

Extraction of Implicit Networks from Unstructured Text in R

Advanced Software Practical (FoPra)

Katja Hauser

7th August 2018

Extraction of Implicit Networks in R

Implement the LOAD algorithm in R and embed it into a highly modular pipeline.

Extraction of Implicit Networks in R

Implement the LOAD algorithm in R and embed it into a highly modular pipeline.



Extraction of Implicit Networks in R

Implement the LOAD algorithm in R and embed it into a highly modular pipeline.

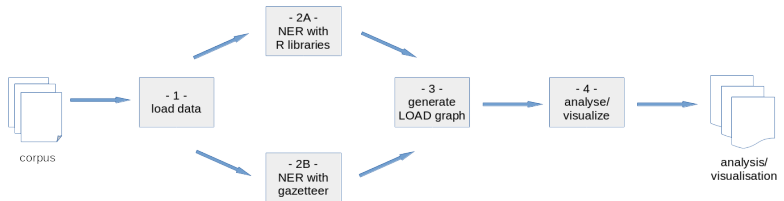


R offers:

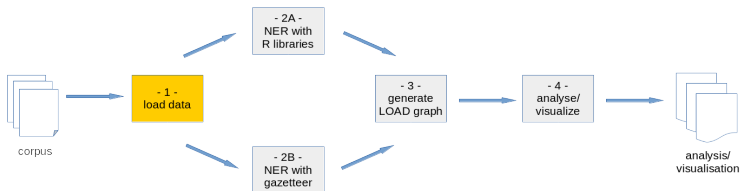
- modular packet structure
- easy redistribution of code via CRAN
- already existing libraries (esp. statistical analysis and network analysis)
- *tidy data* tools

Graphic from "Terms over LOAD: Leveraging Named Entities for Cross-Document Extraction and Summarization of Events", Andreas Spitz, and Michael Gertz, ACM, 2016, p. 6

The Pipeline



Functionality of the Pipeline



Loads data from different sources.

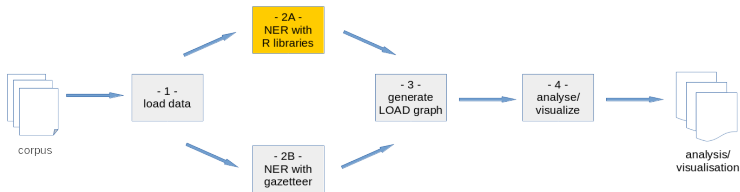
Extensions:

Implemented:

- use txt format

- other formats (csv, json, etc)
- other data sources (data bases)

Functionality of the Pipeline



Create annotations using already existing R libraries.

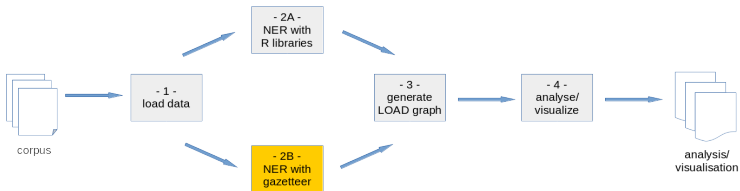
Implemented:

- use *openNLP*

Extensions:

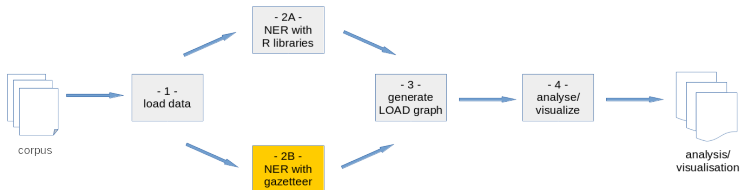
- other libraries

Functionality of the Pipeline



Create annotations using a gazetteer.

Functionality of the Pipeline

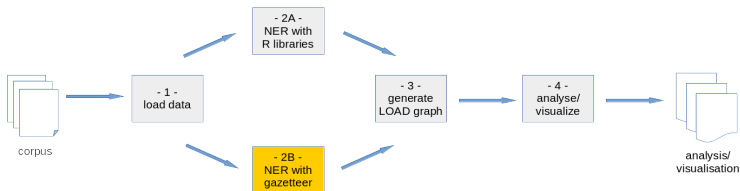


Create annotations using a gazetteer.

Implemented:

- simple sentence splitting
- simple tokenization
- remove stop words
- morphological forms
(genitive and plural s)

Functionality of the Pipeline



Create annotations using a gazetteer.

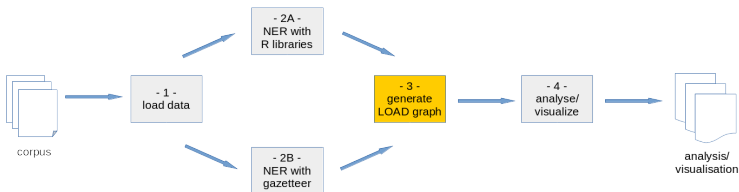
Implemented:

- simple sentence splitting
- simple tokenization
- remove stop words
- morphological forms (genitive and plural s)

Extensions:

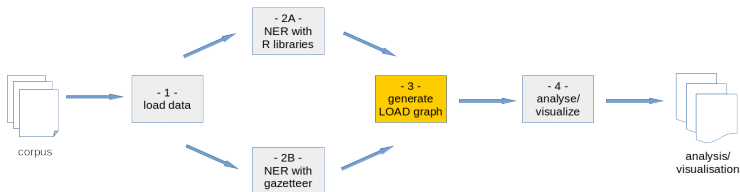
- more sophisticated approaches to sentence splitting, tokenization and creation of morphological forms

Functionality of the Pipeline



Calculate the LOAD network from given annotations.

Functionality of the Pipeline

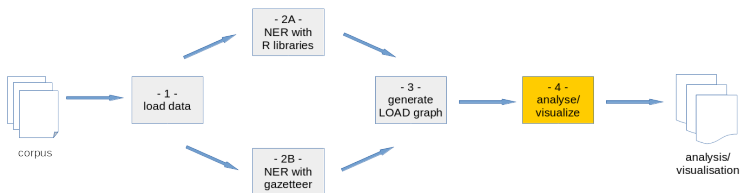


Calculate the LOAD network from given annotations.

No extensions planned, but the user can set:

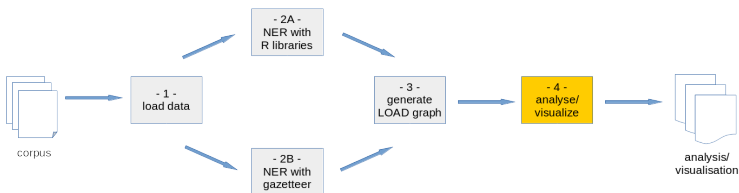
- maximal distance of cooccurrences
- weight edges between named entities and terms?
- allow edges between named entities of the same type?

Functionality of the Pipeline



Analyse or visualize the implicit network.

Functionality of the Pipeline

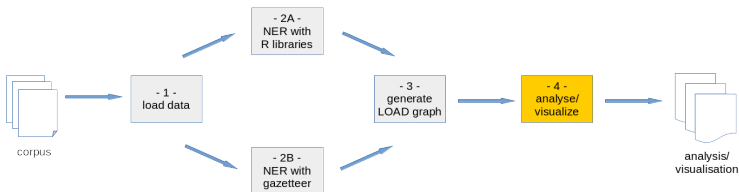


Analyse or visualize the implicit network.

Implemented:

- conversion to *igraph* object
- simple visualization
- simple analysis

Functionality of the Pipeline



Analyse or visualize the implicit network.

Implemented:

- conversion to *igraph* object
- simple visualization
- simple analysis

Extensions:

- more elaborated analysis
- visualization of large networks with other tools

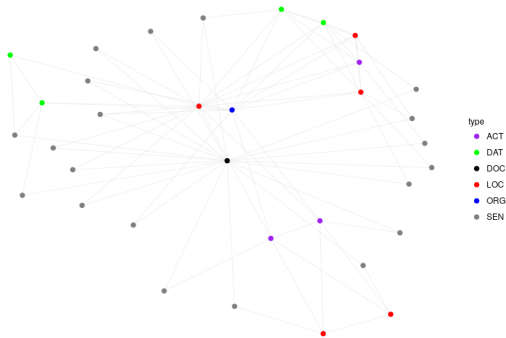
Exemplary Use Case 1

Extract the implicit network of named entities and terms from Jules Verne's *Around the World in 80 Days* and find the most important nodes by PageRank.

passepartout	fix	aouda	london	bombay	fogg
0.095678005	0.049369987	0.013413588	0.008156093	0.007751962	0.007361383

Exemplary Use Case 2

Visualize the LOAD network of the named entities in the first chapter of *Around the World in 80 Days*.



Runtime Measurement

	# docs	# nodes	# edges	gaz. or R lib.	edges NE-TER	viz or analysis
1	37	8,369	143,697	library	weighted	analysis
2	37	8,369	66,493	library	unweighted	analysis
3	37	8,128	89,762	gazetteer	weighted	analysis
4	1	32	84	library	no terms	visualisation
5	1	34	79	gazetteer	no terms	visualisation

Runtime Measurement

	# docs	# nodes	# edges	gaz. or R lib.	edges NE-TER	viz or analysis
1	37	8,369	143,697	library	weighted	analysis
2	37	8,369	66,493	library	unweighted	analysis
3	37	8,128	89,762	gazetteer	weighted	analysis
4	1	32	84	library	no terms	visualisation
5	1	34	79	gazetteer	no terms	visualisation

runtime in seconds

	load corpus	library/gazetteer	LOAD network	analysis/viz	runtime total
1	0.0985 \pm 0.0027	305.1000 \pm 0.5977	317.1600 \pm 0.1576	0.7527 \pm 0.0031	623.1120 \pm 1.0427
2	0.0756 \pm 0.0004	299.9400 \pm 0.2606	123.9600 \pm 0.0069	0.3935 \pm 0.0002	424.3691 \pm 0.3180
3	0.0955 \pm 0.0004	6.5800 \pm 0.0345	2.4300 \pm 0.0005	0.5120 \pm 0.0027	9.6200 \pm 0.0520
4	0.0261 \pm 0.0019	4.5000 \pm 0.3940	0.1860 \pm 0.0169	0.1910 \pm 0.0906	4.9100 \pm 0.9080
5	0.0324 \pm 0.0032	0.0785 \pm 0.0015	0.1750 \pm 0.0136	0.1840 \pm 0.0806	0.4700 \pm 0.1890

Conclusion

Future Work

- extend modules (creation of annotations, analysis, etc.)
- wrapper for external annotations
- named entity disambiguation

Conclusion

Future Work

- extend modules (creation of annotations, analysis, etc.)
- wrapper for external annotations
- named entity disambiguation

Thank you for your attention!