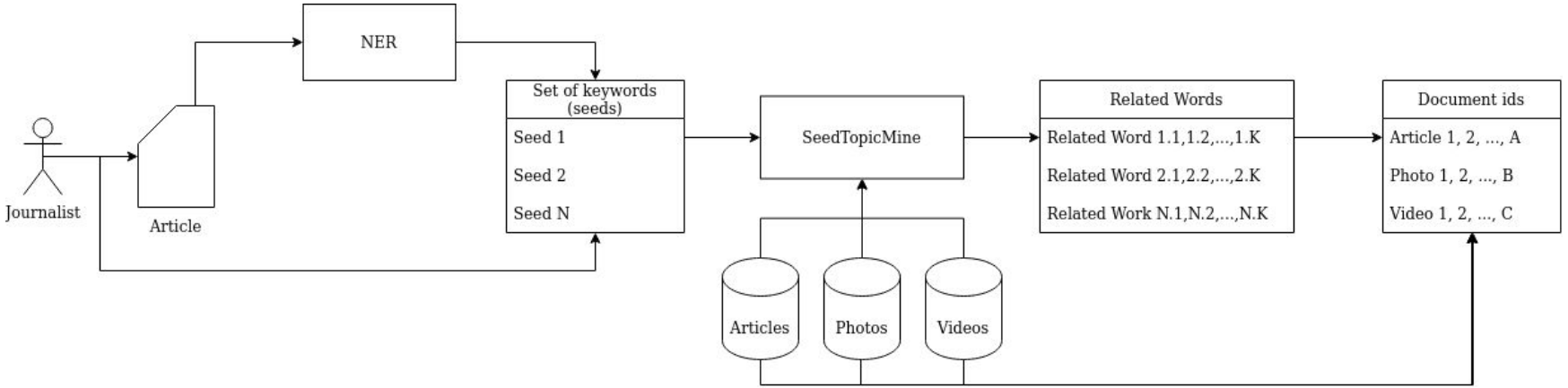
News Linking

Illia Tesliuk
Panpan Liu
Trifebi Shina

Goal

- To prepare a system that would could connect different types of information in Slovenian based on their content/descriptions:
  - News Articles
  - Photos - text descriptions
  - Videos - text descriptions
- Dataset: data collected from API of The Slovenian Press Agency (STA)
- Use cases - a journalist publishes a new article and would like to:
  - Place links to the articles describing the context or preceding events
  - Find photos/videos describing the same event

System Review



- Keywords are entered manually or extracted by NER model
- Set of keywords is fed to SeedTopicMine framework
- SeedTopicMine extracts the words related to each of keywords
- We use dataset to find documents that contain these related words

SeedTopicMine

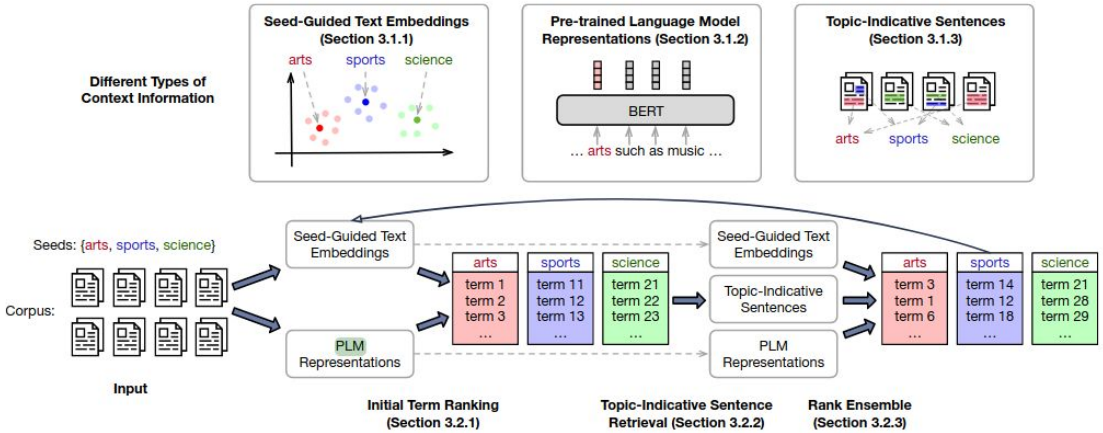


Figure 1: Overview of the SeedTopicMine framework.

- Iterative framework
- Combines different types of contexts
- Receives a set of keywords (seeds) and for each seed returns a list of related words

Data Pre-processing

- We’ve collected STA articles, photos and videos over a period of 1 year
- Filtered out redundant news categories
- Transformed collected data into the format of New York Times dataset (used in the SeedTopicMine paper):
  - All words lowercase
  - Punctuation marks followed and preceded by spaces
  - 1 document / 1 file line
- Some of the documents contain ‘related’ / ‘attachedTo’ fields - links to related articles / media, which we’ll use as ground-truth data

Articles

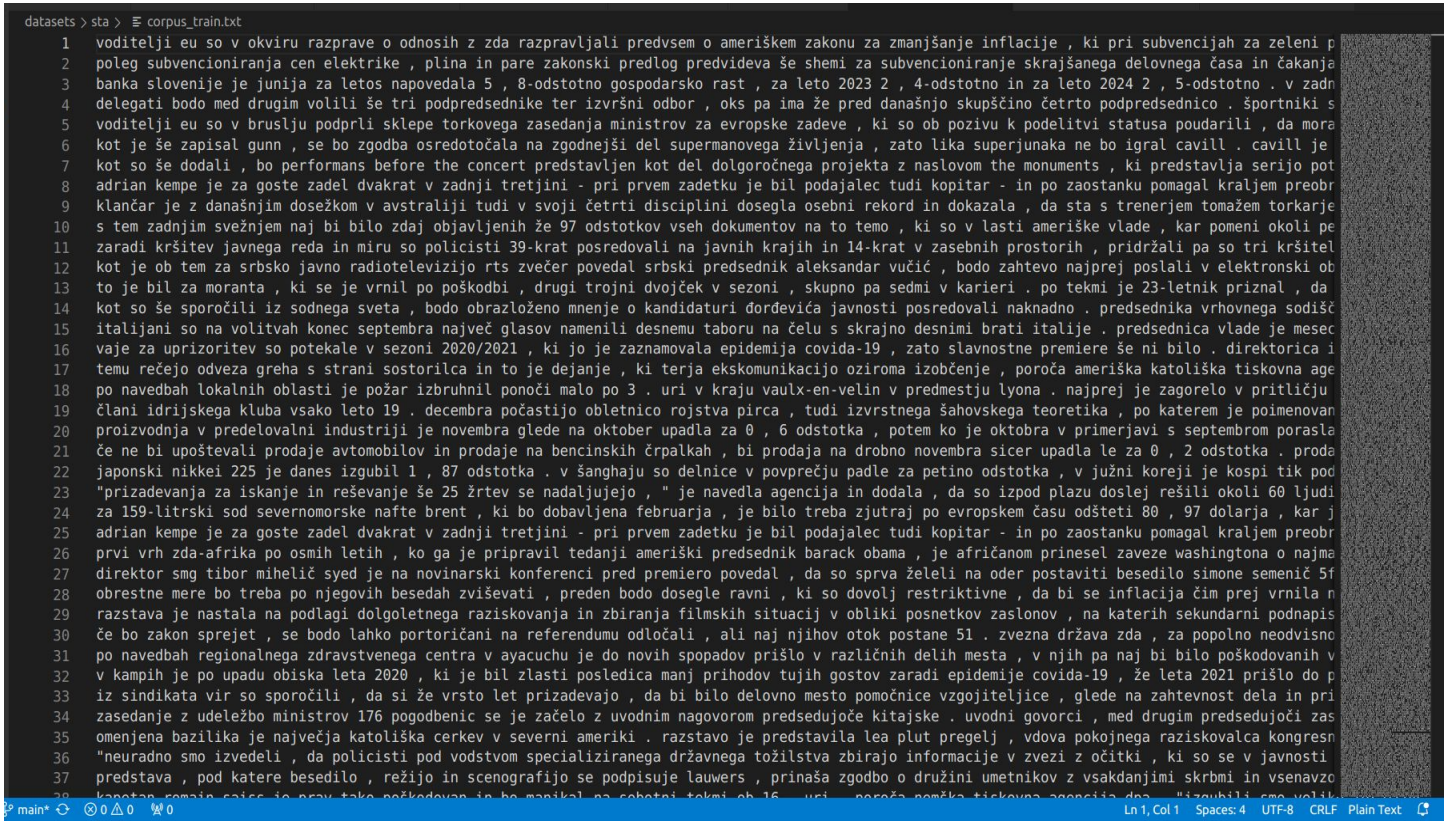
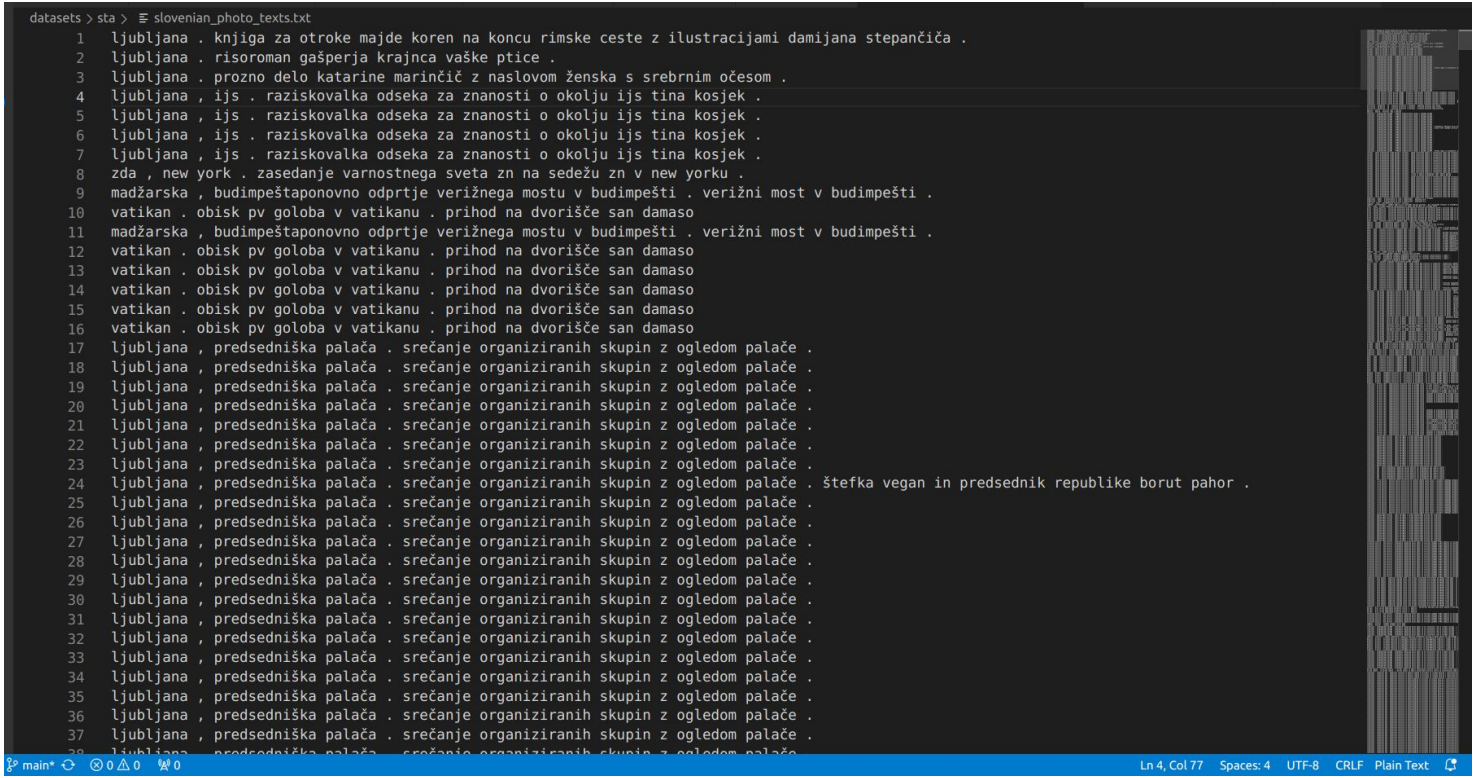

datasets > sta > corpus_train.txt
1 voditelji eu so v okviru razprave o odnosih z zda razpravljali predvsem o ameriškem zakonu za zmanjšanje inflacije , ki pri subvencijah za zeleni p
2 poleg subvencioniranja cen elektrike , plina in pare zakonski predlog predvideva še shemi za subvencioniranje skrajšanega delovnega časa in čakanja
3 banka slovenije je junija za letos napovedala 5 , 8-odstotno gospodarsko rast , za leto 2023 2 , 4-odstotno in za leto 2024 2 , 5-odstotno . v zadn
4 delegati bodo med drugim volili še tri podpredsednike ter izvršni odbor , oks pa ima že pred današnjo skupščino četrto podpredsednico . športniki s
5 voditelji eu so v bruslju podprli sklepe torkovega zasedanja ministrov za evropske zadeve , ki so ob pozivu k podelitvi statusa poudarili , da mora
6 kot je še zapisal gunn , se bo zgodba osredotočala na zgodnejši del supermanovega življenja , zato lika superjunaka ne bo igral cavill . cavill je
7 kot so še dodali , bo performans before the concert predstavljen kot del dolgoročnega projekta z naslovom the monuments , ki predstavlja serijo pot
8 adrian kempe je za goste zadel dvakrat v zadnji tretjini - pri prvem zadetku je bil podajalec tudi kopitar - in po zaostanku pomagal kraljem preobr
9 klančar je z današnjim dosežkom v avstraliji tudi v svoji četrti disciplini dosegla osebni rekord in dokazala , da sta s trenerjem tomažem torkarje
10 s tem zadnjim svežnjem naj bi bilo zdaj objavljenih že 97 odstotkov vseh dokumentov na to temo , ki so v lasti ameriške vlade , kar pomeni okoli pe
11 zaradi kršitev javnega reda in miru so policisti 39-krat posredovali na javnih krajih in 14-krat v zasebnih prostorih , pridržali pa so tri kršitel
12 kot je ob tem za srbsko javno radiotelevizijo rts zvečer povedal srbski predsednik aleksandar vučić , bodo zahtevo najprej poslali v elektronski ob
13 to je bil za moranta , ki se je vrnil po poškodbi , drugi trojni dvojček v sezoni , skupno pa sedmi v karieri . po tekmi je 23-letnik priznal , da
14 kot so še sporočili iz sodnega sveta , bodo obrazloženo mnenje o kandidaturi đorđevića javnosti posredovali naknadno . predsednika vrhovnega sodišč
15 italijani so na volitvah konec septembra največ glasov namenili desnemu taboru na čelu s skrajno desnimi brati italije . predsednica vlade je mesec
16 vaje za uprizoritev so potekale v sezoni 2020/2021 , ki jo je zaznamovala epidemija covida-19 , zato slavnostne premiere še ni bilo . direktorica i
17 temu rečejo odveza greha s strani sostorilca in to je dejanje , ki terja ekskomunikacijo oziroma izobčenje , poroča ameriška katoliška tiskovna age
18 po navedbah lokalnih oblasti je požar izbruhnil ponoči malo po 3 . uri v kraju vaulx-en-velin v predmestju lyona . najprej je zagorelo v pritličju
19 člani idrijskega kluba vsako leto 19 . decembra počastijo obletnico rojstva pirca , tudi izvrstnega šahovskega teoretika , po katerem je poimenovan
20 proizvodnja v predelovalni industriji je novembra glede na oktober upadla za 0 , 6 odstotka , potem ko je oktobra v primerjavi s septembrom porasla
21 če ne bi upoštevali prodaje avtomobilov in prodaje na bencinskih črpalkah , bi prodaja na drobno novembra sicer upadla le za 0 , 2 odstotka . proda
22 japonski nikkei 225 je danes izgubil 1 , 87 odstotka . v šanghaju so delnice v povprečju padle za petino odstotka , v južni koreji je kospi tik pod
23 "prizadevanja za iskanje in reševanje še 25 žrtev se nadaljujejo , " je navedla agencija in dodala , da so izpod plazu doslej rešili okoli 60 ljudi
24 za 159-litrski sod severnomorske nafte brent , ki bo dobavljena februarja , je bilo treba zjutraj po evropskem času odšteti 80 , 97 dolarja , kar j
25 adrian kempe je za goste zadel dvakrat v zadnji tretjini - pri prvem zadetku je bil podajalec tudi kopitar - in po zaostanku pomagal kraljem preobr
26 prvi vrh zda-afrika po osmih letih , ko ga je pripravil tedanji ameriški predsednik barack obama , je afričanom prinesel zaveze washingtona o najma
27 direktor smg tibor mihelič syed je na novinarski konferenci pred premiero povedal , da so sprva želeli na oder postaviti besedilo simone semenič 5f
28 obrestne mere bo treba po njegovih besedah zviševati , preden bodo dosegle ravni , ki so dovolj restriktivne , da bi se inflacija čim prej vrnila n
29 razstava je nastala na podlagi dolgoletnega raziskovanja in zbiranja filmskih situacij v obliki posnetkov zaslonov , na katerih sekundarni podnapis
30 če bo zakon sprejet , se bodo lahko portoričani na referendumu odločali , ali naj njihov otok postane 51 . zvezna država zda , za popolno neodvisno
31 po navedbah regionalnega zdravstvenega centra v ayacuchu je do novih spopadov prišlo v različnih delih mesta , v njih pa naj bi bilo poškodovanih v
32 v kampih je po upadu obiska leta 2020 , ki je bil zlasti posledica manj prihodov tujih gostov zaradi epidemije covida-19 , že leta 2021 prišlo do p
33 iz sindikata vir so sporočili , da si že vrsto let prizadevajo , da bi bilo delovno mesto pomočnice vzgojiteljice , glede na zahtevnost dela in pri
34 zasedanje z udeležbo ministrov 176 pogodbenic se je začelo z uvodnim nagovorom predsedujoče kitajske . uvodni govorci , med drugim predsedujoči zas
35 omenjena bazilika je največja katoliška cerkev v severni ameriki . razstavo je predstavila lea plut pregelj , vdova pokojnega raziskovalca kongresn
36 "neuradno smo izvedeli , da policisti pod vodstvom specializiranega državnega tožilstva zbirajo informacije v zvezi z očitki , ki so se v javnosti
37 predstava , pod katere besedilo , režijo in scenografijo se podpisuje lauwers , prinaša zgodbo o družini umetnikov z vsakdanjimi skrbmi in vsenavzo
main* 0 0 0 Ln 1, Col 1 Spaces: 4 UTF-8 CRLF Plain Text

Photo Descriptions



Ground-truth Links

Articles

sta > slovenian_related.txt

```
3117853, 3118105, 3118162
3114451, 3114800, 3115621, 3115632, 3116488, 3117048, 3117100
3108675, 3114648, 3114677, 3117893
None
3117968, 3118105, 3118110, 3118113, 3118127, 3118162
None
2595868, 3076770
None
None
2442016, 2443591, 2444314, 2446508
None
3116214, 3116347, 3116674, 3116796, 3117307, 3117565, 3117952
None
3096921, 3108647, 3110424, 3112127, 3112377, 3117073
None
3117357
3114090, 3114415
3069745
None
3094629, 3106089
3081727, 3093094, 3106085, 3113008
None
3021113, 3024315, 3027978, 3076312, 3109453
None
None
3116330, 3117121, 3117206, 3117657
3116804
3117880, 3117975
None
1818427, 2397666
3116386, 3116683, 3116824, 3117258, 3117626
None
None
3114438, 3114647, 3116995
3084057, 3084525
None
None
```

Photos

sta > slovenian_photo_related.txt

```
3118477
3118477
3118477
3118725, 3119318, 3129949
3118725, 3118776, 3119318, 3129949
3118725, 3119318, 3129949
3118725, 3119318, 3129949
3122079, 3132061, 3142360, 3143005, 3143113, 3154615, 3162891, 3178598, 3188267,
None
None
None
None
None
None
None
None
None
None
None
None
None
None
None
None
None
None
None
None
None
None
None
None
None
None
None
None
None
```

Problem:
- some documents don’t have ground-truth links

Implementation Details

- Used sl_core_news_sm pipeline from spaCy for NER
- Original paper used word2vec word embeddings
  - No word2vec for Slovenian - used fasttext instead
  - Number of dimensions remained the same - 100
- SeedTopicMine uses several types of embedding, including PLM-based
  - We pre-computed embeddings for the training dataset using pre-trained SloBerta from HuggingFace

spaCy NER Pipeline

```python
nlp = spacy.load("sl_core_news_sm")

with open('./datasets/sta/compact/slovenian_texts.txt','r') as f:
    lines = f.readlines()

doc=nlp(lines[0])
ne = set([(x.text,x.label_) for x in doc.ents])
for entity, entity_type in ne:
    print((entity,entity_type))
```

✓ 1.9s

```
('evropskega sklada za okrepitev industrije čiste tehnologije', 'MISC')
('eu', 'ORG')
('joeja bidna', 'PER')
('von der leyen', 'PER')
('evropi', 'LOC')
('evropska komisija', 'ORG')
```

spaCy NER Pipeline (contd.)

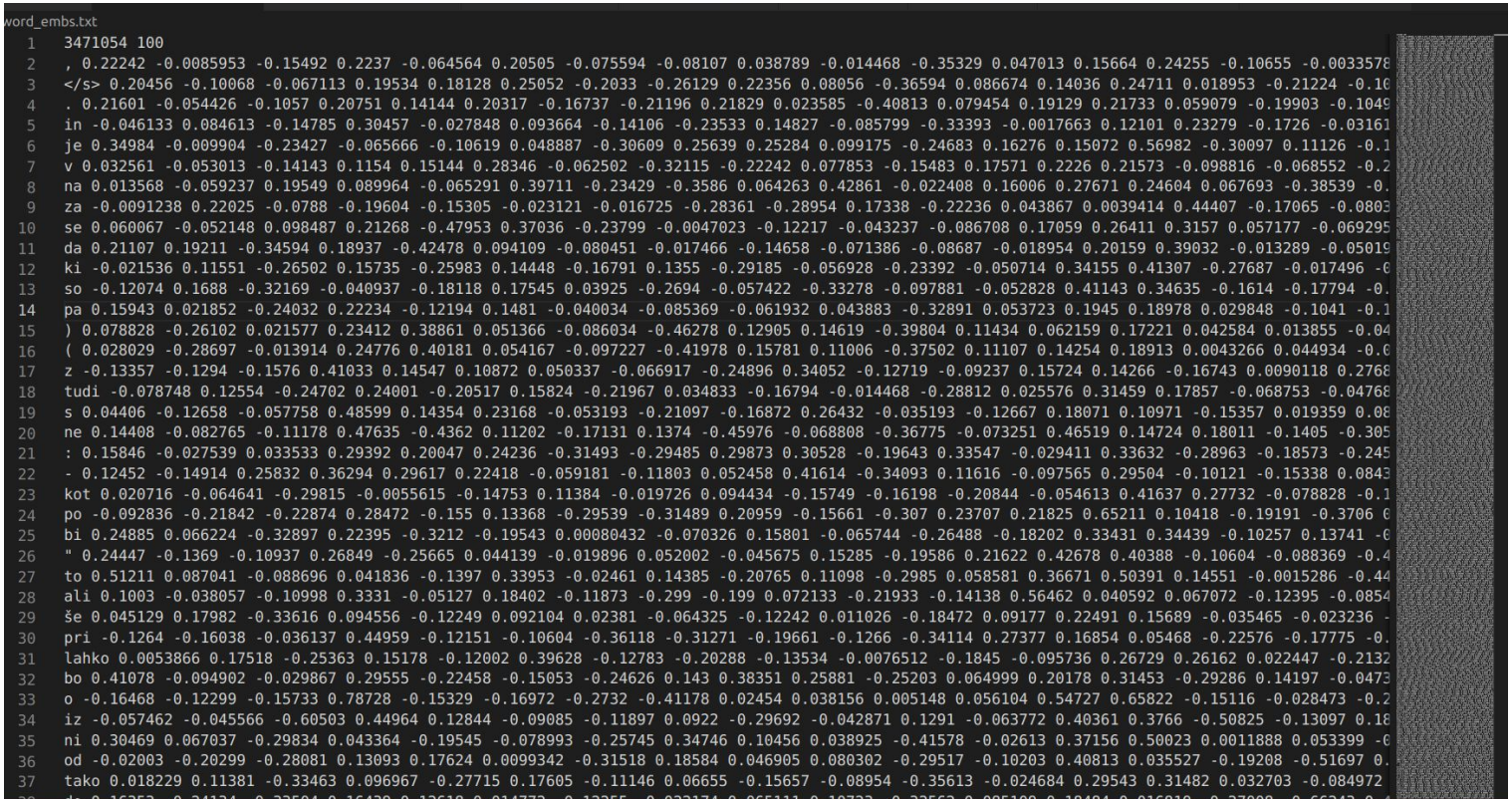
```
get_ner(980)
✓ 0.1s

('zda', 'LOC')
('washington', 'LOC')
('tvn24', 'ORG')
```

```
get_ner(10520)
✓ 0.0s

('srbija', 'LOC')
('oliver varhely', 'PER')
('eu', 'ORG')
('evropo', 'LOC')
('evropska banka', 'ORG')
```

```
get_ner(434)
✓ 0.0s

('andrej stare', 'PER')
('gregor', 'PER')
```

```
get_ner(1000)
✓ 0.0s
```

Fasttext word embeddings (dim=100)


word_embs.txt

```
3471054 100
, 0.22242 -0.0085953 -0.15492 0.2237 -0.064564 0.20505 -0.075594 -0.08107 0.038789 -0.014468 -0.35329 0.047013 0.15664 0.24255 -0.10655 -0.0033578
</s> 0.20456 -0.10068 -0.067113 0.19534 0.18128 0.25052 -0.2033 -0.26129 0.22356 0.08056 -0.36594 0.086674 0.14036 0.24711 0.018953 -0.21224 -0.10
. 0.21601 -0.054426 -0.1057 0.20751 0.14144 0.20317 -0.16737 -0.21196 0.21829 0.023585 -0.40813 0.079454 0.19129 0.21733 0.059079 -0.19903 -0.1049
in -0.046133 0.084613 -0.14785 0.30457 -0.027848 0.093664 -0.14106 -0.23533 0.14827 -0.085799 -0.33393 -0.0017663 0.12101 0.23279 -0.1726 -0.03161
je 0.34984 -0.009904 -0.23427 -0.065666 -0.10619 0.048887 -0.30609 0.25639 0.25284 0.099175 -0.24683 0.16276 0.15072 0.56982 -0.30097 0.11126 -0.1
v 0.032561 -0.053013 -0.14143 0.1154 0.15144 0.28346 -0.062502 -0.32115 -0.22242 0.077853 -0.15483 0.17571 0.2226 0.21573 -0.098816 -0.068552 -0.2
na 0.013568 -0.059237 0.19549 0.089964 -0.065291 0.39711 -0.23429 -0.3586 0.064263 0.42861 -0.022408 0.16006 0.27671 0.24604 0.067693 -0.38539 -0.
za -0.0091238 0.22025 -0.0788 -0.19604 -0.15305 -0.023121 -0.016725 -0.28361 -0.28954 0.17338 -0.22236 0.043867 0.0039414 0.44407 -0.17065 -0.0803
se 0.060067 -0.052148 0.098487 0.21268 -0.47953 0.37036 -0.23799 -0.0047023 -0.12217 -0.043237 -0.086708 0.17059 0.26411 0.3157 0.057177 -0.069295
da 0.21107 0.19211 -0.34594 0.18937 -0.42478 0.094109 -0.080451 -0.017466 -0.14658 -0.071386 -0.08687 -0.018954 0.20159 0.39032 -0.013289 -0.05019
ki -0.021536 0.11551 -0.26502 0.15735 -0.25983 0.14448 -0.16791 0.1355 -0.29185 -0.056928 -0.23392 -0.050714 0.34155 0.41307 -0.27687 -0.017496 -0
so -0.12074 0.1688 -0.32169 -0.040937 -0.18118 0.17545 0.03925 -0.2694 -0.057422 -0.33278 -0.097881 -0.052828 0.41143 0.34635 -0.1614 -0.17794 -0.
pa 0.15943 0.021852 -0.24032 0.22234 -0.12194 0.1481 -0.040034 -0.085369 -0.061932 0.043883 -0.32891 0.053723 0.1945 0.18978 0.029848 -0.1041 -0.1
) 0.078828 -0.26102 0.021577 0.23412 0.38861 0.051366 -0.086034 -0.46278 0.12905 0.14619 -0.39804 0.11434 0.062159 0.17221 0.042584 0.013855 -0.04
( 0.028029 -0.28697 -0.013914 0.24776 0.40181 0.054167 -0.097227 -0.41978 0.15781 0.11006 -0.37502 0.11107 0.14254 0.18913 0.0043266 0.044934 -0.0
z -0.13357 -0.1294 -0.1576 0.41033 0.14547 0.10872 0.050337 -0.066917 -0.24896 0.34052 -0.12719 -0.09237 0.15724 0.14266 -0.16743 0.0090118 0.2768
tudi -0.078748 0.12554 -0.24702 0.24001 -0.20517 0.15824 -0.21967 0.034833 -0.16794 -0.014468 -0.28812 0.025576 0.31459 0.17857 -0.068753 -0.04768
s 0.04406 -0.12658 -0.057758 0.48599 0.14354 0.23168 -0.053193 -0.21097 -0.16872 0.26432 -0.035193 -0.12667 0.18071 0.10971 -0.15357 0.019359 0.08
ne 0.14408 -0.082765 -0.11178 0.47635 -0.4362 0.11202 -0.17131 0.1374 -0.45976 -0.068808 -0.36775 -0.073251 0.46519 0.14724 0.18011 -0.1405 -0.305
: 0.15846 -0.027539 0.033533 0.29392 0.20047 0.24236 -0.31493 -0.29485 0.29873 0.30528 -0.19643 0.33547 -0.029411 0.33632 -0.28963 -0.18573 -0.245
- 0.12452 -0.14914 0.25832 0.36294 0.29617 0.22418 -0.059181 -0.11803 0.052458 0.41614 -0.34093 0.11616 -0.097565 0.29504 -0.10121 -0.15338 0.0843
kot 0.020716 -0.064641 -0.29815 -0.0055615 -0.14753 0.11384 -0.019726 0.094434 -0.15749 -0.16198 -0.20844 -0.054613 0.41637 0.27732 -0.078828 -0.1
po -0.092836 -0.21842 -0.22874 0.28472 -0.155 0.13368 -0.29539 -0.31489 0.20959 -0.15661 -0.307 0.23707 0.21825 0.65211 0.10418 -0.19191 -0.3706 0
bi 0.24885 0.066224 -0.32897 0.22395 -0.3212 -0.19543 0.00080432 -0.070326 0.15801 -0.065744 -0.26488 -0.18202 0.33431 0.34439 -0.10257 0.13741 -0
" 0.24447 -0.1369 -0.10937 0.26849 -0.25665 0.044139 -0.019896 0.052002 -0.045675 0.15285 -0.19586 0.21622 0.42678 0.40388 -0.10604 -0.088369 -0.4
to 0.51211 0.087041 -0.088696 0.041836 -0.1397 0.33953 -0.02461 0.14385 -0.20765 0.11098 -0.2985 0.058581 0.36671 0.50391 0.14551 -0.0015286 -0.44
ali 0.1003 -0.038057 -0.10998 0.3331 -0.05127 0.18402 -0.11873 -0.299 -0.199 0.072133 -0.21933 -0.14138 0.56462 0.040592 0.067072 -0.12395 -0.0854
še 0.045129 0.17982 -0.33616 0.094556 -0.12249 0.092104 0.02381 -0.064325 -0.12242 0.011026 -0.18472 0.09177 0.22491 0.15689 -0.035465 -0.023236 -
pri -0.1264 -0.16038 -0.036137 0.44959 -0.12151 -0.10604 -0.36118 -0.31271 -0.19661 -0.1266 -0.34114 0.27377 0.16854 0.05468 -0.22576 -0.17775 -0.
lahko 0.0053866 0.17518 -0.25363 0.15178 -0.12002 0.39628 -0.12783 -0.20288 -0.13534 -0.0076512 -0.1845 -0.095736 0.26729 0.26162 0.022447 -0.2132
bo 0.41078 -0.094902 -0.029867 0.29555 -0.22458 -0.15053 -0.24626 0.143 0.38351 0.25881 -0.25203 0.064999 0.20178 0.31453 -0.29286 0.14197 -0.0473
o -0.16468 -0.12299 -0.15733 0.78728 -0.15329 -0.16972 -0.2732 -0.41178 0.02454 0.038156 0.005148 0.056104 0.54727 0.65822 -0.15116 -0.028473 -0.2
iz -0.057462 -0.045566 -0.60503 0.44964 0.12844 -0.09085 -0.11897 0.0922 -0.29692 -0.042871 0.1291 -0.063772 0.40361 0.3766 -0.50825 -0.13097 0.18
ni 0.30469 0.067037 -0.29834 0.043364 -0.19545 -0.078993 -0.25745 0.34746 0.10456 0.038925 -0.41578 -0.02613 0.37156 0.50023 0.0011888 0.053399 -0
od -0.02003 -0.20299 -0.28081 0.13093 0.17624 0.0099342 -0.31518 0.18584 0.046905 0.080302 -0.29517 -0.10203 0.40813 0.035527 -0.19208 -0.51697 0.
tako 0.018229 0.11381 -0.33463 0.096967 -0.27715 0.17605 -0.11146 0.06655 -0.15657 -0.08954 -0.35613 -0.024684 0.29543 0.31482 0.032703 -0.084972
```

SeedTopicMine

- We used an official implementation of the SeedTopicMine algorithm
- Combines Python and C code
- We've modified the parts that were using English embeddings to work on Slovenian data
- Algorithm starts with a set of seeds
- At each iteration the framework adds one related word per seed
- Relatedness is assessed using similarities between embeddings computed in different contexts
- At each iteration only the words with the highest ranking score is selected
- Finally, each seed contains a predefined number of the most related words

Partial results

```
Read 10 topics
poljska
ukrajina
japonska
maroko
hrvaška
nemčija
amerika
evropa
romunija
francija
corpus size: 58332
Pre-training for 2 epochs, in total 2 + 10 = 12 epochs
Alpha: 0.018753  Progress: 24.99%  Words/thread/sec: 31.53k
Category (poljska):     poljska slovaška
Category (ukrajina):    ukrajina rusija
Category (japonska):    japonska kitajska
Category (maroko):      maroko tunizija
Category (hrvaška):     hrvaška srbska
Category (nemčija):     nemčija nemčijo
Category (amerika):     amerika oceanija
Category (evropa):      evropa milatovićevo
Category (romunija):    romunija bolgarija
Category (francija):    francija španija
Alpha: 0.016670  Progress: 33.33%  Words/thread/sec: 31.49k
Category (poljska):     poljska slovaška pap
Category (ukrajina):    ukrajina rusija kijev
Category (japonska):    japonska kitajska tajvanska
Category (maroko):      maroko tunizija tunizijo
Category (hrvaška):     hrvaška hina srbska
Category (nemčija):     nemčija nemčijo "nemčija
Category (amerika):     amerika oceanija karibi
Category (evropa):      evropa milatovićevo unija
Category (romunija):    romunija bolgarija latvija
Category (francija):    francija španija portugalska
```

Partial results (contd.)

```
Category (poljska):     poljska slovaška pap poljski duda varšava poljskem varšavi madžarska
Category (ukrajina):    ukrajina "agresor rusija kijev moskva ukrajinska bahmut protiofenzivo kasetnim
Category (japonska):    japonska kitajska tajvanska ameriška južnokorejska japonske japonski xinhua avstralska
Category (maroko):      maroko tunizija bahrajn mavretanija honduras katar tunizijo marokom kuvajt
Category (hrvaška):     hrvaška hina srbska tasr tanjug poročala pisala hine varaždinska
Category (nemčija):     nemčija rheinmetall "nemčija nemčijo scholz nemških berlina nemško habeck
Category (amerika):     amerika oceanija karibi oceanijo azija karibe podsaharska latinska kontinent
Category (evropa):      evropa milatovićevo prosperitetna "moldavija balkan: vključenost" unija vladavina zgojznik
Category (romunija):    romunija estonija bolgarija latvija belgija makedonija litva bosna konjičanin
Category (francija):    francija španija nizozemska češka švica portugalska grčija danska irska
Alpha: 0.002090  Progress: 91.65%  Words/thread/sec: 30.56k
Category (poljska):     poljska slovaška pap poljski madžarska duda varšava poljskem varšavi poljsko
Category (ukrajina):    ukrajina "agresor rusija kijev moskva ukrajinska kasetnim bahmut protiofenzivo "ukrajina
Category (japonska):    japonska kitajska tajvanska ameriška južnokorejska japonske japonski xinhua avstralska kitajski
Category (maroko):      maroko tunizija bahrajn mavretanija honduras katar tunizijo marokom kuvajt turčijo
Category (hrvaška):     hrvaška hina srbska tasr varaždinska tanjug pisala poročala hine italijanska
Category (nemčija):     nemčija rheinmetall "nemčija nemčijo scholz nemških nemško berlina habeck tankov
Category (amerika):     amerika oceanija karibi oceanijo azija podsaharska karibe kontinent latinska afrika
Category (evropa):      evropa milatovićevo prosperitetna "moldavija balkan: vključenost" unija zgojznik vladavina renew
Category (romunija):    romunija estonija bolgarija latvija belgija makedonija litva konjičanin bosna srbija
Category (francija):    francija španija švica češka nizozemska portugalska grčija danska irska malta
Alpha: 0.000007  Progress: 99.98%  Words/thread/sec: 30.37k
Category (poljska):     poljska slovaška pap poljski madžarska duda varšava poljsko poljskem varšavi madžarski
Category (ukrajina):    ukrajina "agresor rusija kijev moskva kasetnim ukrajinska bahmut protiofenzivo "ukrajina ukrajine
Category (japonska):    japonska kitajska tajvanska ameriška južnokorejska japonske japonski xinhua avstralska kitajski bidnova
Category (maroko):      maroko tunizija bahrajn mavretanija honduras katar tunizijo marokom kuvajt turčijo egipt
Category (hrvaška):     hrvaška hina srbska tasr varaždinska tanjug pisala poročala hine italijanska mti
Category (nemčija):     nemčija rheinmetall "nemčija nemčijo scholz nemških nemško berlina tankov habeck proizvajalec
Category (amerika):     amerika oceanija karibi oceanijo podsaharska azija karibe kontinent latinska afrika amerike
Category (evropa):      evropa milatovićevo prosperitetna "moldavija balkan: vključenost" unija zgojznik vladavina renew "postati
Category (romunija):    romunija estonija bolgarija latvija belgija makedonija litva konjičanin bosna srbija hercegovina
Category (francija):    francija španija švica nizozemska češka grčija portugalska danska malta irska luksemburg
Topic mining results written to file ./datasets/sta/res_locations.txt
```

Difficulties & Remaining Work

- We require not just the related words, but also the ids of the documents where they are located - ‘links’
- Scale of required changes in the C code appeared to be far greater than expected
- We planned to use the ground-truth links for prediction evaluation
- Paper authors were evaluating term accuracy of the predicted words with the help of independent annotators
- Annotators were telling whether a prediction is relevant or not and these assessments were used for metrics calculation (P@k, NDCG@k)
- Similar approach may be required in our experiments for predictions on the documents without ground-truth links

Thank you for attention!