



DEPARTMENT OF
**KNOWLEDGE
TECHNOLOGIES**

Jožef Stefan Institute

Ontogen: Application on press agencies' articles about Slovenia



<http://kt.ijs.si>

Matjaž Juršič
Mentor: dr. Nada Lavrač

Initial problem

- What is the public opinion of Slovenia in foreign countries.



Presumption

- Most foreigners never visit Slovenia.
- Their opinions are based on reports of media in their countries.
- Media (TV, radio, newspapers) are mainly reporting just what is published by official press agency in that country.
- Press agency of foreign country selects interesting news from all that STA publishes.



Information flow

Something happens in Slovenia

STA reports about it

Official report about event

Opinion of the news and Slovenia in general

People hear about it

If interesting for them, foreign press agency translates the news

Reports about event in foreign local media

Published

Some event in Slovenia



Data

- All STA published news (18.676)
(slovene articles)
- News published by press agencies in
other countries. (2.711 or 14.5%)
(foreign articles)



Data

slovene articles

Something happens in Slovenia

STA reports about it

Official report about event

Opinion of the news and Slovenia in general

People hear about it

If interesting for them, foreign press agency translates the news

foreign articles

Reports about event in foreign local media

Published

Some event in Slovenia

New problem

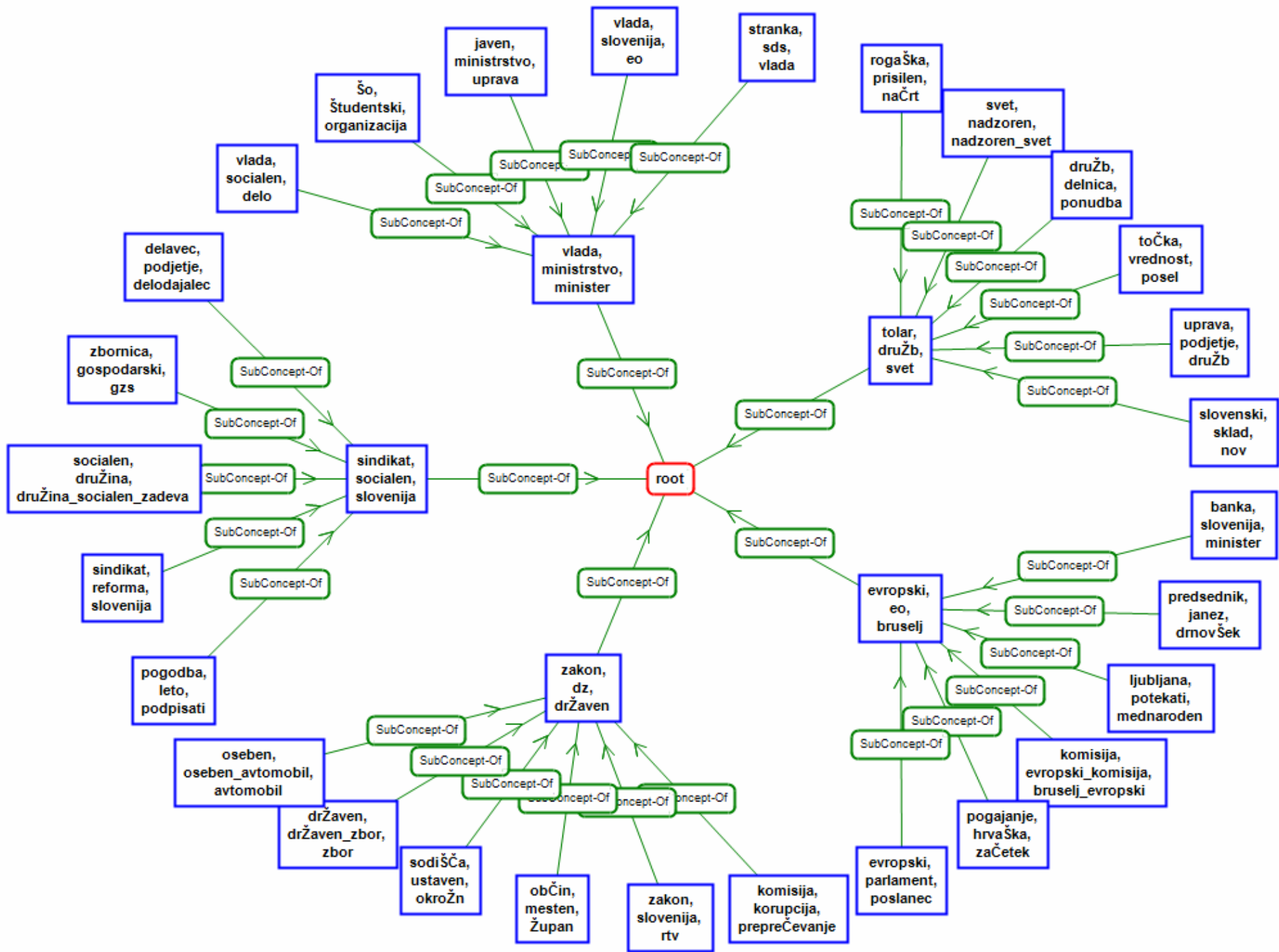
- No questioning foreigners of their opinion.
- We just look at the information they were given.
- We can build our own picture of Slovenia with given data and expect their opinion is similar.

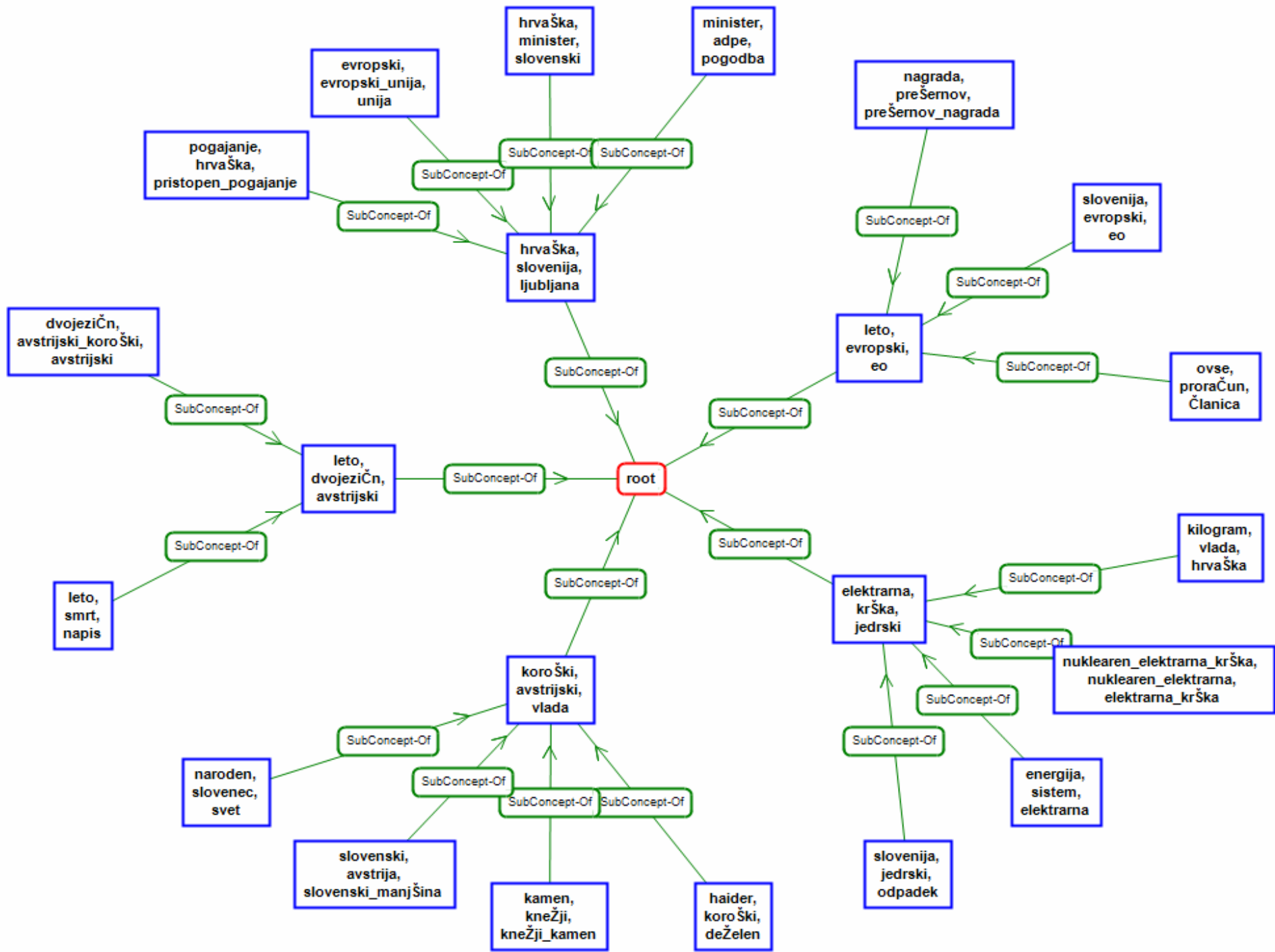


Ontogen

- How can we build “picture” of Slovenia from 1000 articles?
- Practically impossible to read all of them.
- Solution: ONTOGEN
- We create ontology of topics from foreign articles.
- We can compare this ontology with ontology from slovene articles or even with other countries and observe how focus changes on different topics in each









DEPARTMENT OF
**KNOWLEDGE
TECHNOLOGIES**

Jožef Stefan Institute

Ontogen: Data preparation



<http://kt.ijs.si>

Matjaž Juršič
Mentor: dr. Nada Lavrač

Data preparation

- 1) Correct file format
- 2) Lemmatization or stemming and stop word removal

For certain languages already included in Ontogen



Ontogen input formats

- Named Line-Documents
 - Each line of the file is new instance
 - All text instances are stored in one file
- Folder
 - Loads all files in given folder and optionally subfolders
- Bag-Of-Words
 - Text Garden format



Example of Named Line-Doc

006657 !SLO !ALL VRHNIKA Upnik Industrija usnje Vrh...
006665 !SLO !ALL LJUBLJANA Mercator sklad terminski...
006673 !SLO !ALL LJUBLJANA Potekati predstaviteva p...
018193 !SCG !ALL LJUBLJANA Ena glavnih Sloveniji Lj...
018195 !SCG !ALL slovenskem časniku Finance povzela...
018196 !SCG !ALL pisala zunanji minister Rupel nujn...
021227 !HRV !ALL nedeljo pisala predsednik republik...
021229 !HRV !ALL soboto povzela izjavo slovenski te...
016720 !FRA !ALL LJUBLJANA pozornosti namenile pog...
016725 !FRA !ALL povzela izjavo osebnega predstavn...



Example of Named Line-Doc

006657 !SLO !ALL VRHNIKA Upnik Industrija usnje Vrh...
006665 !SLO !ALL LJUBLJANA Mercator sklad terminski...
006673 !SLO !ALL LJUBLJANA Potekati predstaviteva p...
018193 !SCG !ALL LJUBLJANA Ena glavnih Sloveniji Lj...
018195 !SCG !ALL slovenskem časniku Finance povzela...
018196 !SCG !ALL pisala zunanji minister Rupel nujn...
021227 !HRV !ALL nedeljo pisala predsednik republik...
021229 !HRV !ALL soboto povzela izjavo slovenski te...
016720 !FRA !ALL LJUBLJANA pozornosti namenile pog...
016725 !FRA !ALL povzela izjavo osebnega predstavn...

First word of line stands for title of current text instance. Sth like: This_is_the_title.



Example of Named Line-Doc

006657 !SLO !ALL VRHNIKA Upnik Industrija usnje Vrh...
006665 !SLO !ALL LJUBLJANA Mercator sklad terminski...
006673 !SLO !ALL LJUBLJANA Potekati predstaviteva p...
018193 !SCG !ALL LJUBLJANA Ena glavnih Sloveniji Lj...
018195 !SCG !ALL slovenskem časniku Finance povzela...
018196 !SCG !ALL pisala zunanji minister Rupel nujn...
021227 !HRV !ALL nedeljo pisala predsednik republik...
021229 !HRV !ALL soboto povzela izjavo slovenski te...
016720 !FRA !ALL LJUBLJANA pozornosti namenile pog...
016725 !FRA !ALL povzela izjavo osebnega predstavn...

Optionally: instance labels. Each starting with ! (exclamation mark).



Example of Named Line-Doc

```
006657 !SLO !ALL VRHNIKA Upnik Industrija usnje Vrh...
006665 !SLO !ALL LJUBLJANA Mercator sklad terminski...
006673 !SLO !ALL LJUBLJANA Potekati predstaviteva p...
018193 !SCG !ALL LJUBLJANA Ena glavnih Sloveniji Lj...
018195 !SCG !ALL slovenskem časniku Finance povzela...
018196 !SCG !ALL pisala zunanji minister Rupel nujn...
021227 !HRV !ALL nedeljo pisala predsednik republik...
021229 !HRV !ALL soboto povzela izjavo slovenski te...
016720 !FRA !ALL LJUBLJANA pozornosti namenile pog...
016725 !FRA !ALL povzela izjavo osebnega predstavn...
```

**After labels or title and to the end of the line
there is instance text.**



Lemmatization or stemming

- This is the process of converting all words in text to some neutral form.
- Example: cats, catty or catlike to cat
- If your current version of Ontogen doesn't support stemmer for desired language, try downloading new version.
- Else you must find algorithm and manually run it through your texts. This is very important for good results!!



Stop word removal

- Stop words are those words which are so common that they are useless to text mining algorithms. In English some obvious stop words would be "a", "of", "the", "I", "it", "you", and "and".
- Again, if Ontogen doesn't support your language you should find a list of stop words and somehow remove them manually.

