Word2Vec and Metadata Records

Felix Sasaki

April 2019

Overview

- Motivation
- Aspects of word2vec
- Processing of metadata records: input and pre-preprocessing
- Metadata records and word2vec: example output & optimization
- Application: recommendation based on metadata semantic similarity
- Next steps

Overview

- Motivation
- Aspects of word2vec
- Processing of metadata records: input and pre-preprocessing
- Metadata records and word2vec: example output & optimization
- Application: recommendation based on metadata semantic similarity
- Next steps

Motivation – from the perspective of applications in digital libraries

- Libraries usually provide metadata records e.g. to facility keyword based search
- Similarity of metadata records is hard to compute on a semantic level
    - “Germany” is closely related to “Europe”
    - This relation is not explicit in metadata records

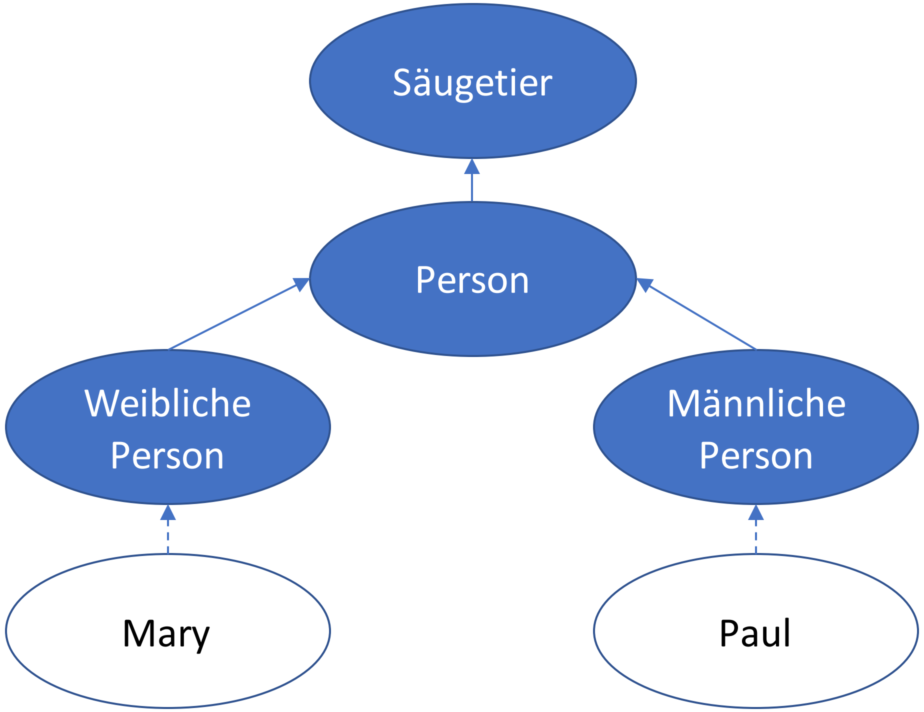
Motivation – from the perspective of applications in digital libraries

- word2vec is an unsupervised learning approach that calculates semantic similarity
- Goals
  - Apply word2vec to metadata records
  - Evaluate outcome using various word2vec settings
  - Show example application: word2vec based recommender system

Motivation – from the perspective of applied computational linguistics research

- Symbolic AI

  Knowledge crafted by humans



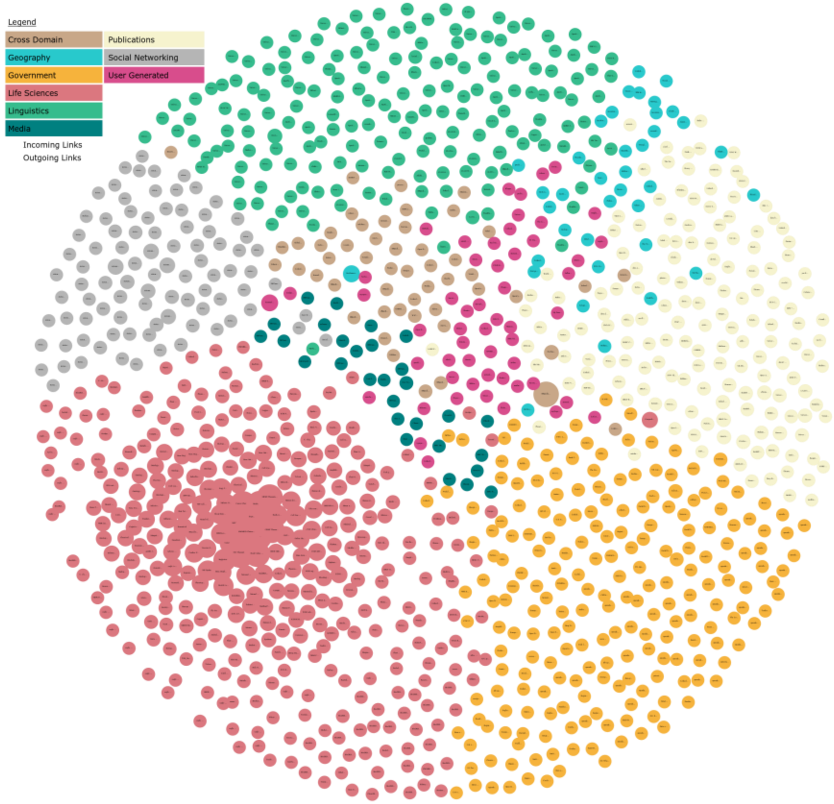
- **Machine Learning**

  Example: computer learns „what is a cat“?

Source: https://arxiv.org/pdf/1112.6209v3.pdf

Motivation – from the perspective of applied computational linguistics research

Linked (open) data sources



Input to machine learning

Overview

- Motivation
- Aspects of word2vec
- Processing of metadata records: input and pre-preprocessing
- Metadata records and word2vec: example output & optimization
- Application: recommendation based on metadata semantic similarity
- Next steps

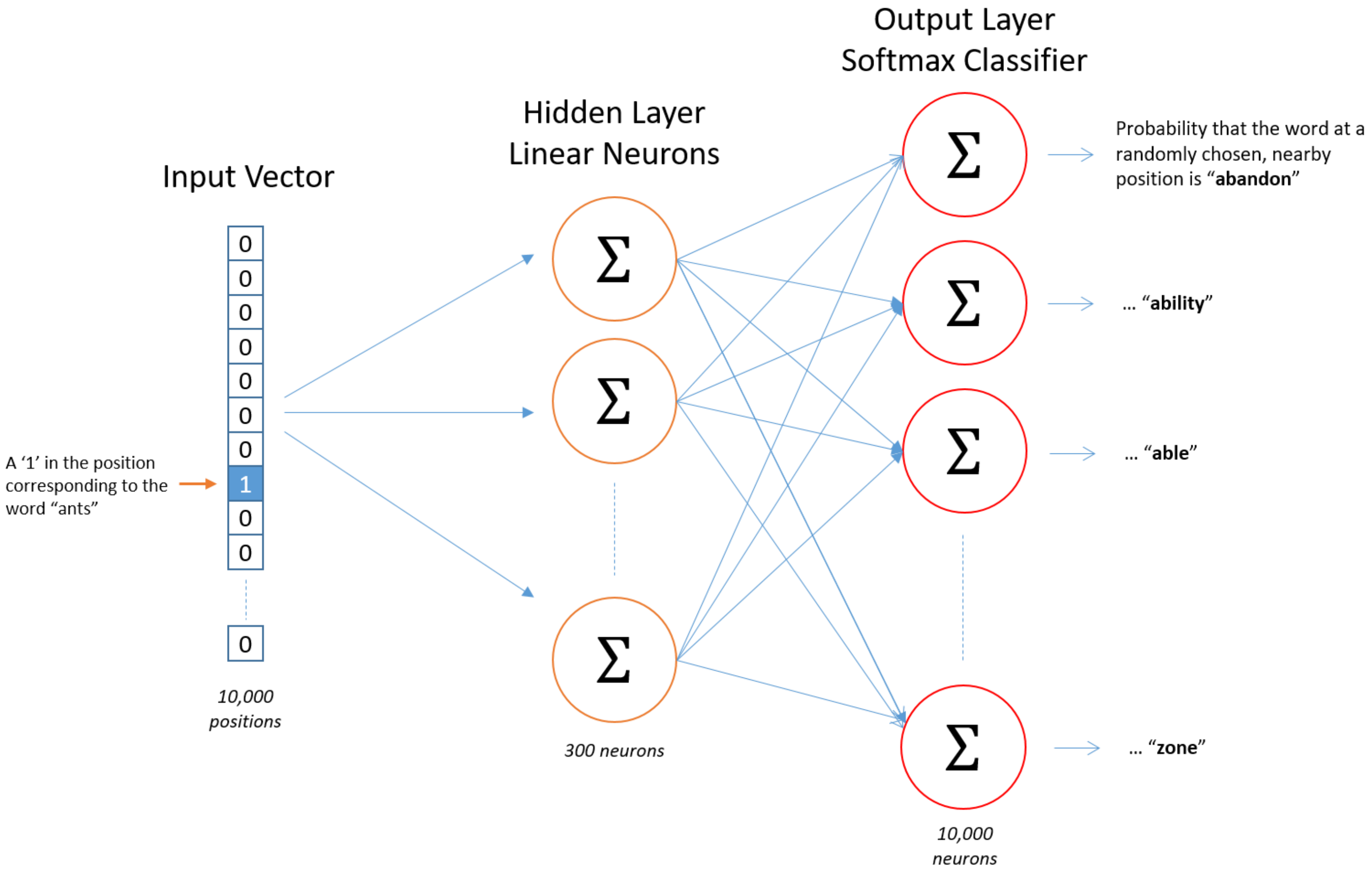
word2vec

- Unsupervised machine learning approach
- Used to calculate word embeddings
  - Representation of a vocabulary of words as numeric vectors
- Pivotal paper: Mikolov et al. (2013)
  - “Efficient Estimation of Word Representations in Vector Space”. In: proceedings of International Conference on Learning Representations 2013
- Basis: neural network – but not learning of weights for creating output
  - Learning of weights as the means to represent words
  - Using a single hidden layer
  - Similar to autoencoder approach

word2vec task

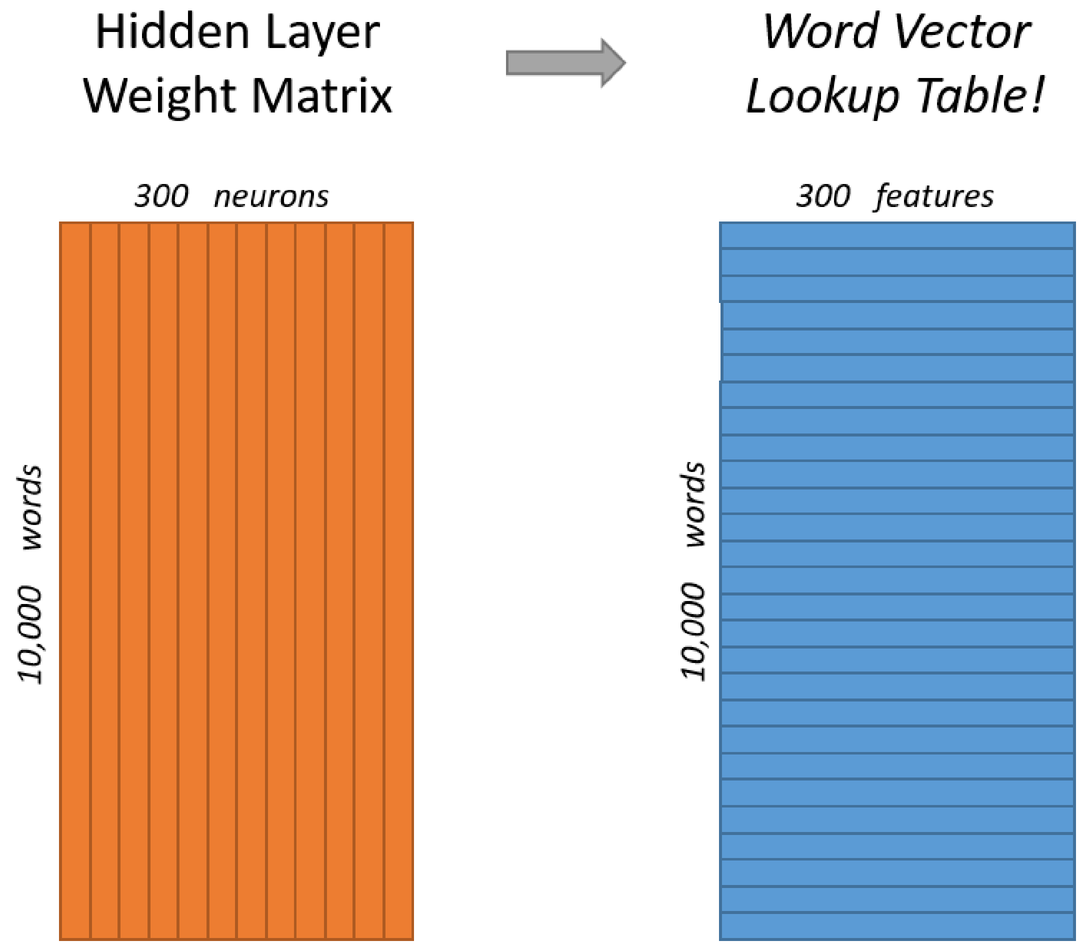
- Input: set of sentences
- Each word in sentence is processed with regards to neighbouring words – the window size
- Example sentence with window size 2
  - “The <u>cat</u> sat on the mat.”
    - the: the, cat; the, sat
    - cat: the, cat; cat, sat; cat, on
    - sat: the, sat; cat, sat; sat, on; sat, the
    - …
- Similarity = higher for words with the same context
  - “The <u>dog</u> sat on the mat.”

word2vec architecture



Source http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Learning in word2vec: using hidden layer as lookup table



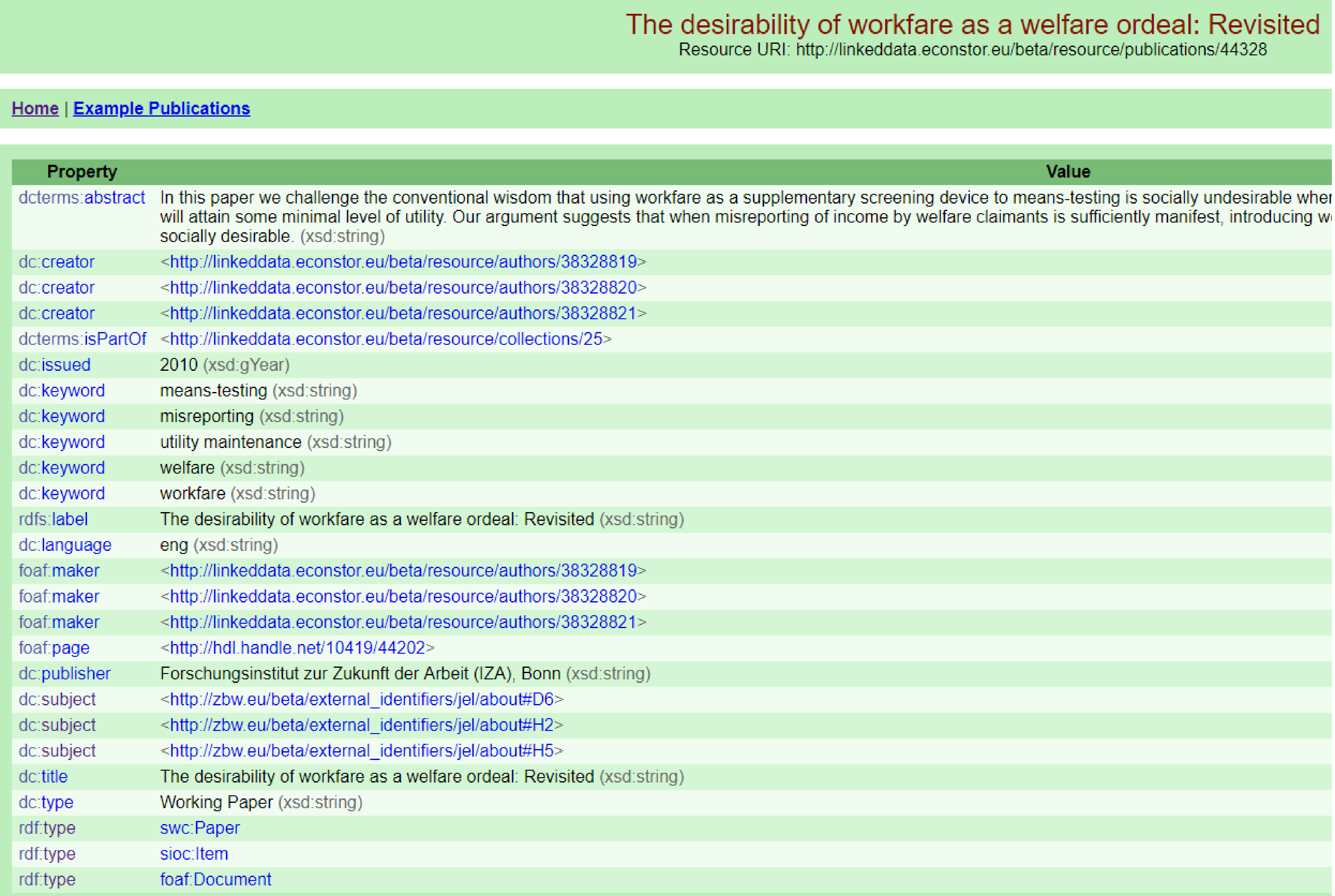
Source http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Overview

- Motivation
- Aspects of word2vec
- Processing of metadata records: input and pre-preprocessing
- Metadata records and word2vec: example output & optimization
- Application: recommendation based on metadata semantic similarity
- Next steps

Processing of metadata records: input

- Using metadata available as linked data via EconStor LOD http://zbw.eu/labs/en/project/econstor-lod
    - Basis: Data from Leibniz Information Centre for Economics (ZBW)
    - Available as linked data
    - Keywords in German and English
    - Example record, see next slide
- Processing of keywords, not textual description
- Approach: keywords in similar context have related meanings

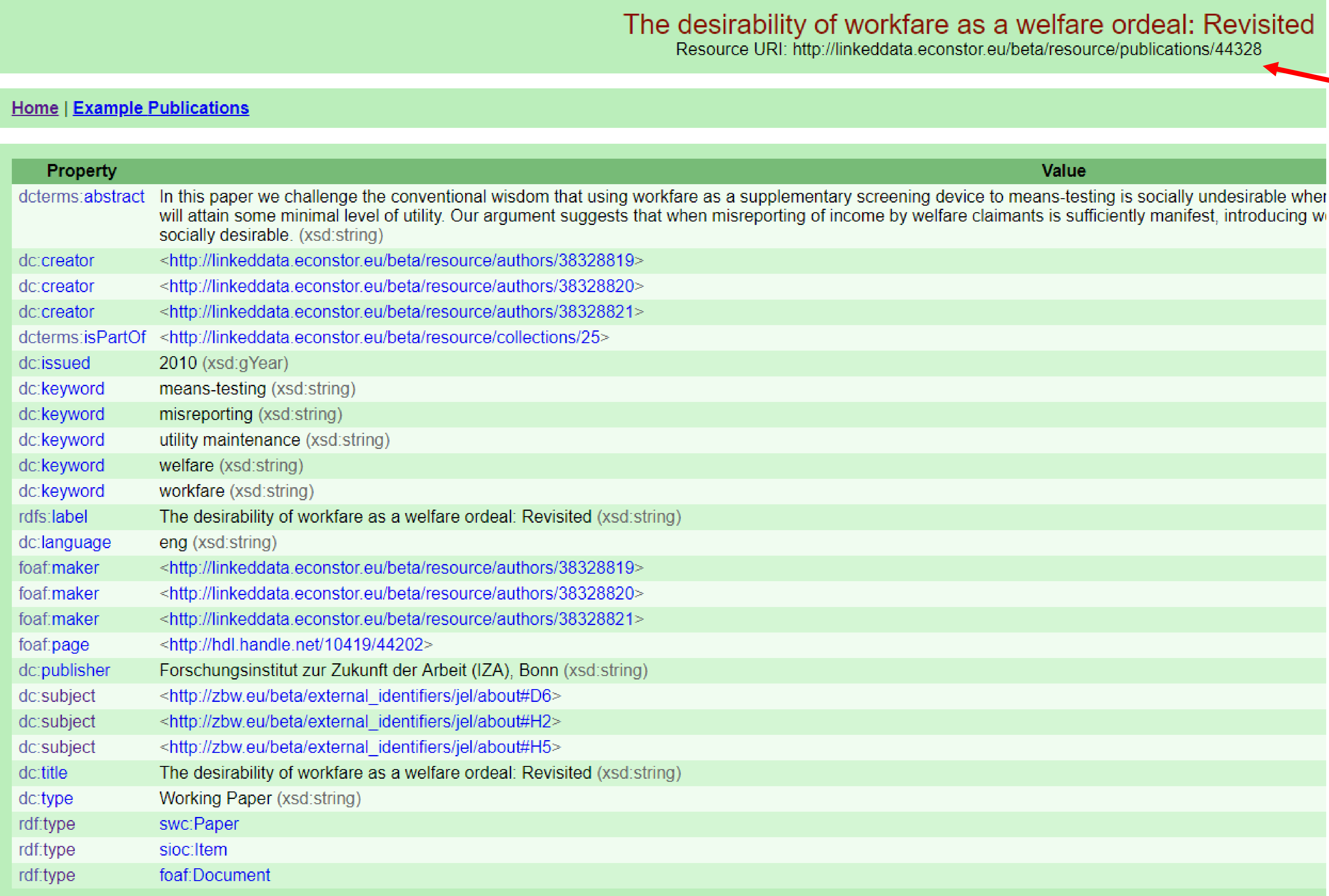


The desirability of workfare as a welfare ordeal: Revisited

Resource URI: http://linkeddata.econstor.eu/beta/resource/publications/44328

Home | Example Publications

Property	Value
dcterms:abstract	In this paper we challenge the conventional wisdom that using workfare as a supplementary screening device to means-testing is socially undesirable whe will attain some minimal level of utility. Our argument suggests that when misreporting of income by welfare claimants is sufficiently manifest, introducing w socially desirable. (xsd:string)
dc:creator	<http://linkeddata.econstor.eu/beta/resource/authors/38328819>
dc:creator	<http://linkeddata.econstor.eu/beta/resource/authors/38328820>
dc:creator	<http://linkeddata.econstor.eu/beta/resource/authors/38328821>
dcterms:isPartOf	<http://linkeddata.econstor.eu/beta/resource/collections/25>
dc:issued	2010 (xsd:gYear)
dc:keyword	means-testing (xsd:string)
dc:keyword	misreporting (xsd:string)
dc:keyword	utility maintenance (xsd:string)
dc:keyword	welfare (xsd:string)
dc:keyword	workfare (xsd:string)
rdfs:label	The desirability of workfare as a welfare ordeal: Revisited (xsd:string)
dc:language	eng (xsd:string)
foaf:maker	<http://linkeddata.econstor.eu/beta/resource/authors/38328819>
foaf:maker	<http://linkeddata.econstor.eu/beta/resource/authors/38328820>
foaf:maker	<http://linkeddata.econstor.eu/beta/resource/authors/38328821>
foaf:page	<http://hdl.handle.net/10419/44202>
dc:publisher	Forschungsinstitut zur Zukunft der Arbeit (IZA), Bonn (xsd:string)
dc:subject	<http://zbw.eu/beta/external_identifiers/jel/about#D6>
dc:subject	<http://zbw.eu/beta/external_identifiers/jel/about#H2>
dc:subject	<http://zbw.eu/beta/external_identifiers/jel/about#H5>
dc:title	The desirability of workfare as a welfare ordeal: Revisited (xsd:string)
dc:type	Working Paper (xsd:string)
rdf:type	swc:Paper
rdf:type	sioc:Item
rdf:type	foaf:Document



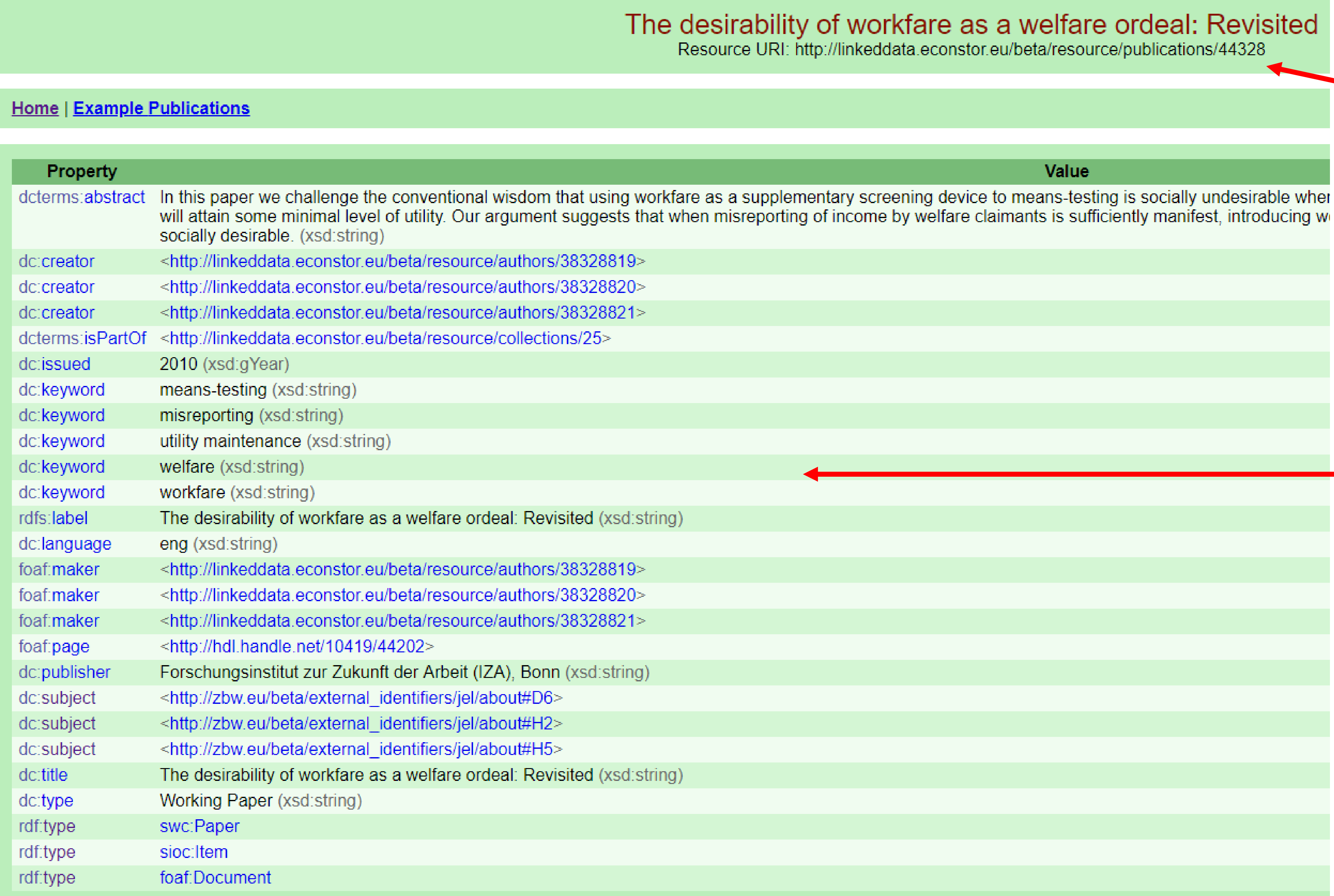
The desirability of workfare as a welfare ordeal: Revisited

Resource URI: http://linkeddata.econstor.eu/beta/resource/publications/44328

Home | Example Publications

Property	Value
dcterms:abstract	In this paper we challenge the conventional wisdom that using workfare as a supplementary screening device to means-testing is socially undesirable wher will attain some minimal level of utility. Our argument suggests that when misreporting of income by welfare claimants is sufficiently manifest, introducing w socially desirable. (xsd:string)
dc:creator	<http://linkeddata.econstor.eu/beta/resource/authors/38328819>
dc:creator	<http://linkeddata.econstor.eu/beta/resource/authors/38328820>
dc:creator	<http://linkeddata.econstor.eu/beta/resource/authors/38328821>
dcterms:isPartOf	<http://linkeddata.econstor.eu/beta/resource/collections/25>
dc:issued	2010 (xsd:gYear)
dc:keyword	means-testing (xsd:string)
dc:keyword	misreporting (xsd:string)
dc:keyword	utility maintenance (xsd:string)
dc:keyword	welfare (xsd:string)
dc:keyword	workfare (xsd:string)
rdfs:label	The desirability of workfare as a welfare ordeal: Revisited (xsd:string)
dc:language	eng (xsd:string)
foaf:maker	<http://linkeddata.econstor.eu/beta/resource/authors/38328819>
foaf:maker	<http://linkeddata.econstor.eu/beta/resource/authors/38328820>
foaf:maker	<http://linkeddata.econstor.eu/beta/resource/authors/38328821>
foaf:page	<http://hdl.handle.net/10419/44202>
dc:publisher	Forschungsinstitut zur Zukunft der Arbeit (IZA), Bonn (xsd:string)
dc:subject	<http://zbw.eu/beta/external_identifiers/jel/about#D6>
dc:subject	<http://zbw.eu/beta/external_identifiers/jel/about#H2>
dc:subject	<http://zbw.eu/beta/external_identifiers/jel/about#H5>
dc:title	The desirability of workfare as a welfare ordeal: Revisited (xsd:string)
dc:type	Working Paper (xsd:string)
rdf:type	swc:Paper
rdf:type	sioc:Item
rdf:type	foaf:Document

Canonical URI



The desirability of workfare as a welfare ordeal: Revisited

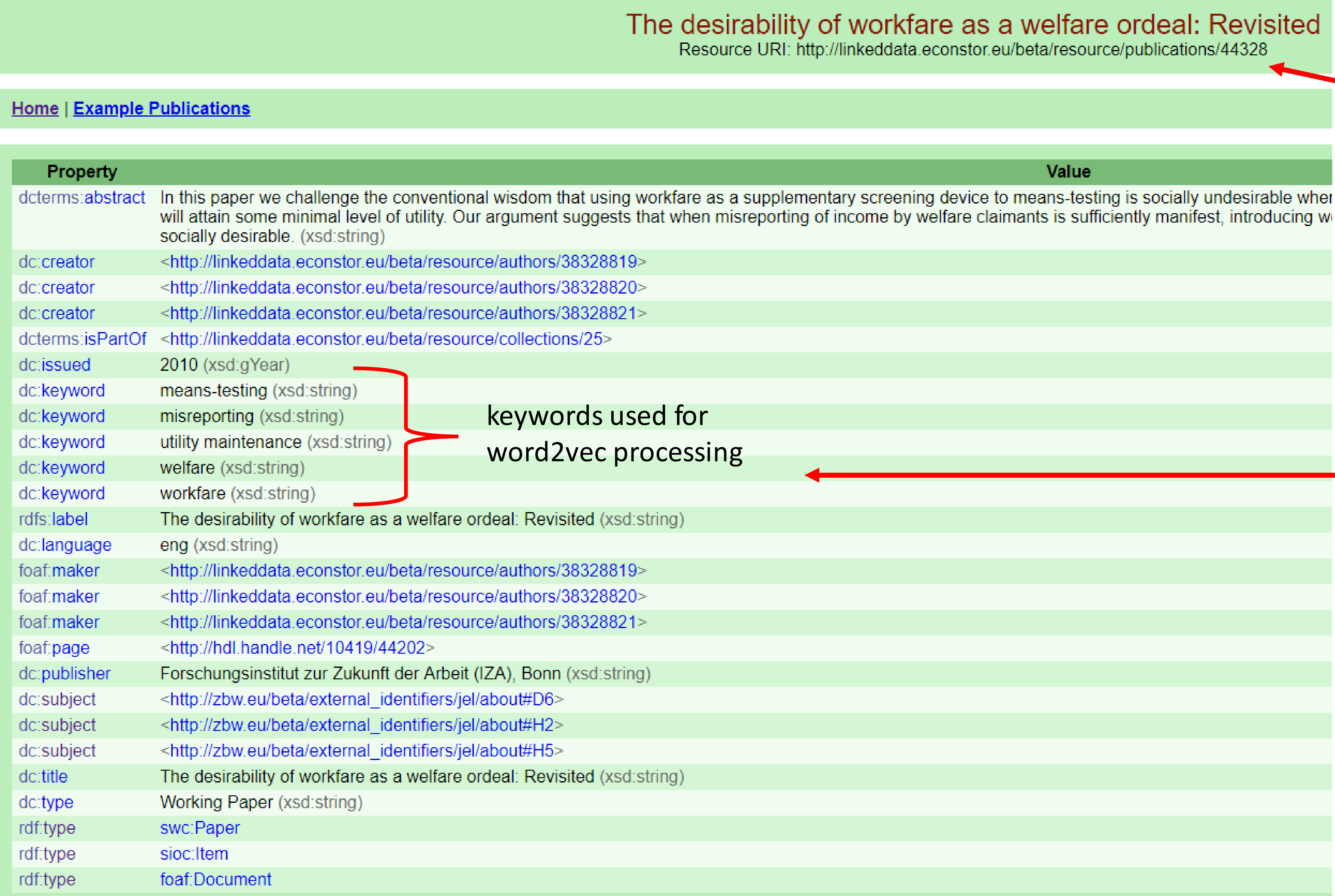
Resource URI: http://linkeddata.econstor.eu/beta/resource/publications/44328

Home | Example Publications

Property	Value
dcterms:abstract	In this paper we challenge the conventional wisdom that using workfare as a supplementary screening device to means-testing is socially undesirable wher will attain some minimal level of utility. Our argument suggests that when misreporting of income by welfare claimants is sufficiently manifest, introducing w socially desirable. (xsd:string)
dc:creator	<http://linkeddata.econstor.eu/beta/resource/authors/38328819>
dc:creator	<http://linkeddata.econstor.eu/beta/resource/authors/38328820>
dc:creator	<http://linkeddata.econstor.eu/beta/resource/authors/38328821>
dcterms:isPartOf	<http://linkeddata.econstor.eu/beta/resource/collections/25>
dc:issued	2010 (xsd:gYear)
dc:keyword	means-testing (xsd:string)
dc:keyword	misreporting (xsd:string)
dc:keyword	utility maintenance (xsd:string)
dc:keyword	welfare (xsd:string)
dc:keyword	workfare (xsd:string)
rdfs:label	The desirability of workfare as a welfare ordeal: Revisited (xsd:string)
dc:language	eng (xsd:string)
foaf:maker	<http://linkeddata.econstor.eu/beta/resource/authors/38328819>
foaf:maker	<http://linkeddata.econstor.eu/beta/resource/authors/38328820>
foaf:maker	<http://linkeddata.econstor.eu/beta/resource/authors/38328821>
foaf:page	<http://hdl.handle.net/10419/44202>
dc:publisher	Forschungsinstitut zur Zukunft der Arbeit (IZA), Bonn (xsd:string)
dc:subject	<http://zbw.eu/beta/external_identifiers/jel/about#D6>
dc:subject	<http://zbw.eu/beta/external_identifiers/jel/about#H2>
dc:subject	<http://zbw.eu/beta/external_identifiers/jel/about#H5>
dc:title	The desirability of workfare as a welfare ordeal: Revisited (xsd:string)
dc:type	Working Paper (xsd:string)
rdf:type	swc:Paper
rdf:type	sioc:Item
rdf:type	foaf:Document

Canonical URI of publication

statements about publication



The desirability of workfare as a welfare ordeal: Revisited

Resource URI: http://linkeddata.econstor.eu/beta/resource/publications/44328

Home | Example Publications

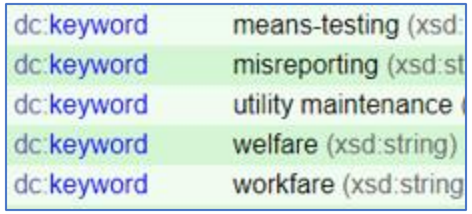
Property	Value
dcterms:abstract	In this paper we challenge the conventional wisdom that using workfare as a supplementary screening device to means-testing is socially undesirable wher will attain some minimal level of utility. Our argument suggests that when misreporting of income by welfare claimants is sufficiently manifest, introducing w socially desirable. (xsd:string)
dc:creator	<http://linkeddata.econstor.eu/beta/resource/authors/38328819>
dc:creator	<http://linkeddata.econstor.eu/beta/resource/authors/38328820>
dc:creator	<http://linkeddata.econstor.eu/beta/resource/authors/38328821>
dcterms:isPartOf	<http://linkeddata.econstor.eu/beta/resource/collections/25>
dc:issued	2010 (xsd:gYear)
dc:keyword	means-testing (xsd:string)
dc:keyword	misreporting (xsd:string)
dc:keyword	utility maintenance (xsd:string)
dc:keyword	welfare (xsd:string)
dc:keyword	workfare (xsd:string)
rdfs:label	The desirability of workfare as a welfare ordeal: Revisited (xsd:string)
dc:language	eng (xsd:string)
foaf:maker	<http://linkeddata.econstor.eu/beta/resource/authors/38328819>
foaf:maker	<http://linkeddata.econstor.eu/beta/resource/authors/38328820>
foaf:maker	<http://linkeddata.econstor.eu/beta/resource/authors/38328821>
foaf:page	<http://hdl.handle.net/10419/44202>
dc:publisher	Forschungsinstitut zur Zukunft der Arbeit (IZA), Bonn (xsd:string)
dc:subject	<http://zbw.eu/beta/external_identifiers/jel/about#D6>
dc:subject	<http://zbw.eu/beta/external_identifiers/jel/about#H2>
dc:subject	<http://zbw.eu/beta/external_identifiers/jel/about#H5>
dc:title	The desirability of workfare as a welfare ordeal: Revisited (xsd:string)
dc:type	Working Paper (xsd:string)
rdf:type	swc:Paper
rdf:type	sioc:Item
rdf:type	foaf:Document

keywords used for word2vec processing

Canonical URI

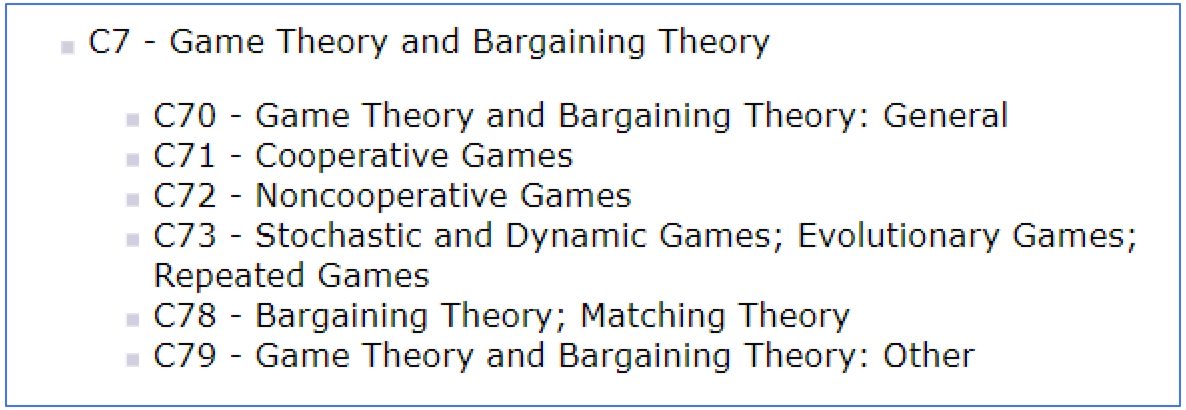
statements about publication

Why keywords?



dc:keyword	means-testing (xsd:
dc:keyword	misreporting (xsd:st
dc:keyword	utility maintenance (
dc:keyword	welfare (xsd:string)
dc:keyword	workfare (xsd:string

- word2vec processes tokens
- sets of keywords can be considered as sets of tokens, to be processed by word2vec
- Classification (part of EconStor) not used for word2vec: missing means to derive set of tokens from classification

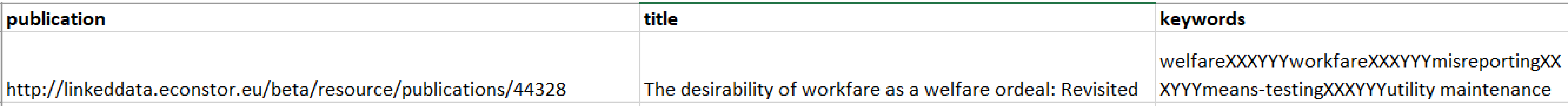


- C7 - Game Theory and Bargaining Theory
  - C70 - Game Theory and Bargaining Theory: General
  - C71 - Cooperative Games
  - C72 - Noncooperative Games
  - C73 - Stochastic and Dynamic Games; Evolutionary Games; Repeated Games
  - C78 - Bargaining Theory; Matching Theory
  - C79 - Game Theory and Bargaining Theory: Other

Pre-processing of data

- SPARQL Query to linked data endpoint at http://linkeddata.econstor.eu/beta/sparql
- Result: list of publications – for each publication
  - title
  - URI
  - keywords
- Sample output next slide

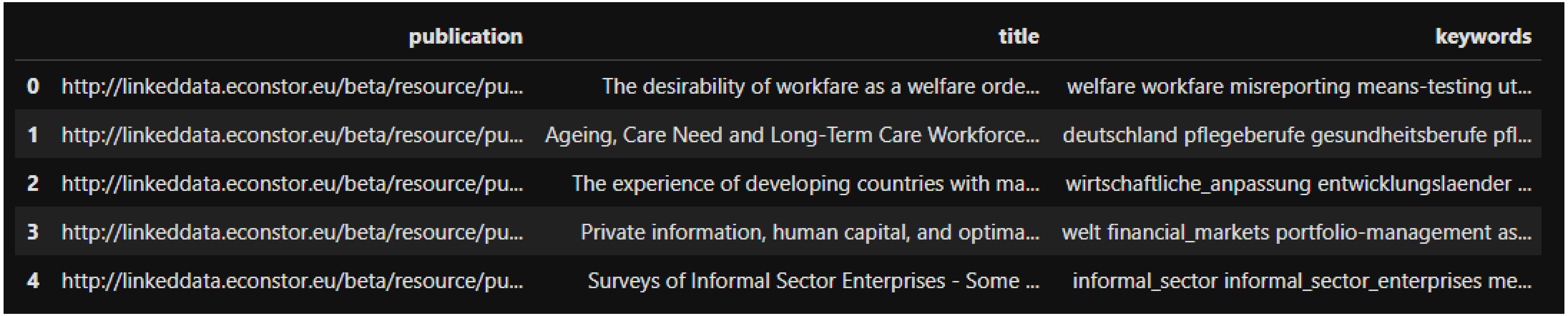
Sample output of pre-processing



publication	title	keywords
http://linkeddata.econstor.eu/beta/resource/publications/44328	The desirability of workfare as a welfare ordeal: Revisited	welfareXXXYYYworkfareXXXYYYmisreportingXXXYYYmeans-testingXXXYYYutility maintenance

- Keywords are temporarily separated via delimiter
- Before processing with word2vec, multi-token keywords are merged
  - Before: “wirtschaftliche anpassung”
  - After: “wirtschaftliche_anpassung”
- In that way, word2vec can calculate similarly per keyword, independent of the keyword internal segmentation
  - No need to remove stopwords, since we assume no internal structure for a keyword
- Input to word2vec (next slide)

Input to word2vec



	publication	title	keywords
0	http://linkeddata.econstor.eu/beta/resource/pu...	The desirability of workfare as a welfare orde...	welfare workfare misreporting means-testing ut...
1	http://linkeddata.econstor.eu/beta/resource/pu...	Ageing, Care Need and Long-Term Care Workforce...	deutschland pflegeberufe gesundheitsberufe pfl...
2	http://linkeddata.econstor.eu/beta/resource/pu...	The experience of developing countries with ma...	wirtschaftliche_anpassung entwicklungslaender ...
3	http://linkeddata.econstor.eu/beta/resource/pu...	Private information, human capital, and optima...	welt financial_markets portfolio-management as...
4	http://linkeddata.econstor.eu/beta/resource/pu...	Surveys of Informal Sector Enterprises - Some ...	informal_sector informal_sector_enterprises me...

Overview

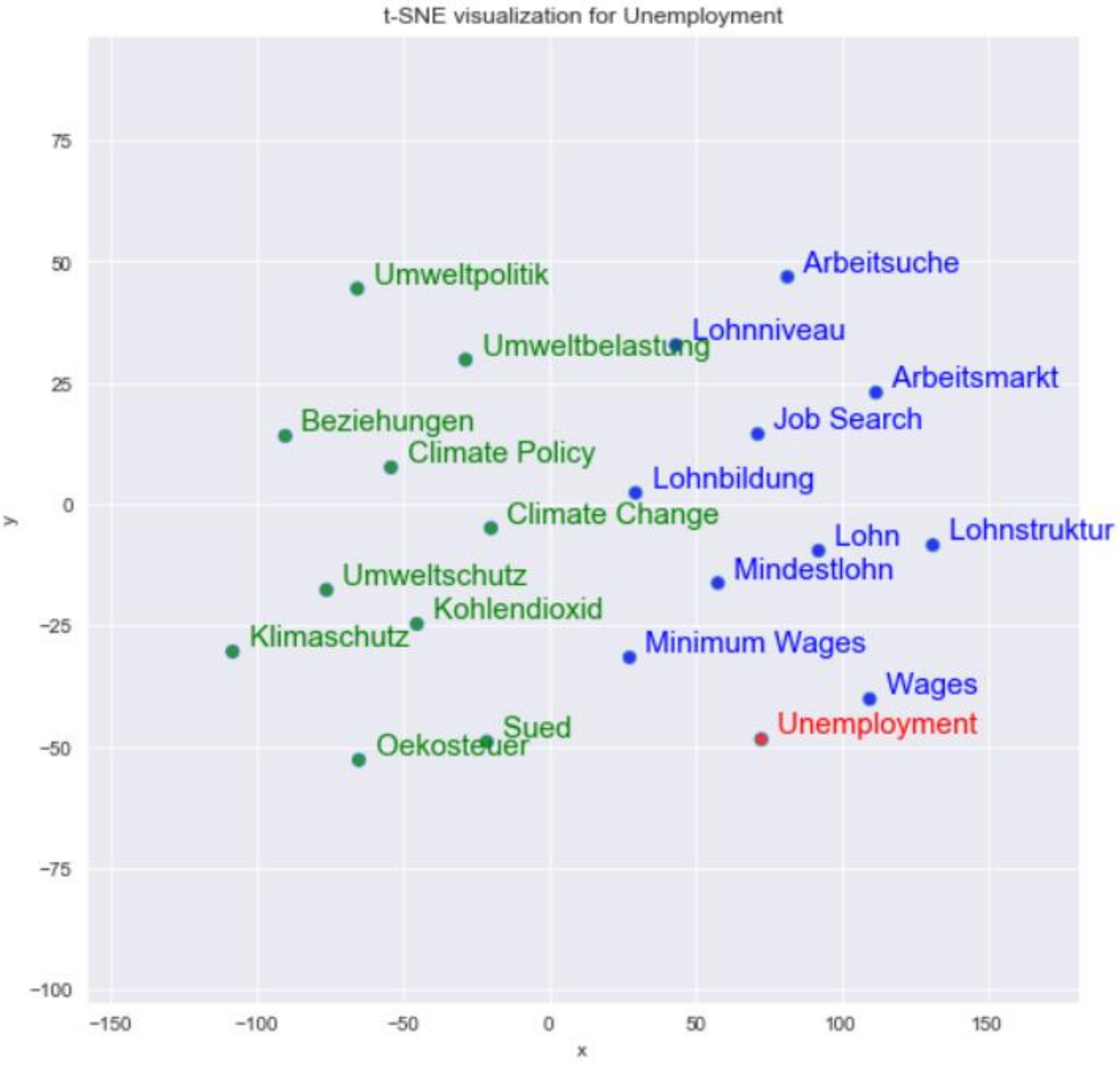
- Motivation
- Aspects of word2vec
- Processing of metadata records: input and pre-preprocessing
- Metadata records and word2vec: example output & optimization
- Application: recommendation based on metadata semantic similarity
- Next steps

Library: gensim

- Widely used word2vec library
- See https://radimrehurek.com/gensim/index.html

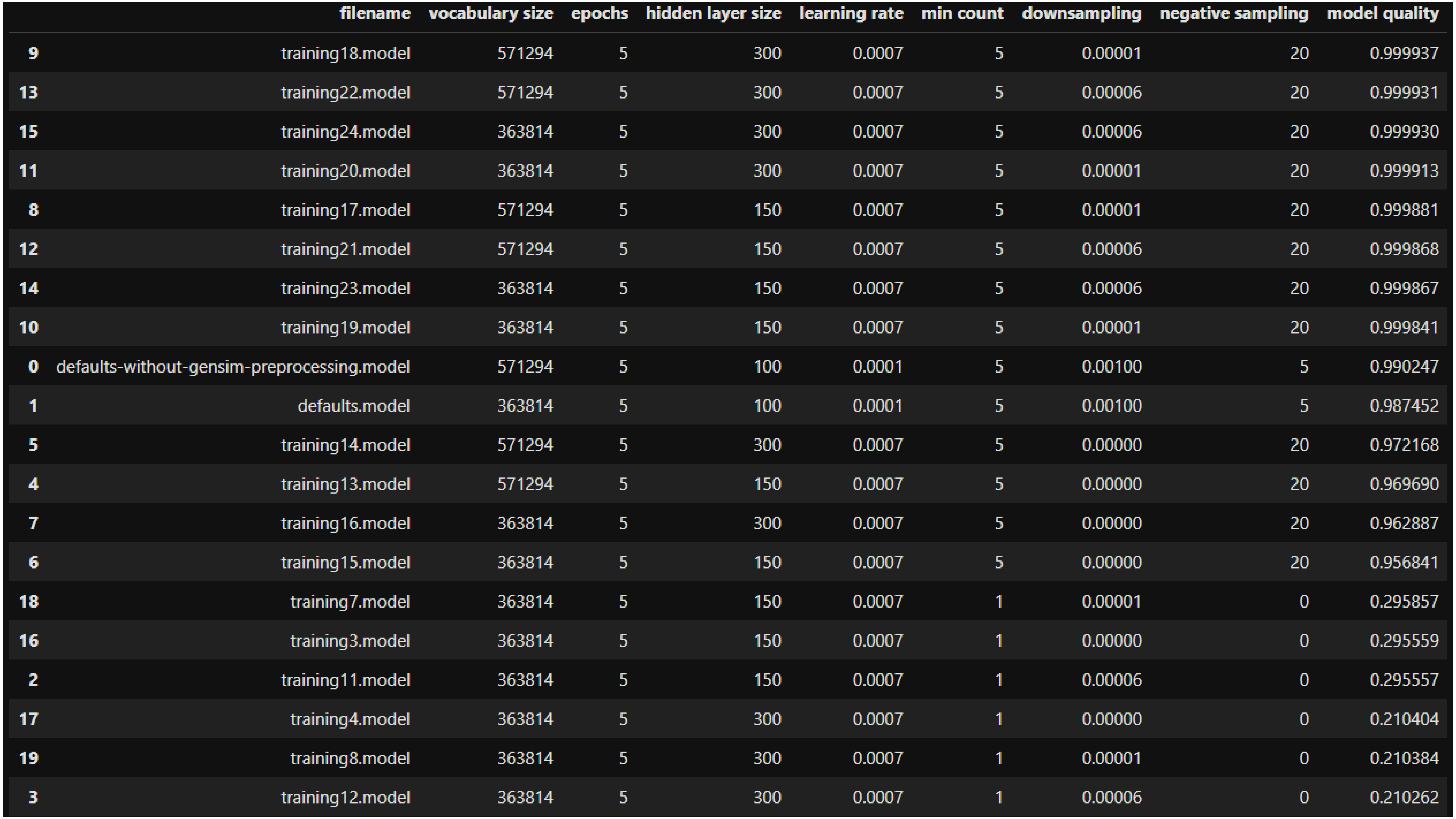
Example output – words similar to "Unemployment"

- Using t-SNE (distributed stochastic neighbor embedding) for dimensionality reduction
- Keywords are in English and German
- word2vec calculates similarity independent of language
- Blue keywords are similar, green keywords are not similar



Gutter search for optimizing word2vec parameters

- training gutter 2 * 3 * 2 * 2 = 24
  - size of matrix (using 150 or 300) = 2
  - sample (0, 1-e5, 6-e-5) = 3
  - using preprocessing (true or false) = 3
  - using mincout=5 and negative=20, or defaults = 2
- In addititition, default processing with or without gensim preprocessing
- Model quality overview – see next slide



	filename	vocabulary size	epochs	hidden layer size	learning rate	min count	downsampling	negative sampling	model quality
9	training18.model	571294	5	300	0.0007	5	0.00001	20	0.999937
13	training22.model	571294	5	300	0.0007	5	0.00006	20	0.999931
15	training24.model	363814	5	300	0.0007	5	0.00006	20	0.999930
11	training20.model	363814	5	300	0.0007	5	0.00001	20	0.999913
8	training17.model	571294	5	150	0.0007	5	0.00001	20	0.999881
12	training21.model	571294	5	150	0.0007	5	0.00006	20	0.999868
14	training23.model	363814	5	150	0.0007	5	0.00006	20	0.999867
10	training19.model	363814	5	150	0.0007	5	0.00001	20	0.999841
0	defaults-without-gensim-preprocessing.model	571294	5	100	0.0001	5	0.00100	5	0.990247
1	defaults.model	363814	5	100	0.0001	5	0.00100	5	0.987452
5	training14.model	571294	5	300	0.0007	5	0.00000	20	0.972168
4	training13.model	571294	5	150	0.0007	5	0.00000	20	0.969690
7	training16.model	363814	5	300	0.0007	5	0.00000	20	0.962887
6	training15.model	363814	5	150	0.0007	5	0.00000	20	0.956841
18	training7.model	363814	5	150	0.0007	1	0.00001	0	0.295857
16	training3.model	363814	5	150	0.0007	1	0.00000	0	0.295559
2	training11.model	363814	5	150	0.0007	1	0.00006	0	0.295557
17	training4.model	363814	5	300	0.0007	1	0.00000	0	0.210404
19	training8.model	363814	5	300	0.0007	1	0.00001	0	0.210384
3	training12.model	363814	5	300	0.0007	1	0.00006	0	0.210262

Issue with evalution

- No gold standard of word vectors for our input data is available
- Current evaluation calculates only the average of similarities scores for 10% of the vocabulary
  - E.g. for 363814 word = with 3638 words
- Future step: build evaluation vocabulary

Overview

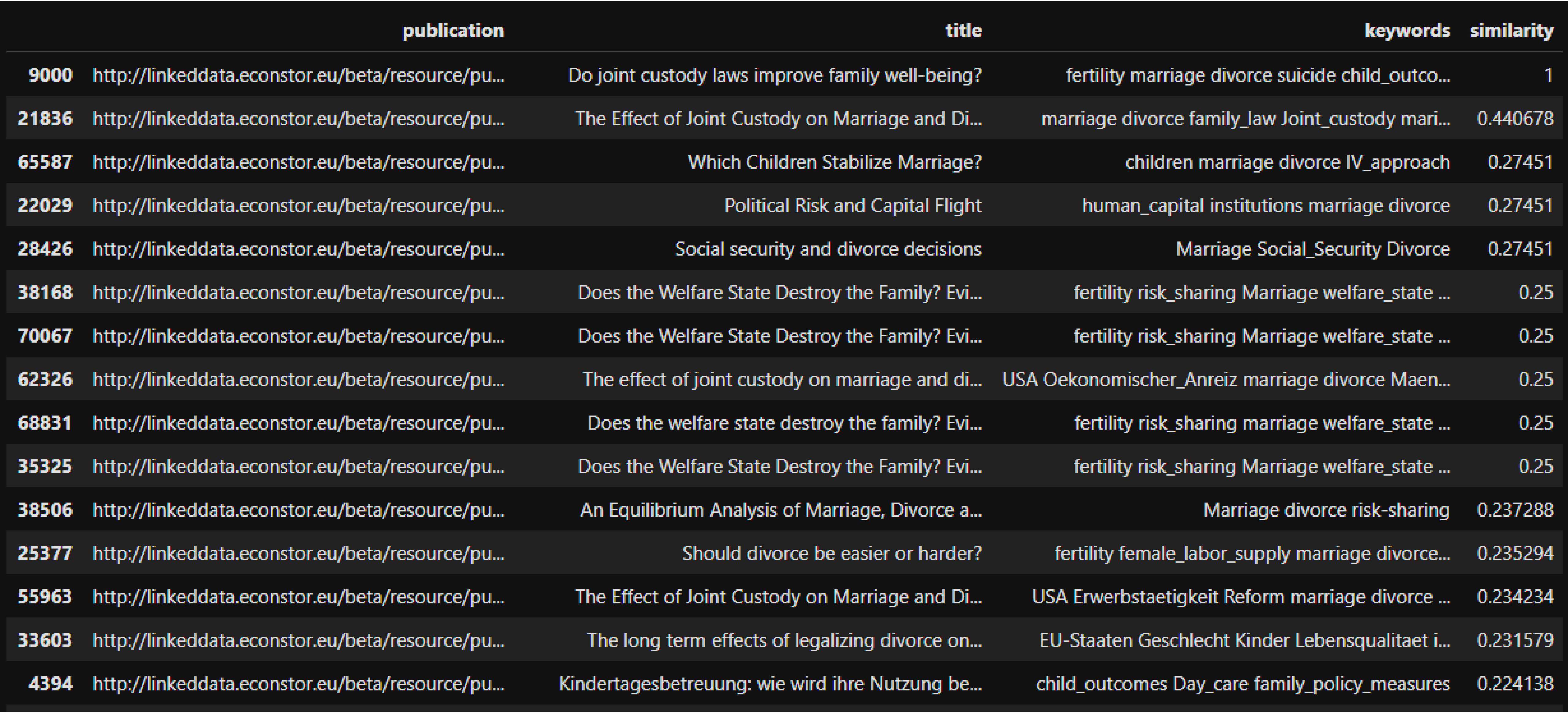
- Motivation
- Aspects of word2vec
- Processing of metadata records: input and pre-preprocessing
- Metadata records and word2vec: example output & optimization
- Application: recommendation based on metadata semantic similarity
- Next steps

Application: recommender system

- Input: a document and its keywords
- Output: a list of similar documents
- Calculation of Jaccard similarity – first without word2vec: size of intersection divided by total size of set
  - doc 1: Germany labour
  - doc 2: Germany workforce
  - intersection = 1
  - total set = 3
  - Jaccard similarity = 1 / 3 = 0.33
- Recommendation approach: content-based filtering recommendation approach: given an input doc, ordering other document based on Jaccard similarity

Application: recommender system with word2vec

- Input: a document and its keywords
- Output: a list of similar documents
- Calculation of Jaccard similarity – with word2vec: size of intersection divided by total size of set
  - doc 1: Germany labour (workforce 0.9 similar to labour)
  - doc 2: Germany workforce
  - intersection = 2
  - total set = 3
  - Jaccard similarity = 2 / 3 = 0.666
- Example output (next slide)



	publication	title	keywords	similarity
9000	http://linkeddata.econstor.eu/beta/resource/pu...	Do joint custody laws improve family well-being?	fertility marriage divorce suicide child_outco...	1
21836	http://linkeddata.econstor.eu/beta/resource/pu...	The Effect of Joint Custody on Marriage and Di...	marriage divorce family_law Joint_custody mari...	0.440678
65587	http://linkeddata.econstor.eu/beta/resource/pu...	Which Children Stabilize Marriage?	children marriage divorce IV_approach	0.27451
22029	http://linkeddata.econstor.eu/beta/resource/pu...	Political Risk and Capital Flight	human_capital institutions marriage divorce	0.27451
28426	http://linkeddata.econstor.eu/beta/resource/pu...	Social security and divorce decisions	Marriage Social_Security Divorce	0.27451
38168	http://linkeddata.econstor.eu/beta/resource/pu...	Does the Welfare State Destroy the Family? Evi...	fertility risk_sharing Marriage welfare_state ...	0.25
70067	http://linkeddata.econstor.eu/beta/resource/pu...	Does the Welfare State Destroy the Family? Evi...	fertility risk_sharing Marriage welfare_state ...	0.25
62326	http://linkeddata.econstor.eu/beta/resource/pu...	The effect of joint custody on marriage and di...	USA Oekonomischer_Anreiz marriage divorce Maen...	0.25
68831	http://linkeddata.econstor.eu/beta/resource/pu...	Does the welfare state destroy the family? Evi...	fertility risk_sharing marriage welfare_state ...	0.25
35325	http://linkeddata.econstor.eu/beta/resource/pu...	Does the Welfare State Destroy the Family? Evi...	fertility risk_sharing Marriage welfare_state ...	0.25
38506	http://linkeddata.econstor.eu/beta/resource/pu...	An Equilibrium Analysis of Marriage, Divorce a...	Marriage divorce risk-sharing	0.237288
25377	http://linkeddata.econstor.eu/beta/resource/pu...	Should divorce be easier or harder?	fertility female_labor_supply marriage divorce...	0.235294
55963	http://linkeddata.econstor.eu/beta/resource/pu...	The Effect of Joint Custody on Marriage and Di...	USA Erwerbstaetigkeit Reform marriage divorce ...	0.234234
33603	http://linkeddata.econstor.eu/beta/resource/pu...	The long term effects of legalizing divorce on...	EU-Staaten Geschlecht Kinder Lebensqualitaet i...	0.231579
4394	http://linkeddata.econstor.eu/beta/resource/pu...	Kindertagesbetreuung: wie wird ihre Nutzung be...	child_outcomes Day_care family_policy_measures	0.224138

Overview

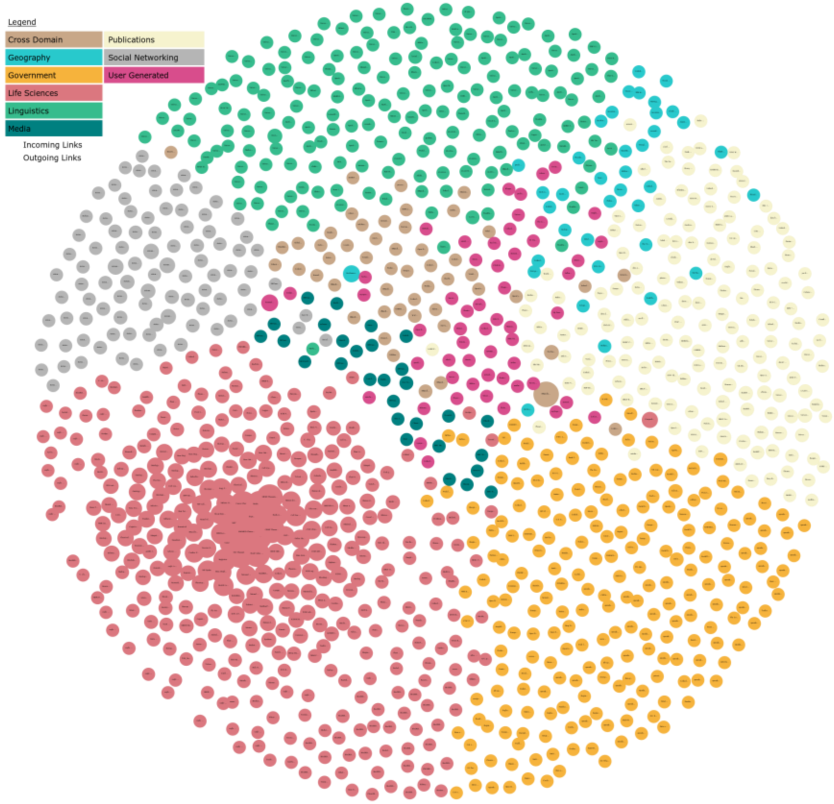
- Motivation
- Aspects of word2vec
- Processing of metadata records: input and pre-preprocessing
- Metadata records and word2vec: example output & optimization
- Application: recommendation based on metadata semantic similarity
- Next steps

Open questions

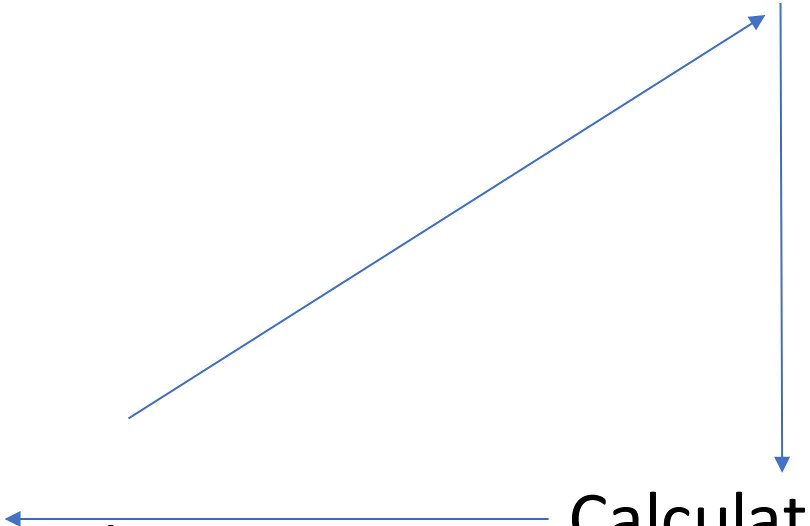
- Dealing with rare keywords
  - word2vec drops several words
  - Gutter search shows that keeping all words leads to lower quality of the model
  - An approach for rare words is needed
- Evaluating the model
  - Build samples, use perplexity for evaluation
  - Compare e.g. to doc2vec based processing and tf-idf
- Using existing models and disambiguation
  - E.g. Google BERT provides pre-trained models to use disambiguation
    - E.g. German “Bank” from Bank”
  - Fine-tuning of models like BERT could be a basis for implementing disambiguation
- Using the similarities to enhance the linked open data sources – **building a circle between linked data and machine learning**

Motivation – from the perspective of applied computational linguistics research

Linked (open) data sources



Input to machine learning



Calculation of data similarities

Enhancement of data models

Word2Vec and Metadata Records

Felix Sasaki

April 2019