učnem gradivu smo tako prikazali uporabo tehnike LDA, ki je ena od najbolj priljubljenih tehnik za tematsko modeliranje. Analizo smo izvedli v prostodostopnem programu za vizualno programiranje Orange, ki omogoča napredno obdelavo podatkov tudi tistim, ki niso vešči pisanja programske kode. Za potrebe analize smo uporabili korpus ParlaMint-SI, ki vsebuje slovenske parlamentarne razprave.

Slika : Prikaz celotnega delotoka iz učnega gradiva.

Učno gradivo je zasnovano za samostojno učenje in vsebuje enostavne opise korakov, ki so pospremljeni s številnimi posnetki zaslona za lažje sledenje (Slika 24). Poleg tega so dodana navodila za dodatno samostojno delo, kar pomaga utrjevati znanje in spodbuja k samostojni rabi programa Orange. V učnem gradivu je posebna pozornost namenjena predstavitvi bistvenih značilnosti obravnavanih podatkov, opisane pa so tudi zakonitosti in omejitve metode, kar uporabnike spodbuja k primerno kritični uporabi metode in ustrezni interpretaciji rezultatov.

V učnem gradivu smo za potrebe analize sicer uporabili slovenske parlamentarne podatke, vendar je analizo enostavno razširiti ne le na druge jezike, ampak tudi na druge primerne besedilne žanre. Družina korpusov ParlaMint namreč združuje številne primerljive parlamentarne korpuse, predstavljena metoda tematskega modeliranja pa ni jezikovnospecifična, zato jo lahko uporabimo na katerem koli jeziku. Vzporedno s slovenskim učnim gradivom smo pripravili tudi gradivo z govori britanskega parlamenta, ki ga lahko vzamete za primerjavo. Vrednost učnega gradiva za študente in raziskovalce na področju družboslovja in humanistike torej vsekakor sega prek specifičnih raziskovalnih vprašanj, obravnavanih v tem učnem gradivu.

**Zahvala**

Delo, opisano v pričujočem gradivu, je financirala Slovenska raziskovalna agencija, raziskovalni program P6-0436: Digitalna humanistka: viri, orodja, metode (2022-2027), projekt Social Sciences & Humanities Open Cloud (SSHOC) (<https://www.sshopencloud.eu/>), projekt CLARIN ERIC ParlaMint (<https://www.clarin.eu/parlamint>) in raziskovalna infrastruktura DARIAH-SI. Posebna zahvala gre tudi Çağrıju Çöltekinu, Marti Kołczyński, Jiřini Popelikovi, Mladenu Zobcu in Juretu Skubicu za strokovno recenzijo in testiranje učnega gradiva ter vse njihove koristne komentarje.

# Reference

Abercrombie, G., & Batista-Navarro, R. (2020). Sentiment and position-taking analysis of parliamentary debates: A systematic literature review. *Journal of Computational Social Science*, *3*(1), 245–270.

Albalawi, R., Yeap, T. H., & Benyoucef, M. (2020). Using topic modeling methods for short-text data: A comparative analysis. *Frontiers in Artificial Intelligence*, *3*, 42.

Allen, C., & Murdock, J. (2020). LDA topic modelling: Context for the history & philosophy of science [Preprint]. http://philsci-archive.pitt.edu/17261/

Arun, R., Suresh, V., Veni Madhavan, C., & Murthy, N. (2010). *On finding the natural number of topics with latent dirichlet allocation: Some observations*. 391–402.

Bayley, P. (2004). *Cross-cultural perspectives on parliamentary discourse* (Vol. 10). John Benjamins Publishing.

Bergmann, H., Geese, L., Koss, C., & Schwemmer, C. (2018). *Using legislative speech to unveil conflict between coalition parties* [Preprint]. SocArXiv. https://doi.org/10.31235/osf.io/pgnwa

Blätte, A., Gehlhar, S., & Leonhardt, C. (2020). *The Europeanization of Parliamentary Debates on Migration in Austria, France, Germany, and the Netherlands*. 66–74.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Brezovšek, M., Haček, M., Ferfila, B., & Zajc, D. (2012). *Politični sistem Republike Slovenije*. Fakulteta za družbene vede.

Chizhik, A. V., & Sergeyev, D. A. (2021). *Exploring the Parliamentary Discourse of the Russian Federation Using Topic Modeling Approach*. 403–416.

Curran, B., Higham, K., Ortiz, E., & Vasques Filho, D. (2018). Look who’s talking: Two-mode networks as representations of a topic model of New Zealand parliamentary speeches. *PLOS ONE*, *13*(6), e0199072. https://doi.org/10.1371/journal.pone.0199072

de Campos, L. M., Fernandez-Luna, J. M., Huete, J. F., & Redondo-Expósito, L. (2021). LDA-based term profiles for expert finding in a political setting. *Journal of Intelligent Information Systems*, *56*(3), 529–559.

Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., & Starič, A. (2013). Orange: Data mining toolbox in Python. *The Journal of Machine Learning Research*, *14*(1), 2349–2353.

DZ RS (Državni zbor Republike Slovenija) (2020). ZIUZEOP: Zakon o interventnih ukrepih za zajezitev epidemije COVID-19 in omilitev njenih posledic za državljane in gospodarstvo (Uradni list RS, št. 49/20, 61/20, 67/20, 80/20 - ZIUOOPE, 101/20 - skl. US, 152/20 - ZZUOOP, 175/20 - ZIUOPDVE, 203/20 - ZIUPOPDVE, 15/21 - ZDUOP). http://www.pisrs.si/Pis.web/pregledPredpisa?id=ZAKO8190

Erjavec, T., et al. (2021). *Multilingual comparable corpora of parliamentary debates ParlaMint 2.1*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, <http://hdl.handle.net/11356/1432>.

Erjavec, T., Ogrodniczuk, M., Osenova, P., & et al. (2022). The ParlaMint corpora of parliamentary proceedings. *Lang Resources & Evaluation*. https://doi.org/10.1007/s10579-021-09574-0

Erjavec, T., & Pančur, A. (2019). Parla-CLARIN: TEI guidelines for corpora of parliamentary proceedings. *Book of Abstracts of the TEI2019: What Is Text, Really*.

Fišer, D., & Pahor de Maiti, K. (2021). » Prvič, sem političarka in ne politik, drugič pa…«. *Contributions to Contemporary History*, *61*(1). https://doi.org/10.51663/pnz.61.1.07

Gkoumas, D., Pontiki, M., Papanikolaou, K., & Papageorgiou, H. (2018). *Exploring the Political Agenda of the Greek Parliament Plenary Sessions* (D. Fišer, M. Eskevich, & F. de Jong, Eds.).

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297.

Høyland, B., & Søyland, M. G. (2019). Electoral reform and parliamentary debates. *Legislative Studies Quarterly*, *44*(4), 593–615.

Ilie, C. (2010). *European parliaments under scrutiny: Discourse strategies and interaction practices* (Vol. 38). John Benjamins Publishing.

Jacobs, T., & Tschötschel, R. (2019). Topic models meet discourse analysis: A quantitative tool for a qualitative approach. *International Journal of Social Research Methodology*, *22*(5), 469–485.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*.

Kilroy, D. (2021). All the king’s men? A demographic study of opinion in the first English Parliament of James I, 1604–10. *Parliaments, Estates and Representation*, *41*(1), 1–23.

Martin, F., & Johnson, M. (2015). *More efficient topic modelling through a noun only approach*. 111–115.

Meeks, E., & Weingart, S. B. (2012). The digital humanities contribution to topic modeling. *Journal of Digital Humanities*, *2*(1), 1–6.

Mollin, S. (2007). The Hansard hazard: Gauging the accuracy of British parliamentary transcripts. *Corpora*, *2*(2), 187–210.

Morstatter, F., Shao, Y., Galstyan, A., & Karunasekera, S. (2018). *From alt-right to alt-rechts: Twitter analysis of the 2017 German federal election*. 621–628.

Müller-Hansen, F., Callaghan, M. W., Lee, Y. T., Leipprand, A., Flachsland, C., & Minx, J. C. (2021). Who cares about coal? Analyzing 70 years of German parliamentary debates on coal with dynamic topic modeling. *Energy Research & Social Science*, *72*, 101869.

Norton, P. (2002). *Parliaments and citizens in Western Europe* (Vol. 3). Psychology Press.

Pančur, A., & Šorn, M. (2016). Digitalni pristop k parlamentarni zgodovini: Uporaba gradiva Državnega zbora v digitalni humanistiki. *Četrt stoletja Republike Slovenije - izzivi, dileme, pričakovanja*, 115–126.

Petukhova, V., Malchanau, A., & Bunt, H. (2015). *Modelling argumentation in parliamentary debates* (M. Baldoni & et al., Eds.). Springer.

Piersma, H., Tames, I., Buitinck, L., Van Doornik, J., & Marx, M. (2014). War in parliament: What a digital approach can add to the study of parliamentary history. *Digital Humanities Quarterly*, *8*(1).

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959.

Proksch, S.-O., & Slapin, J. B. (2010). Position taking in European Parliament speeches. *British Journal of Political Science*, *40*(3), 587–611.

RS RS (Računsko sodišče Republike Slovenije) (2021). Revizijsko poročilo: Učinkovitost nabav zaščitne in medicinske opreme za obvladovanje širjenja virusa SARS-CoV-2, 17. 3. 2021. https://www.rs-rs.si/fileadmin/user\_upload/Datoteke/Revizije/2021/ZascitnaOprema/ZascitnaOprema2020\_RSP\_RevizijskoP.pdf.

Rheault, L., Beelen, K., Cochrane, C., & Hirst, G. (2016). Measuring emotion in parliamentary debates with automated textual analysis. *PloS One*, *11*(12), e0168843.

Rheault, L., & Cochrane, C. (2020). Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis*, *28*(1), 112–133.

Rosa, A. B., Gudowsky, N., & Repo, P. (2021). Sensemaking and lens-shaping: Identifying citizen contributions to foresight through comparative topic modelling. *Futures*, *129*, 102733.

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., & Sedlmair, M. (n.d.). *Supervised Sentiment Analysis of Parliamentary Speeches and News Reports*.

Schmidt, B. M. (2012). Words alone: Dismantling topic models in the humanities. *Journal of Digital Humanities*, *2*(1), 49–65.

Schuler, P. (2020). Position taking or position ducking? A theory of public debate in single-party legislatures. *Comparative Political Studies*, *53*(9), 1493–1524.

Serrano, J. C. M., Shahrezaye, M., Papakyriakopoulos, O., & Hegelich, S. (2019). *The rise of Germany’s AfD: A social media analysis*. 214–223.

Shadrova, A. (2021). Topic models do not model topics: Epistemological remarks and steps towards best practices. *Journal of Data Mining & Digital Humanities*, *2021*.

Sieberer, U., Müller, W. C., & Heller, M. I. (2011). Reforming the rules of the parliamentary game: Measuring and explaining changes in parliamentary rules in Austria, Germany, and Switzerland, 1945–2010. *West European Politics*, *34*(5), 948–975.

Sievert, C., & Shirley, K. (2014). *LDAvis: A method for visualizing and interpreting topics*. 63–70.

Smith, N., & Graham, T. (2019). Mapping the anti-vaccination movement on Facebook. *Information, Communication & Society*, *22*(9), 1310–1327.

Truan, N., & Romary, L. (2021). Building, Encoding, and Annotating a Corpus of Parliamentary Debates in XML-TEI: A Cross-Linguistic Account. *Journal of the Text Encoding Initiative*.

van der Zwaan, J. M., Marx, M., & Kamps, J. (2016). *Validating Cross-Perspective Topic Modeling for Extracting Political Parties’ Positions from Parliamentary Proceedings.* 28–36.

Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. *Information Systems*, *94*, 101582.

Vlada Republike Slovenije (2020). Obrazložitev splošnega dela predloga rebalansa proračuna Republike Slovenije za leto 2020. https://www.gov.si/assets/ministrstva/MF/Proracun-direktorat/Drzavni-proracun/Sprejeti-proracun/Rebalans-2020/Obr-splosni-del-in-politike/REB20\_obrsplosnidel.pdf.

Wiedemann, G. (2016). *Text mining for qualitative data analysis in the social sciences* (Vol. 1). Springer.

Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015). *A heuristic approach to determine an appropriate number of topics in topic modeling*. *16*(13), 1–10.

1. V učnem gradivu uporabljamo generični moški spol. [↑](#footnote-ref-1)
2. Gradniki predstavljajo posamezne korake analize. [↑](#footnote-ref-2)
3. Ključna in zavezujoča pravila delovanja, ki urejajo organizacijo in delo parlamenta ter pravice in dolžnosti parlamentarcev, so zapisana v poslovnikih posameznega parlamenta, npr. [v poslovniku slovenskega državnega zbora](http://www.pisrs.si/Pis.web/pregledPredpisa?id=POSL34), [poslovniku britanskega parlamenta](https://erskinemay.parliament.uk/), [poslovniku nemškega zveznega parlamenta](https://www.btg-bestellservice.de/pdf/80060000.pdf) itd. [↑](#footnote-ref-3)
4. Diahrono analizo tem, torej razvoj tem skozi čas, je mogoče analizirati tudi le s tehniko LDA, vendar je pri tem treba za vsako časovno obdobje narediti ločen tematski model, interpretirati dobljene teme in jih nato primerjati med različnimi časovnimi obdobji. Tak pristop zahteva več ročnega, subjektivnega dela, kar lahko negativno vpliva na rezultate. Druga možnost za časovno analizo tem je *dinamična nenegativna matrična faktorizacija* (*dynamic* *NMF*) (Gkoumas idr., 2018), ki pa tematike med različnimi obdobji poveže zgolj posredno. (Müller-Hansen idr., 2021) [↑](#footnote-ref-4)
5. Če ste Orange že kdaj uporabljali, svetujemo, da ponastavite gradnike, in sicer z možnostjo *Options 🡪 Reset widget settings.* Tako boste lažje sledili analizi, kot je opisana v tem učnem gradivu. [↑](#footnote-ref-5)
6. Prikazani so prednaloženi podatki. [↑](#footnote-ref-6)
7. Predhodno smo preverili, da sta tako časovno zamejeni obdobji primerljivi tudi z vidika količine govorov in sej, ki jih vsebujeta. [↑](#footnote-ref-7)
8. Če bi želeli analizirati celoten korpus ali uporabiti drugačno časovno obdobje, je celoten korpus na voljo v repozitoriju CLARIN.SI v [datoteki s podatki *ParlaMint-SI.ana.tgz*](http://hdl.handle.net/11356/1431). [↑](#footnote-ref-8)
9. Potrebujemo tako datoteke v formatu CoNLL-Ukot tudi tiste v formatu TSV. Datoteke CoNLL-Uvsebujejo jezikoslovno označene govore, datoteke TSVpa metapodatke o govorcih in govorih (glejte poglavje 3.3). [↑](#footnote-ref-9)
10. Število pojavnic označuje število vseh besed, števil in ločil, ki so v korpusu. Število različnic pa je število unikatnih pojavnic v korpusu. [↑](#footnote-ref-10)
11. Pri določanju iskalnega izraza lahko uporabimo pravila [regularnih izrazov](https://www.sketchengine.eu/guide/regular-expressions/), ki nam omogočajo iskanje zelo specifičnih zadetkov oziroma hkratno iskanje različnih oblik z enim iskalnim izrazom. Tako lahko npr. z iskalnim izrazom *epidemij\** zajamemo vse sklanjatve besede *epidemija*. [↑](#footnote-ref-11)
12. Nekateri metapodatki lahko manjkajo, če so bile izvorne parlamentarne evidence nepopolne. Jezikoslovne oznake so bile pripisane avtomatsko, kar pomeni, da so v korpusu možne napake, vendar so te minimalne, saj so bila uporabljena orodja z visoko natančnostjo: lematizacija 98–99 % pravilnost, oblikoslovne oznake 94–77 %, skladenjske oznake 87–94 % (Erjavec idr., 2022). [↑](#footnote-ref-12)
13. Nabor metapodatkov je enak za celotno družino korpusov ParlaMint (glejte poglavje 3.3), vendar vsi korpusi ne vsebujejo vseh metapodatkov. [↑](#footnote-ref-13)
14. Če želimo onemogočiti posodobitev ob vsaki spremembi parametra, odznačimo možnost *Apply Automatically* levo spodaj, nato pa, ko smo vnesli vse želene nastavitve, kliknemo *Apply*. [↑](#footnote-ref-14)
15. Možno je, da dobite nekoliko drugačne rezultate, kot so navedeni v gradivu. LDA je namreč generativni model, ki deluje naključno. Na svojem računalniku bi morali vedno znova dobiti enake rezultate, medtem ko se rezultati lahko razlikujejo med različicami programa Orange ter med operacijskimi sistemi. [↑](#footnote-ref-15)
16. Mejo pri 30 odstotkih smo izbrali, ker je to najnižja vrednost, ki vrne vsaj nekaj dokumentov. Pri tem se je treba zavedati, da je meja precej nizka, torej dokumenti izkazujejo nizko zastopanost teme. Vrednost lahko poljubno prilagajate. [↑](#footnote-ref-16)