Paweł Lula, Janusz Tuchowski, Urszula Cieraszewska, Magdalena Talaga
Cracow University of Economics

Clustering in ontology-based exploratory analysis of scientific productivity

9th International Conference "Distributed Computing and Grid Technologies in Science and Education" (GRID'2021)

5-9 July 2021

CRACOW UNIVERSITY OF ECONOMICS

Outline

- Goals
- Ontologies and automatic annotation process
- Distance between concepts defined in ontologies
- Document as a set of concepts
- Distances between documents
- Cluster analysis of annotated documents
- Annotation with more than one ontology
- Cluster analysis of documents annotated with more then one ontology
- Practical aspects of cluster analysis of annotated documents (in the context of research productivity analysis)
- Conclusions

Goals

- Presentation of theoretical aspects of:
  - ontology-based exploratory text analysis,
  - cluster analysis of documents annotated with an ontology (ontologies),
- Ontologies for research productivity analysis
- Presentation of results of exemplary calculations

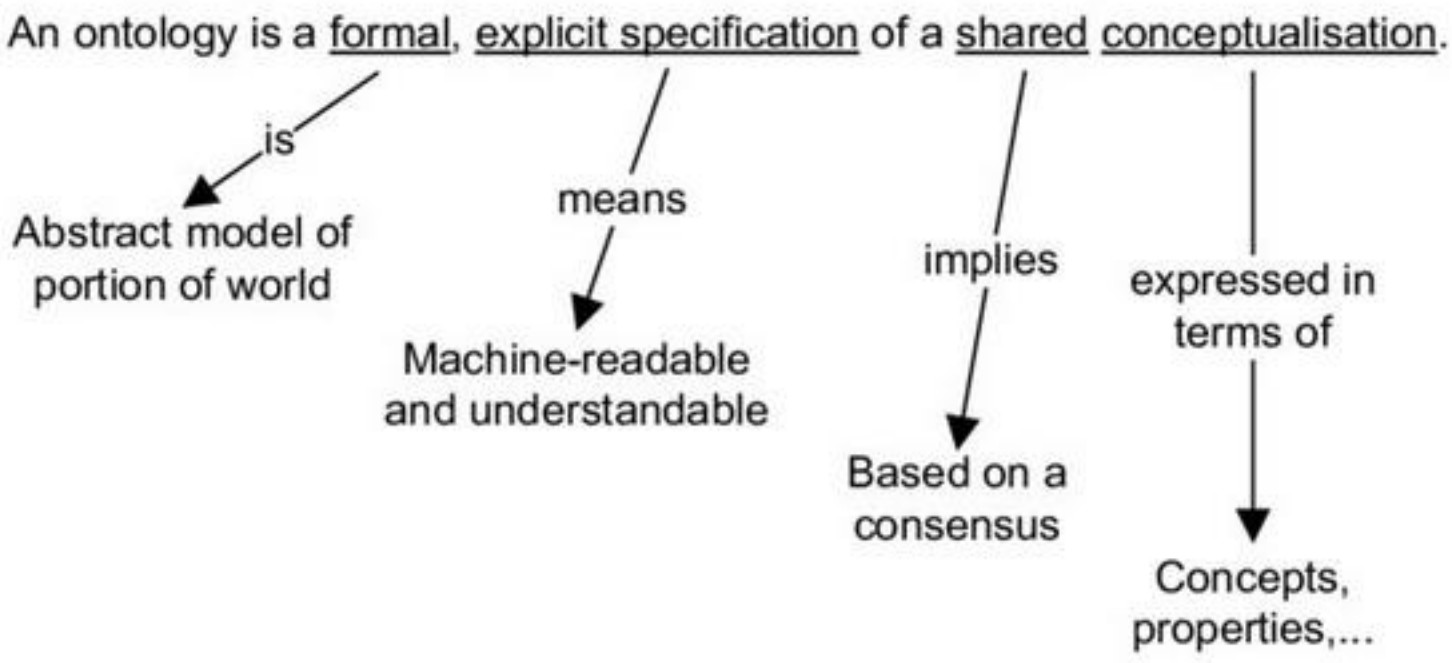
Ontologies
and automatic annotation process

Ontology

Thomas Robert "Tom" Gruber (born 1959):

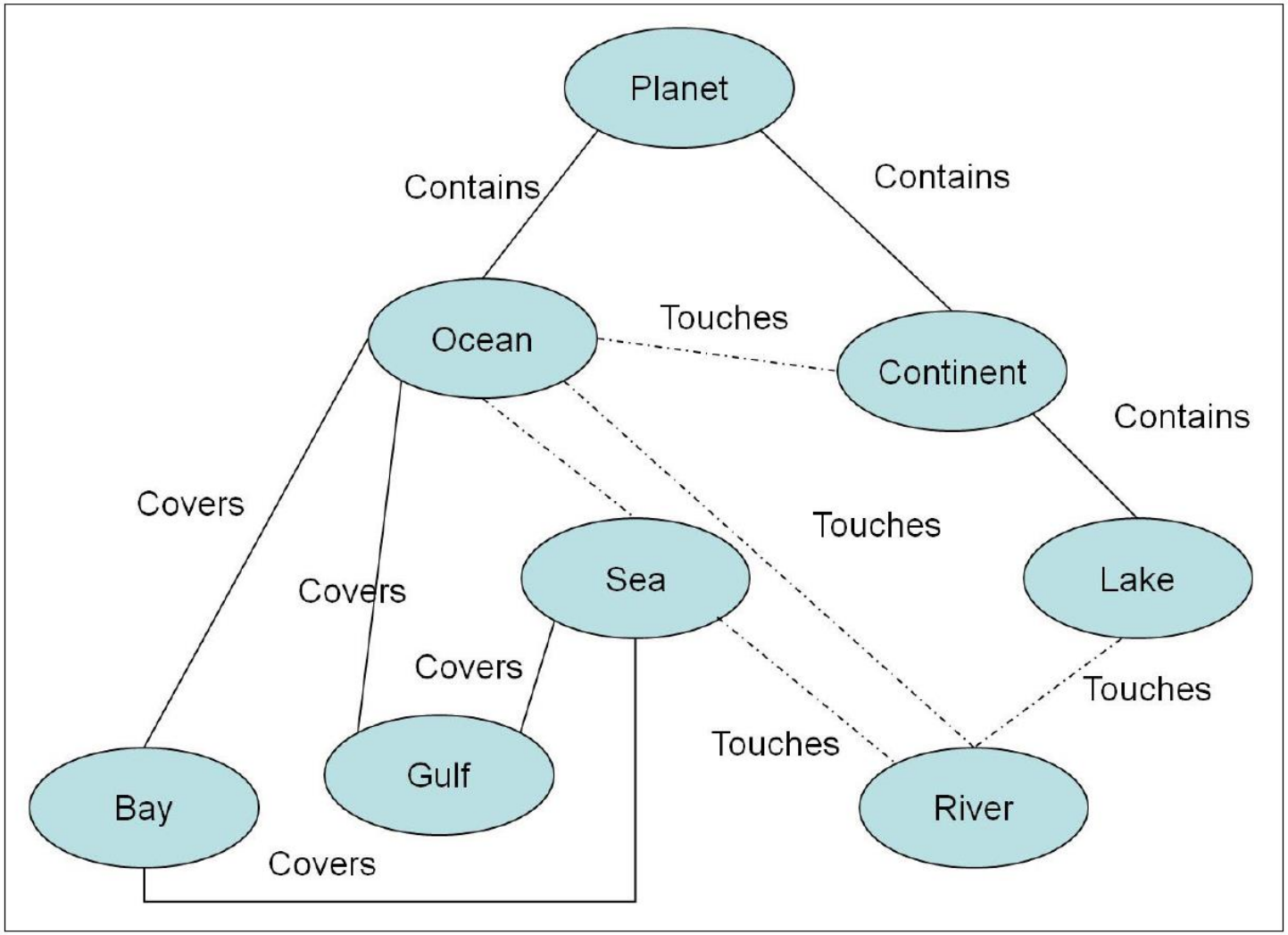
"An ontology is a formal, explicit specification of a shared conceptualization"

Ontologies



Ontology = the method of knowledge representation

Ontologies


Planet
Contains
Contains
Ocean
Touches
Continent
Contains
Covers
Touches
Sea
Lake
Covers
Covers
Touches
Touches
Gulf
Bay
River
Covers

Competency model (competency ontology)

YAML format

```yaml
competencies:
    # kompetencje społeczne
    social:
        communication_skills: # S_COM
            description: kompetencje komunikacyjne
            phrases:
            - '#swobodna #komunikacja'
            - '#komunikuje się #swobodnie'
            - 'wywieranie #wpływu na #rozmówcę'
            - '#(umiejętności|zdolności|predyspozycje) #komunikacyjne'
            - '#potrafi się #komunikować'
            - 'łatwo #nawiązuje #(kontakt|kontakty|relację|relacje)'
        negociation_skills: # S_NEG
            description: negocjowanie
            phrases:
            - '#(umiejętności|zdolności) #negocjacyjne'
            - '#negocjowanie'
            - 'umiejętności związane z #prowadzeniem #negocjacji'
        assertiveness: # S_ASSERT
            description: asertywność
            phrases:
            - '#asertywność'
            - '#asertywny'
        loyalty: # S_LOY
            description: lojalność
            phrases:
            - '#lojalność'
            - '#lojalny'
        teamworking: # S_TEAM
            description: umiejętność pracy w grupie
            phrases:
            - 'umiejętność #pracy w #grupie'
            - '#praca #zespołowa'
```

Types of ontologies

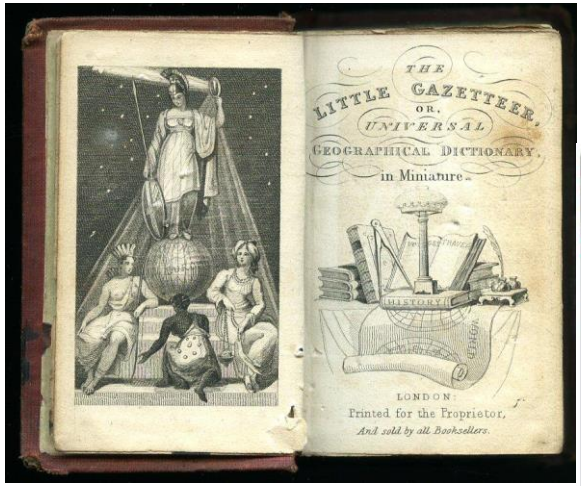
Types of ontologies

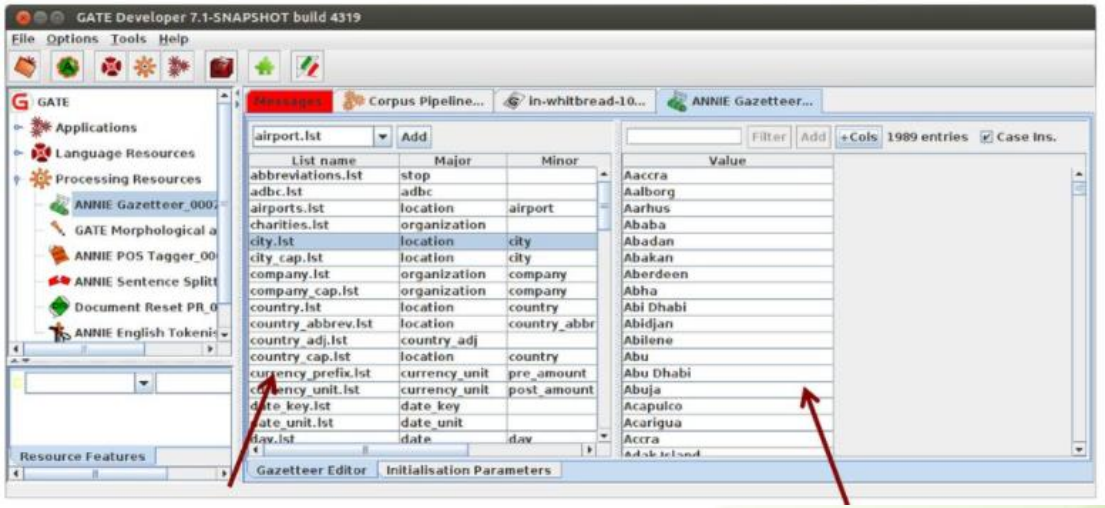
- linear
- hierarchical
- network

Linear ontologies

- linear ontology = gazetteer = a list of concepts
- used for named entity recognition (identification of names)

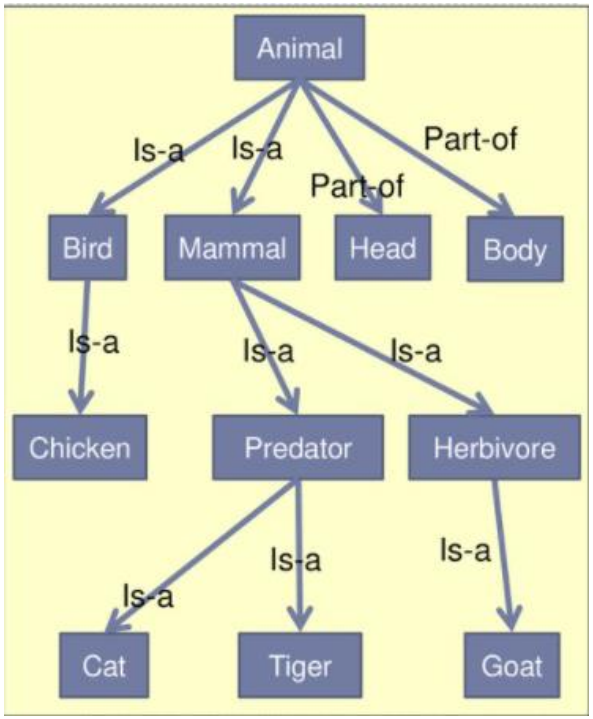
1. A geographical dictionary. 2. A directory in which the entries are arranged by geographical location. For example, a gazetteer of restaurants. can be abbreviated as gazetteer.

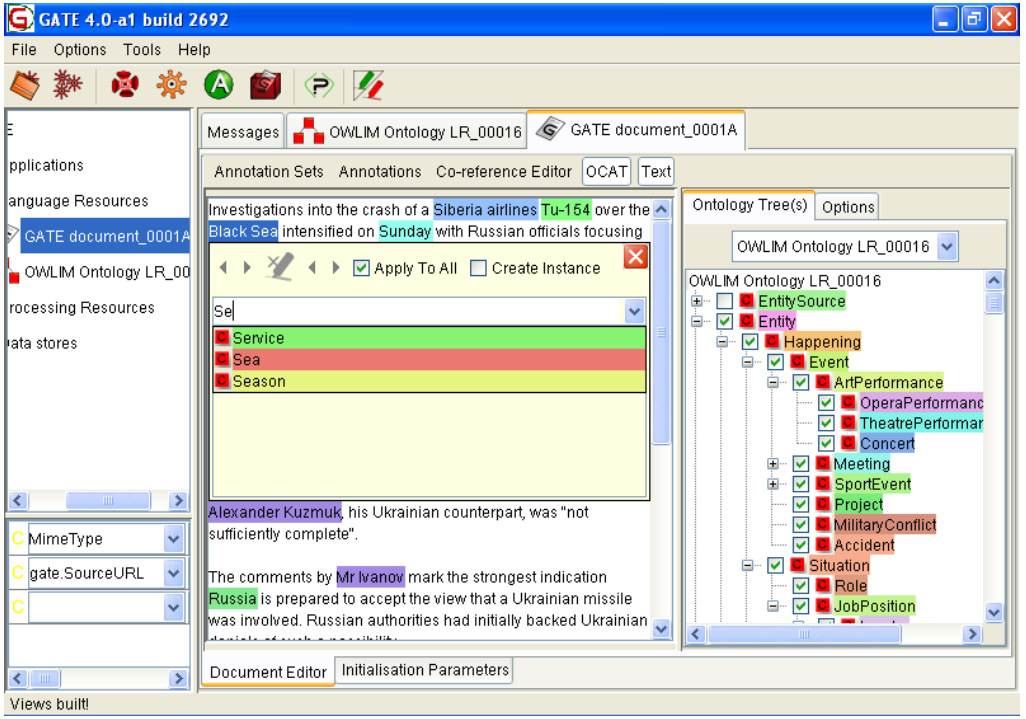




Hierarchical ontology

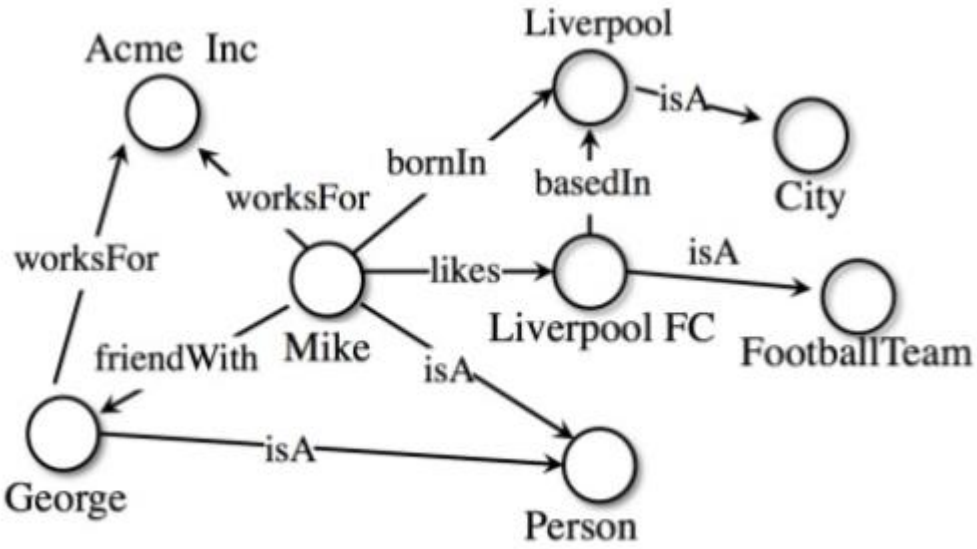
- concepts form tree structure





Network ontology

- represented as a graph
- concepts are represented by vertices, relations by edges



JEL as a hierarchical ontology

General Categories

- **A** General Economics and Teaching
- **B** History of Economic Thought, Methodology, and Heterodox Approaches
- **C** Mathematical and Quantitative Methods
- **D** Microeconomics
- **E** Macroeconomics and Monetary Economics
- **F** International Economics
- **G** Financial Economics
- **H** Public Economics
- **I** Health, Education, and Welfare
- **J** Labor and Demographic Economics
- **K** Law and Economics
- **L** Industrial Organization
- **M** Business Administration and Business Economics • Marketing • Accounting • Personnel Economics
- **N** Economic History
- **O** Economic Development, Innovation, Technological Change, and Growth
- **P** Economic Systems
- **Q** Agricultural and Natural Resource Economics • Environmental and Ecological Economics
- **R** Urban, Rural, Regional, Real Estate, and Transportation Economics
- **Y** Miscellaneous Categories
- **Z** Other Special Topics

M1 Business Administration
- M10 General
- M11 Production Management
- M12 Personnel Management • Executives; Executive Compensation
- M13 New Firms • Startups
- M14 Corporate Culture • Diversity • Social Responsibility
- M15 IT Management
- M16 International Business Administration
- M19 Other

M2 Business Economics
- M20 General
- M21 Business Economics
- M29 Other

M3 Marketing and Advertising
- M30 General
- M31 Marketing
- M37 Advertising
- M38 Government Policy and Regulation
- M39 Other

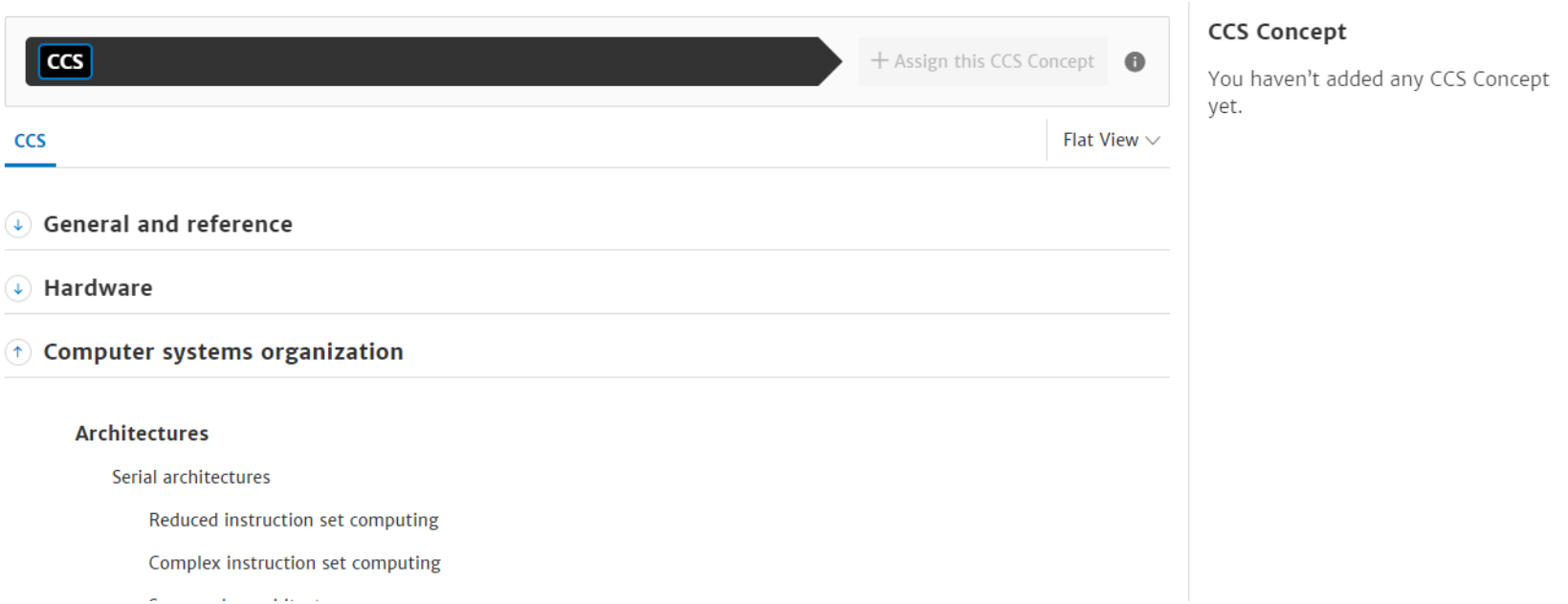
AMERICAN ECONOMIC ASSOCIATION

JEL Classification System / EconLit Subject Descriptors

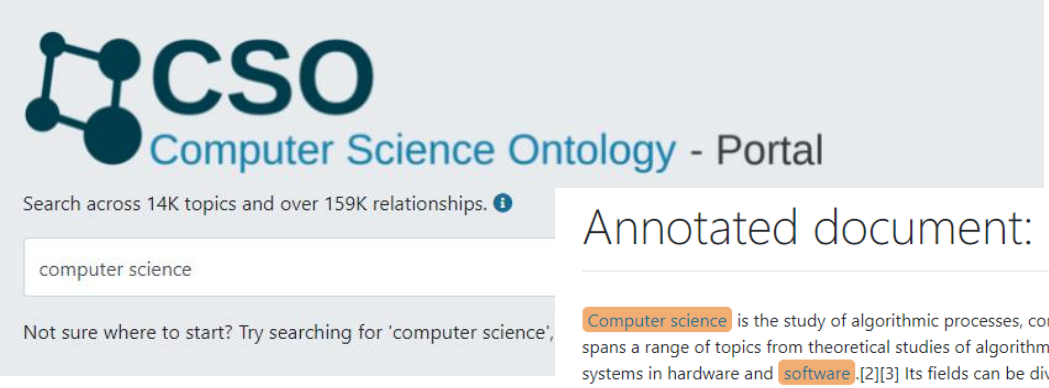
ACM as a hierarchical ontology

ACM Computing Classification System

The 2012 ACM Computing Classification System has been developed as a poly-hierarchical ontology that can be utilized in semantic web applications. It replaces the traditional 1998 version of the ACM Computing Classification System (CCS), which has served as the de facto standard classification system for the computing field. It is being integrated into the search capabilities and visual topic displays of the ACM Digital Library. It relies on a semantic vocabulary as the single source of categories and concepts that reflect the state of the art of the computing discipline and is receptive to structural change as it evolves in the future. ACM provides a tool within the visual display format to facilitate the application of 2012 CCS categories to forthcoming papers and a process to ensure that the CCS stays ... (More)



CSO as a graph ontology



Annotated document:

Computer science is the study of algorithmic processes, computational machines and computation itself.[1] As a discipline, computer science spans a range of topics from theoretical studies of algorithms, computation and information to the practical issues of implementing computational systems in hardware and software.[2][3] Its fields can be divided into theoretical and practical disciplines. For example, the theory of computation concerns abstract models of computation and general classes of problems that can be solved using them, while computer graphics or computational geometry emphasize more specific applications. Algorithms and data structures have been called the heart of computer science.[4] Programming language theory considers approaches to the description of computational processes, while computer programming involves the use of them to create complex systems. Computer architecture describes construction of computer components and computer-operated equipment. Artificial intelligence aims to synthesize goal-orientated processes such as problem-solving, decision-making, environmental adaptation, planning and learning found in humans and animals. A digital computer is capable of simulating various information processes.[5] The fundamental concern of computer science is determining what can and cannot be automated.[6] Computer scientists usually focus on academic research. The Turing Award is generally recognized as the highest distinction in computer sciences.

Topics found:

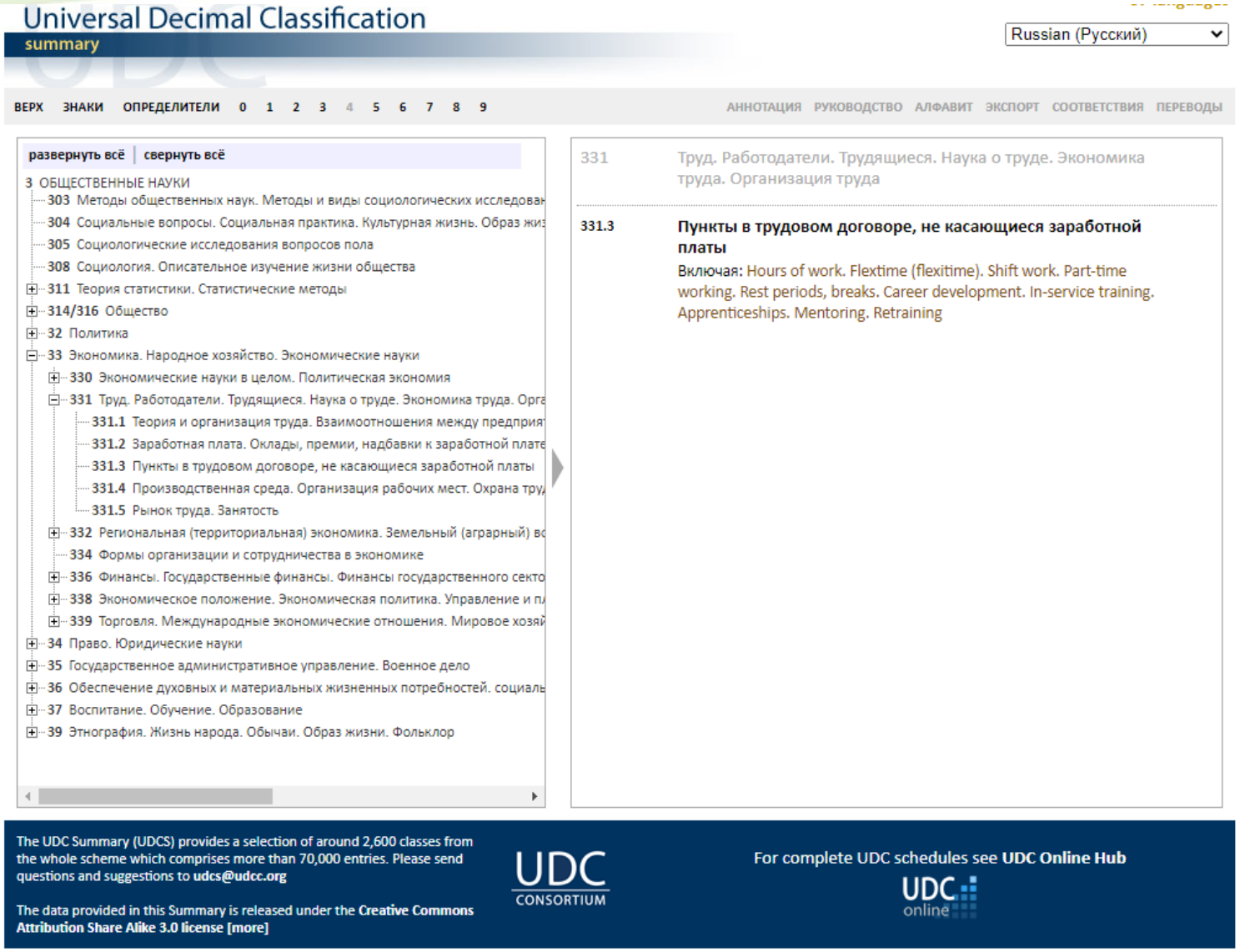
Extracted topics within the document, using the **Syntactic module**.

artificial intelligence	computer hardware	computer programming	computer science
geometry	learning	programming languages	software

Extracted topics within the document, using the **Semantic module**.

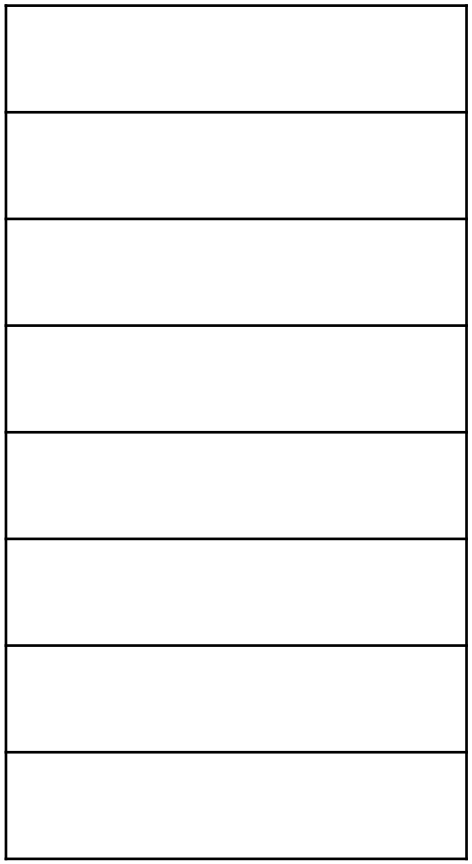
artificial intelligence	computer hardware	computer programming	computer science

UDC as a hybrid ontology



Distances between concepts defined in ontologies

Distances for linear ontologies (gazetteers)

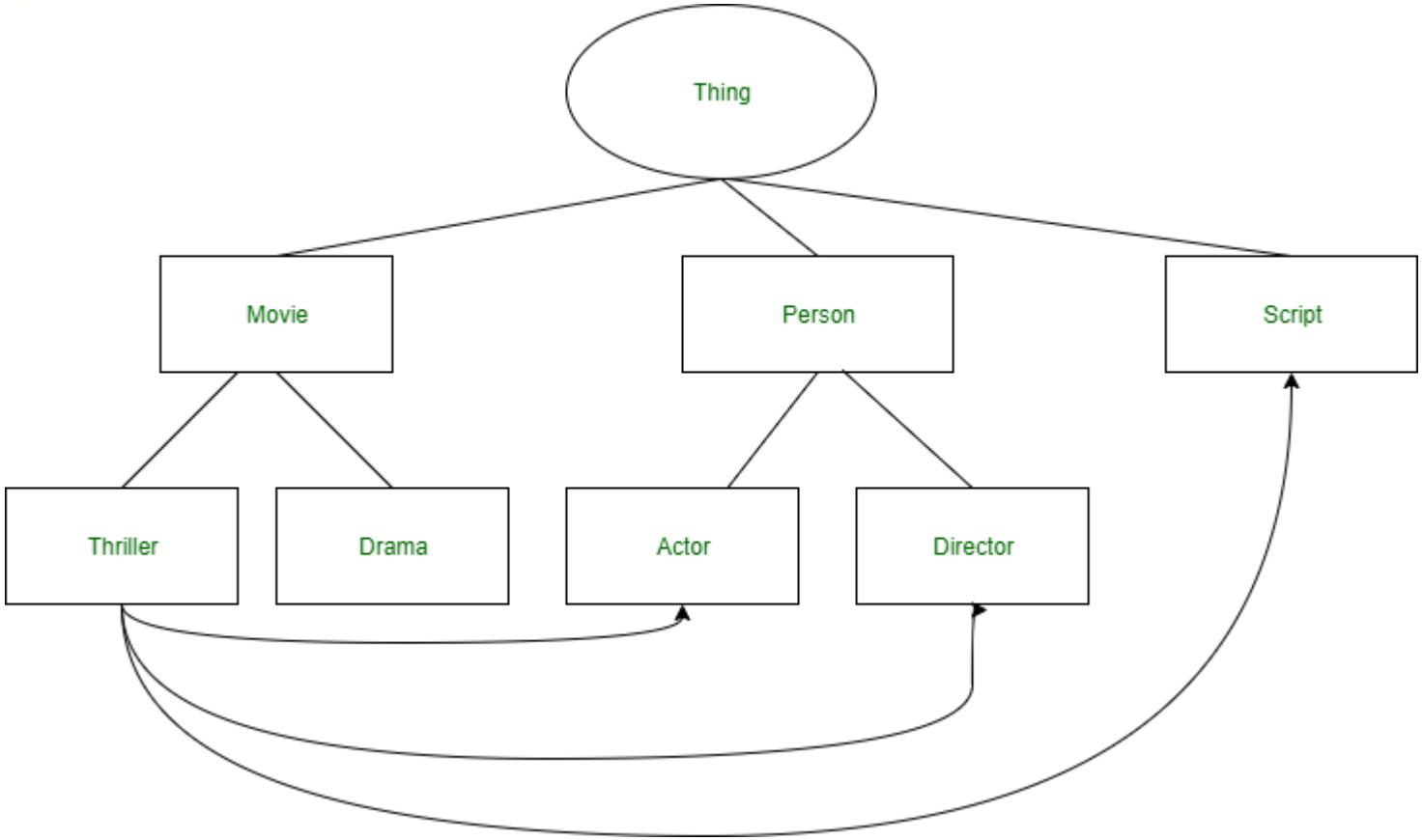


$$C_i = C_j$$

or

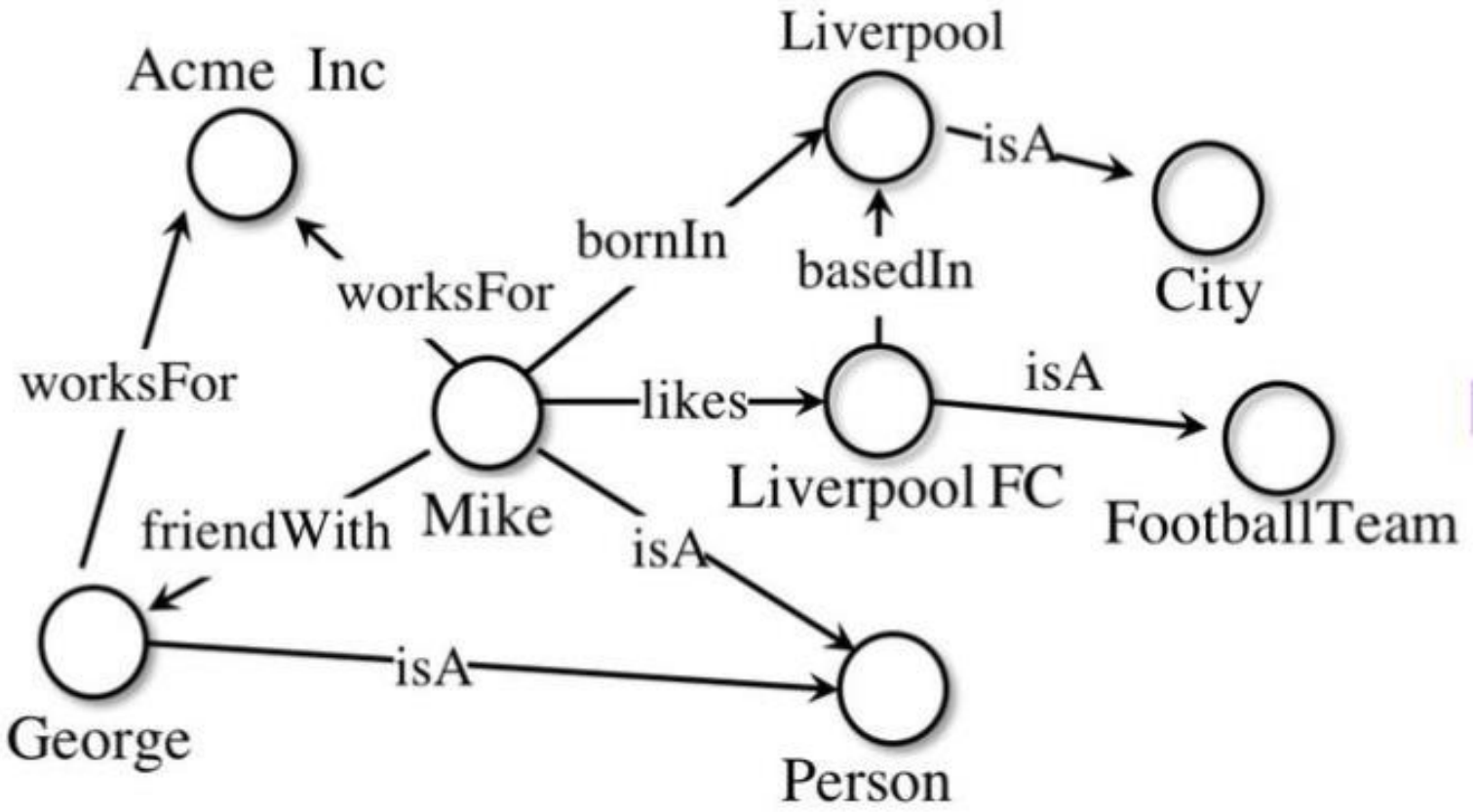
$$C_i \neq C_j$$

Distances for hierarchical ontologies



- based on path length,
- based on information theory

Distances for network ontologies



- based on path length,
- *weights play crucial role!*

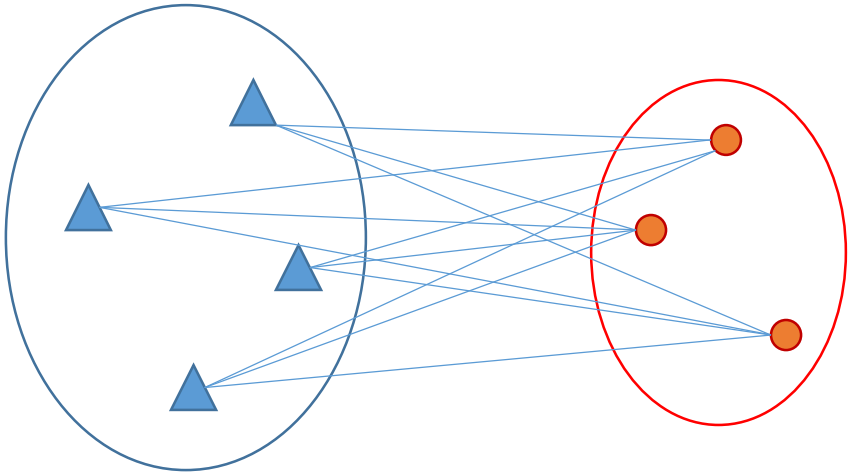
Distances between documents

Document as a set of concepts

Computer science is the study of algorithmic processes, computational machines and computation itself.[1] As a discipline, computer science spans a range of topics from theoretical studies of algorithms, computation and information to the practical issues of implementing computational systems in hardware and software.[2][3] Its fields can be divided into theoretical and practical disciplines. For example, the theory of computation concerns abstract models of computation and general classes of problems that can be solved using them, while computer graphics or computational geometry emphasize more specific applications. Algorithms and data structures have been called the heart of computer science.[4] Programming language theory considers approaches to the description of computational processes, while computer programming involves the use of them to create complex systems. Computer architecture describes construction of computer components and computer-operated equipment. Artificial intelligence aims to synthesize goal-orientated processes such as problem-solving, decision-making, environmental adaptation, planning and learning found in humans and animals. A digital computer is capable of simulating various information processes.[5] The fundamental concern of computer science is determining what can and cannot be automated.[6] Computer scientists usually focus on academic research. The Turing Award is generally recognized as the highest distinction in computer sciences.

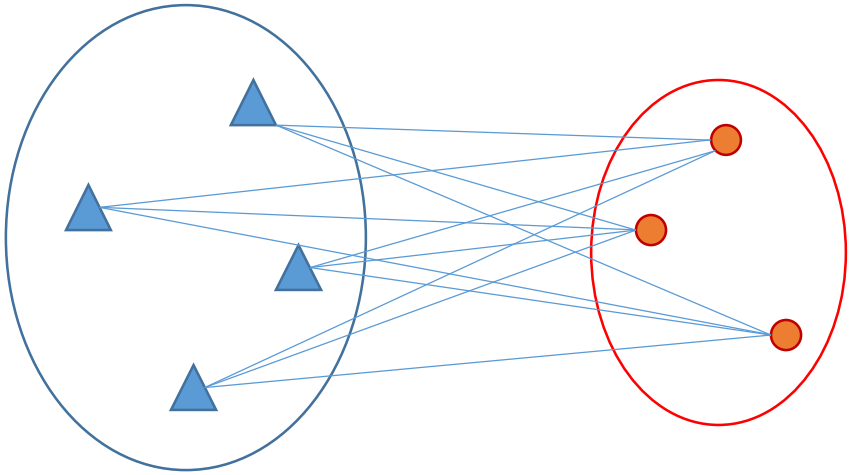
In computing, a database is an organized collection of data stored and accessed electronically from a computer system. Where databases are more complex they are often developed using formal design and modeling techniques. The database management system (DBMS) is the software that interacts with end users, applications, and the database itself to capture and analyze the data. The DBMS software additionally encompasses the core facilities provided to administer the database. The sum total of the database, the DBMS and the associated applications can be referred to as a "database system". Often the term "database" is also used to loosely refer to any of the DBMS, the database system or an application associated with the database. Computer scientists may classify database-management systems according to the database models that they support. Relational databases became dominant in the 1980s. These model data as rows and columns in a series of tables, and the vast majority use SQL for writing and querying data. In the 2000s, non-relational databases became popular, referred to as NoSQL because they use different query languages.

Distances between documents



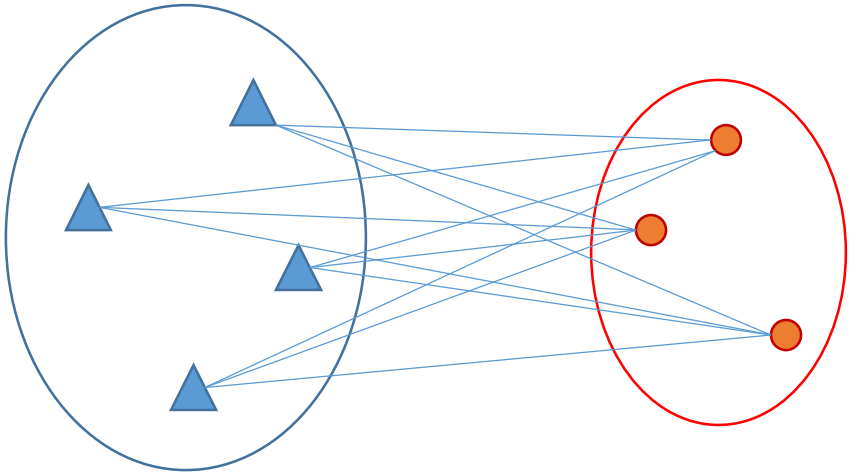
$$sim(Zb_1, \mathrm{Zb}_2) = avg\left(C_i, C_j\right), C_i \in Zb_1, C_j \in Zb_2$$

Distances between documents



$$sim(expr_1, expr_2) = \frac{\sum_{i-1}^{N} \min_j \left(sim(c_i, c_j)\right) + \sum_{j-1}^{M} \min_i \left(sim(c_i, c_j)\right)}{N + M}$$

Distances between documents



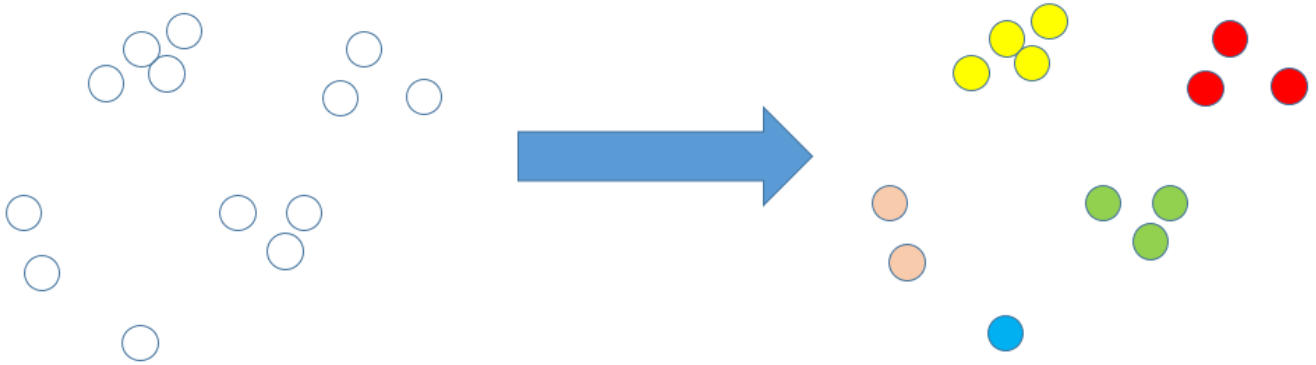
$$sim(expr_1, expr_2) = \arg\min_{c_i, c_j} \sum sim(c_i, c_j)$$

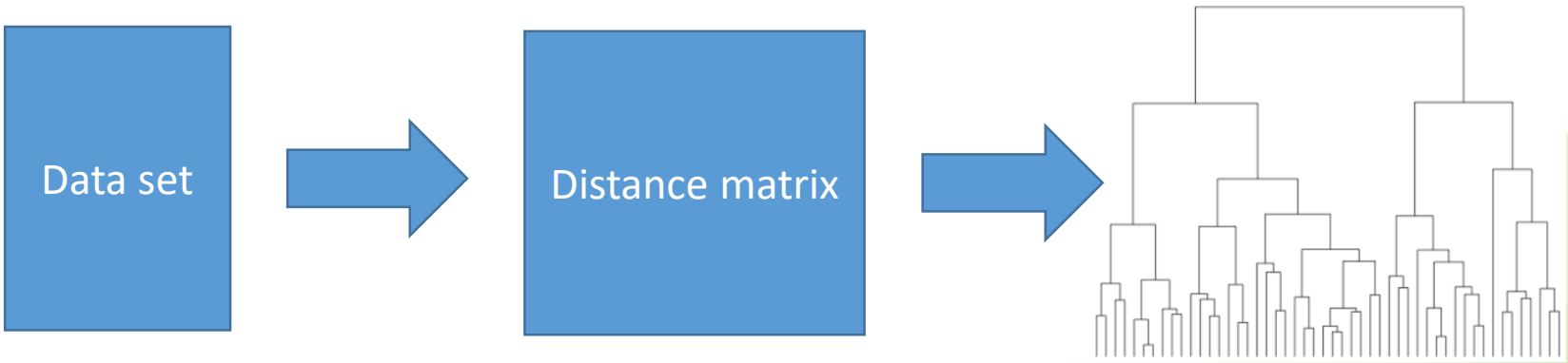
optimal alignment problem – Hungarian algorithm

Cluster analysis of annotated documents

Cluster analysis of documents

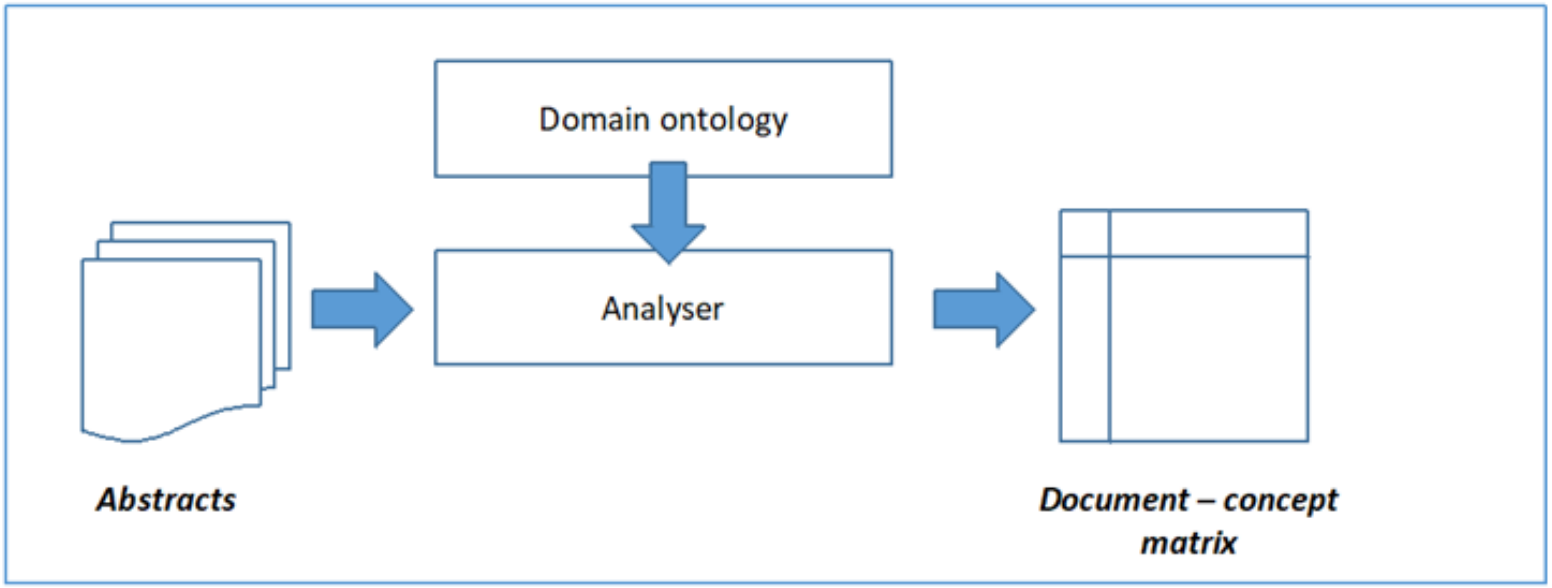
- Two main approaches in cluster analysis:
  - **distance-based,**
  - model-based.





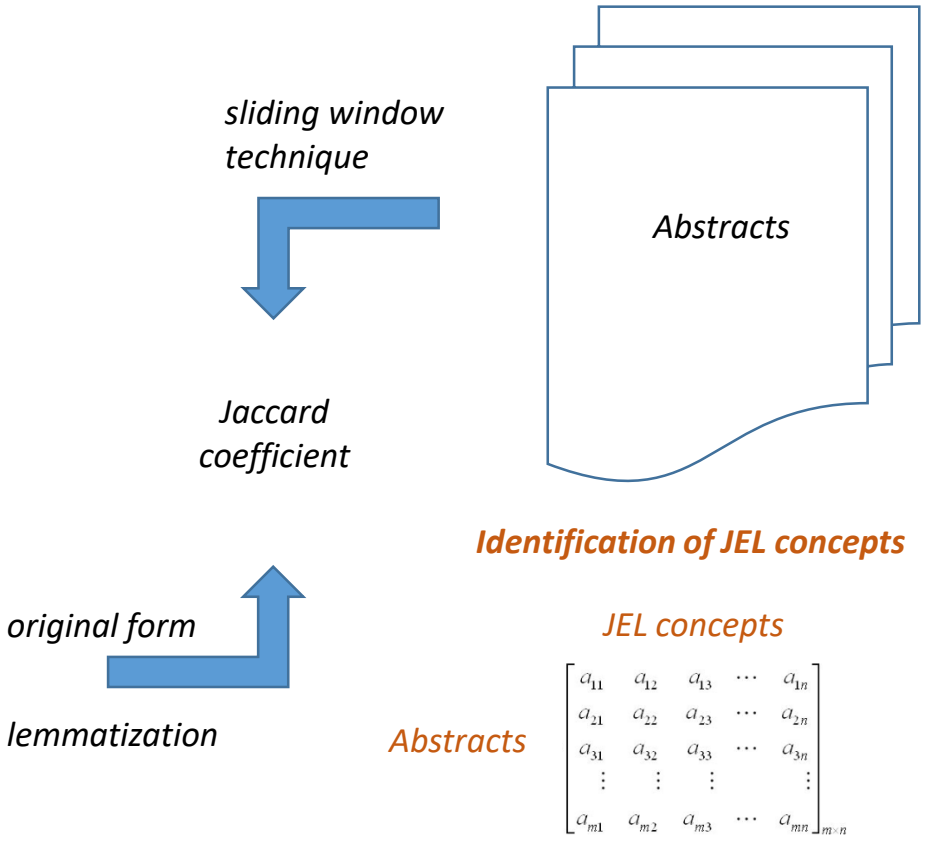
Practical aspects of cluster analysis of annotated documents (in the context of research productivity analysis)

The structure of the ontology-based analyzer

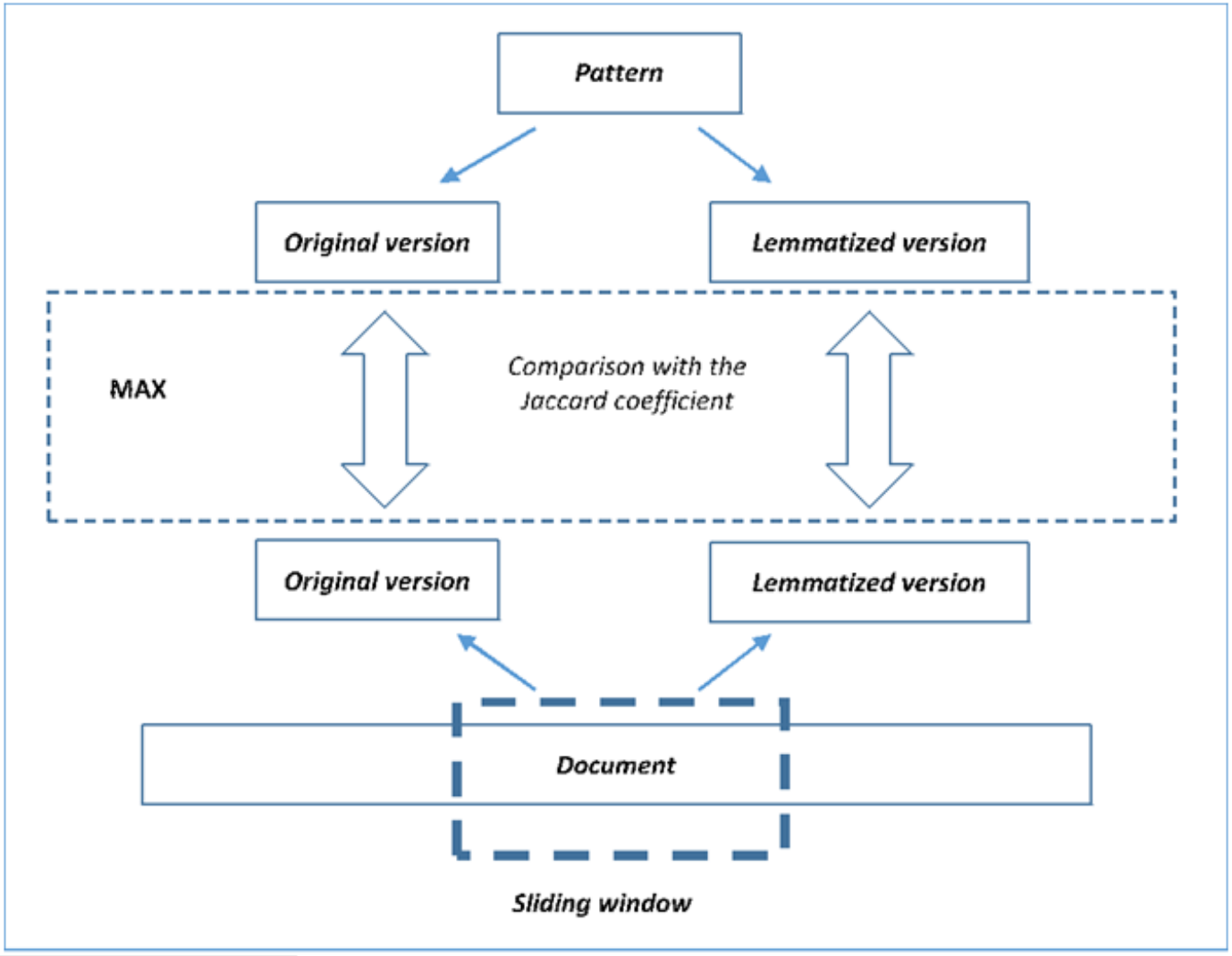


The structure of the ontology-based analyzer

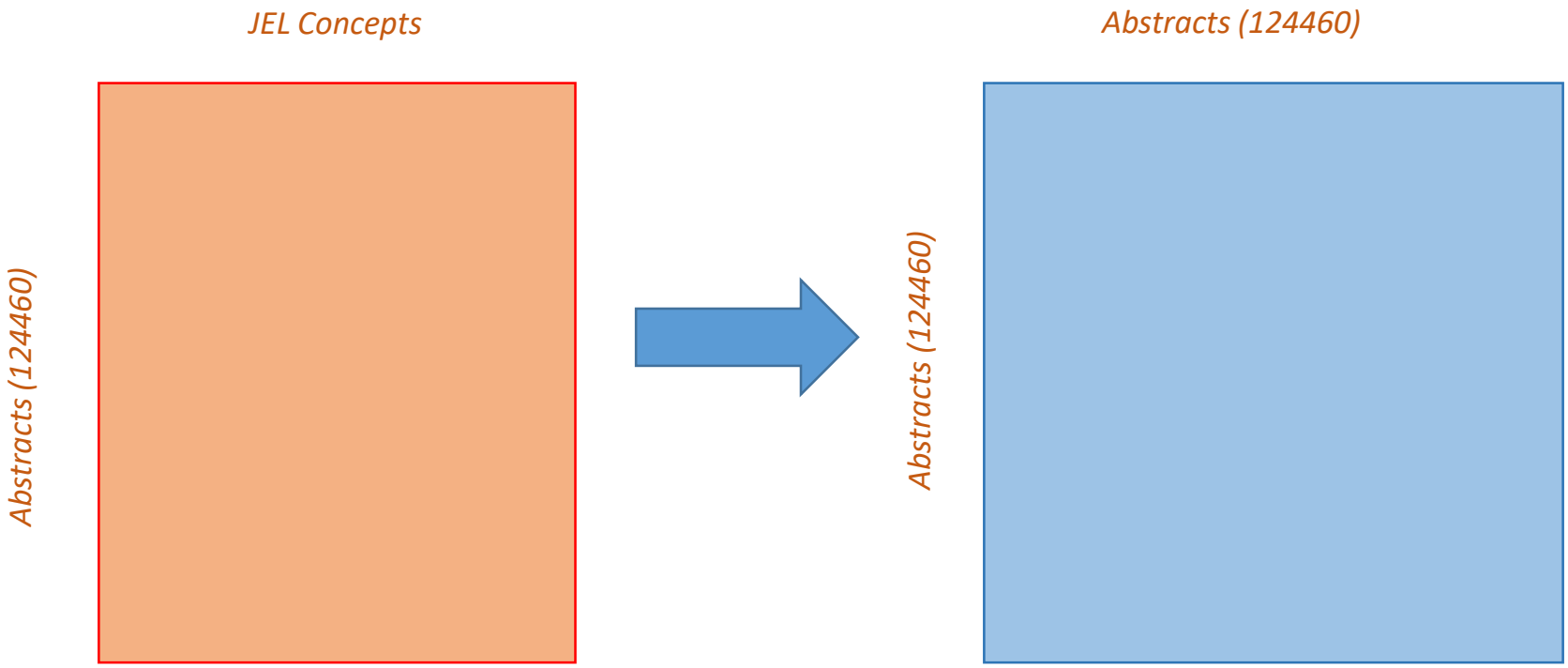
```
JEL_C.1:
    Description:
        - "Econometric and Statistical Methods and Methodology: General"
    Keywords_EN:
        - "#Correlation"
        - "#Coskewness"
        - "#Covariance"
        - "#Econometric #(Methods|Models|Analysis)"
        - "#Econometric #Theory"
        - "#Econometrics"
        - "#Probabilities"
        - "#[Statistical Methods]"
        - "#[Statistical analysis]"
        - "#[Statistical research]"
    JEL_C.1.0:
        Description:
            - "General"
        Keywords_EN:
            - "#[Econometric Methods]"
            - "#[Econometric Theory]"
            - "#Econometrics"
            - "#[Statistical Methods]"
    JEL_C.1.1:
        Description:
            - "Bayesian Analysis: General"
        Keywords_EN:
            - "#Bayesian"
    JEL_C.1.2:
        Description:
            - "Hypothesis Testing: General"
        Keywords_EN:
            - "#Chi-Squared #Test"
            - "#Dickey-Fuller #Test"
            - "#F #Test"
            - "#[Hypothesis Testing]"
            - "#Lagrange Multiplier #Test"
            - "#Likelihood #Ratio #Test"
            - "#Nonparametric #Test"
            - "#Statistical #Testing"
            - "#Test #Statistics"
```



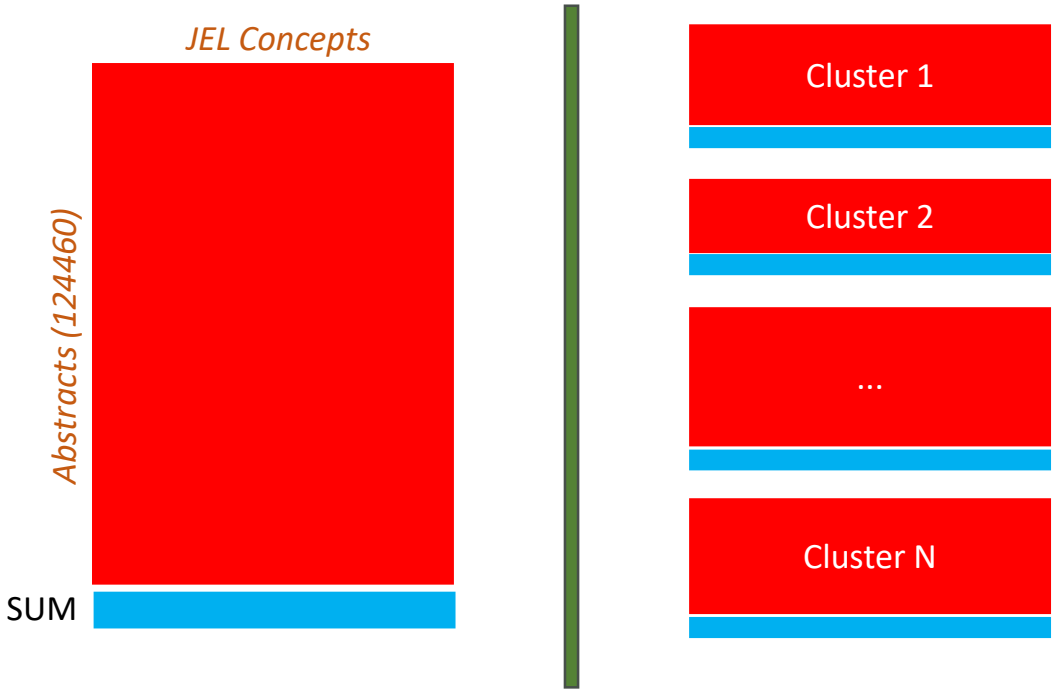
Linguistic issues in pattern analysis



Similarity of dodocuments



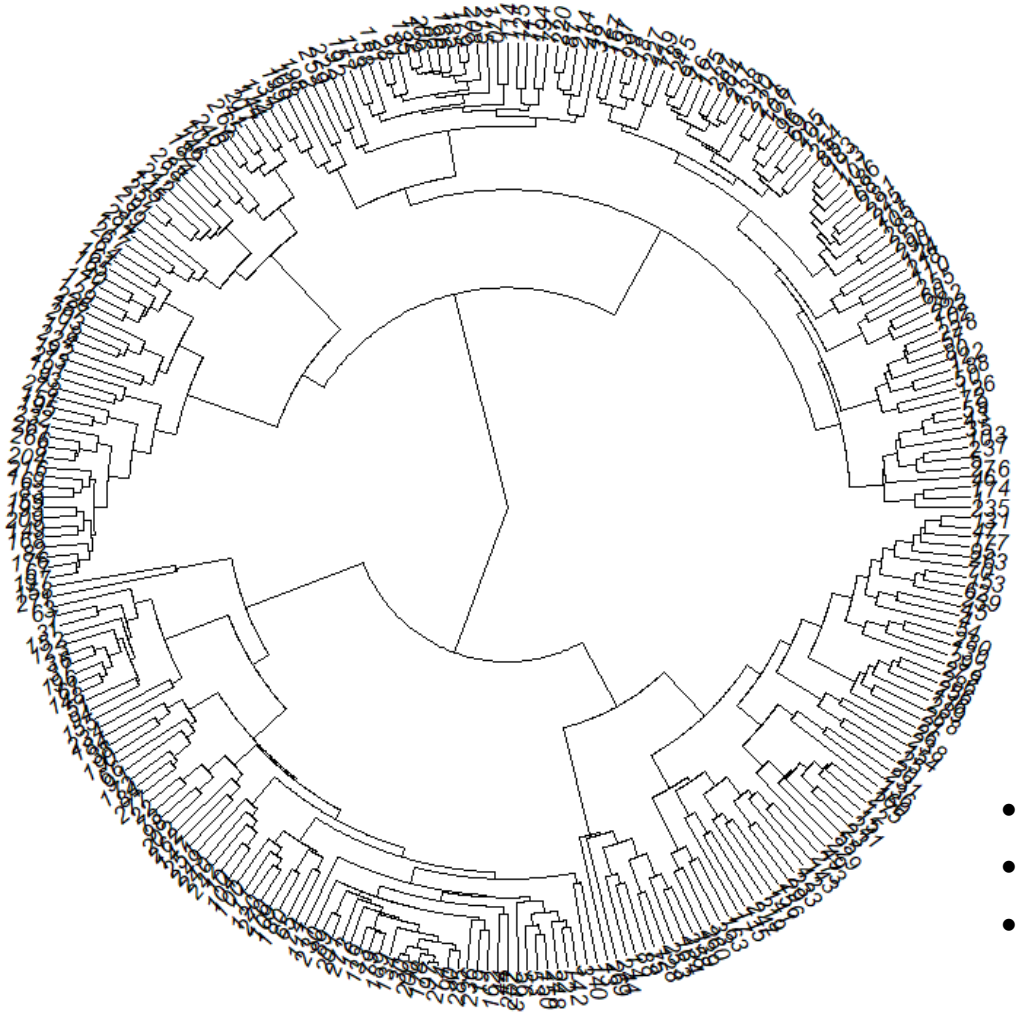
Cluster analysis of abstracts



Projects – dataset

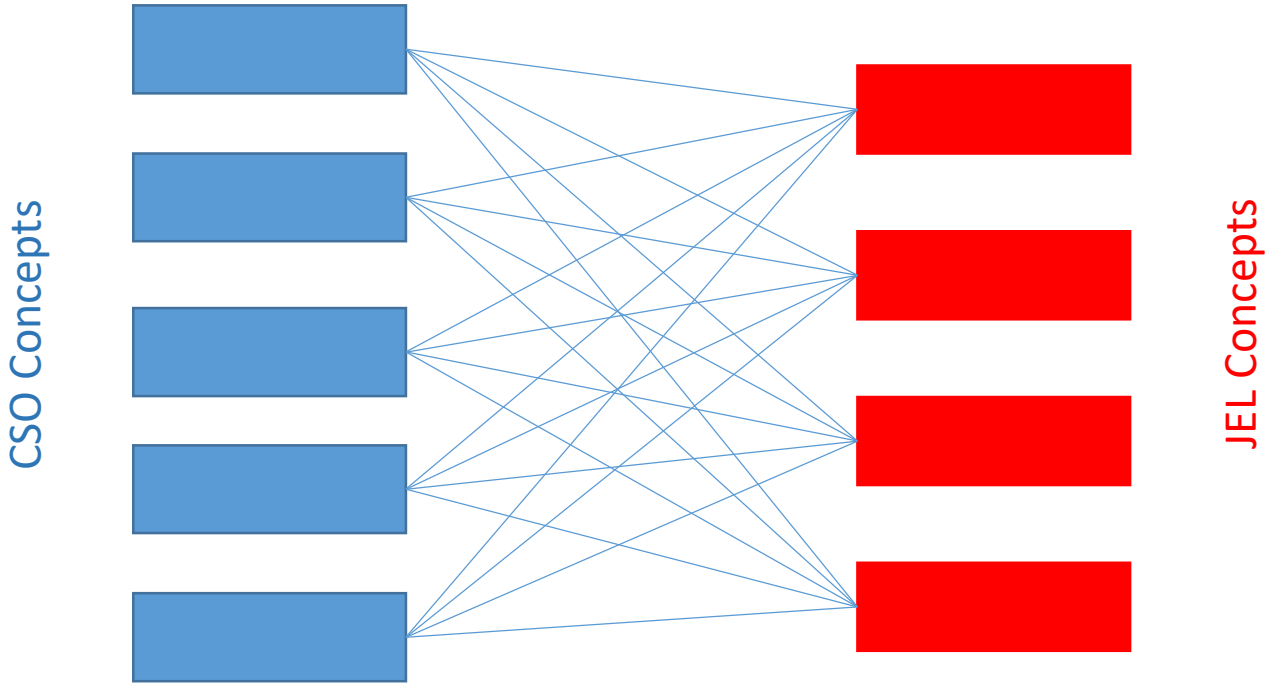
- Cordis database was used.
- Projects related to the area of economics (started in 2019 – 2021).
- Horizon and Framework Program projects were included.

Cluster analysis of projects' description



- annotated with JEL ontology
- Hamming distance
- Ward's method

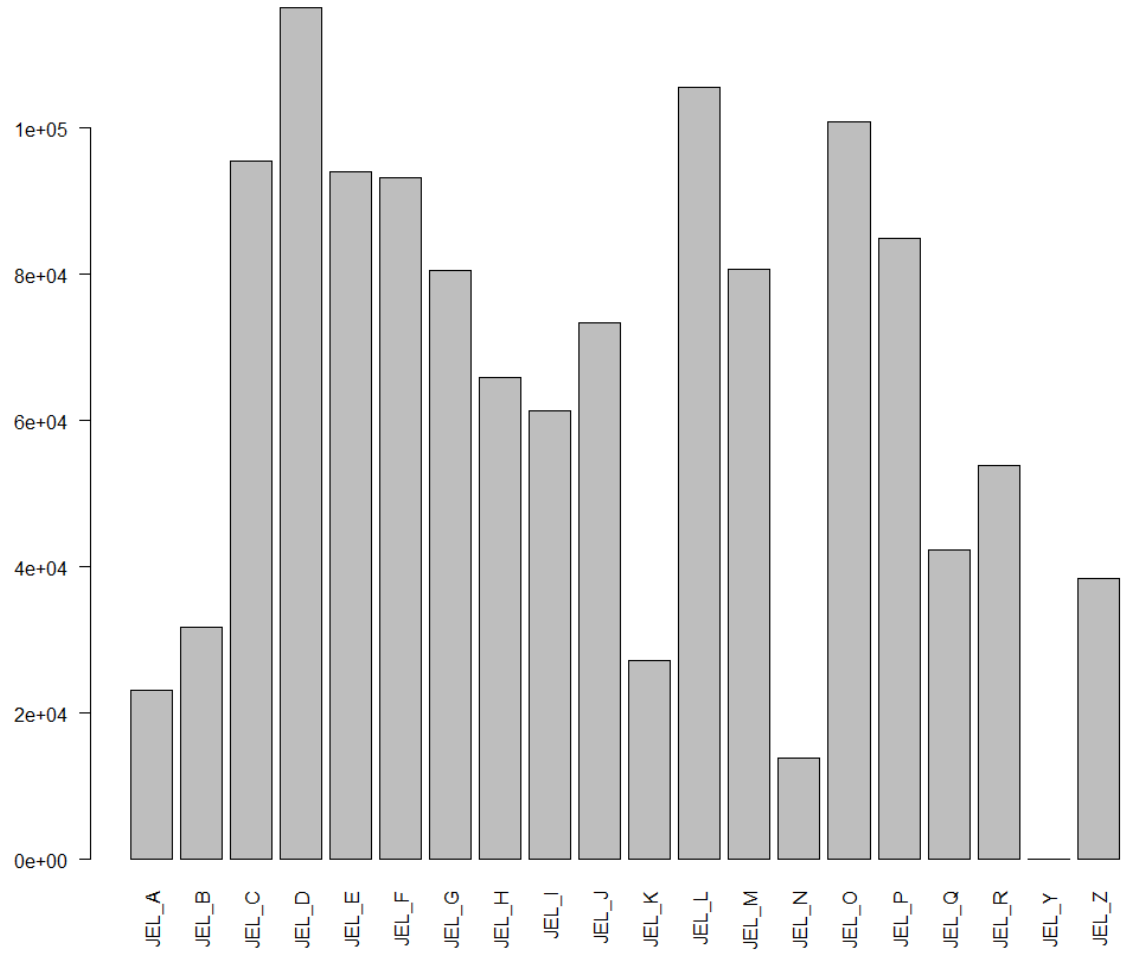
Bipartite model of interdisciplinarity



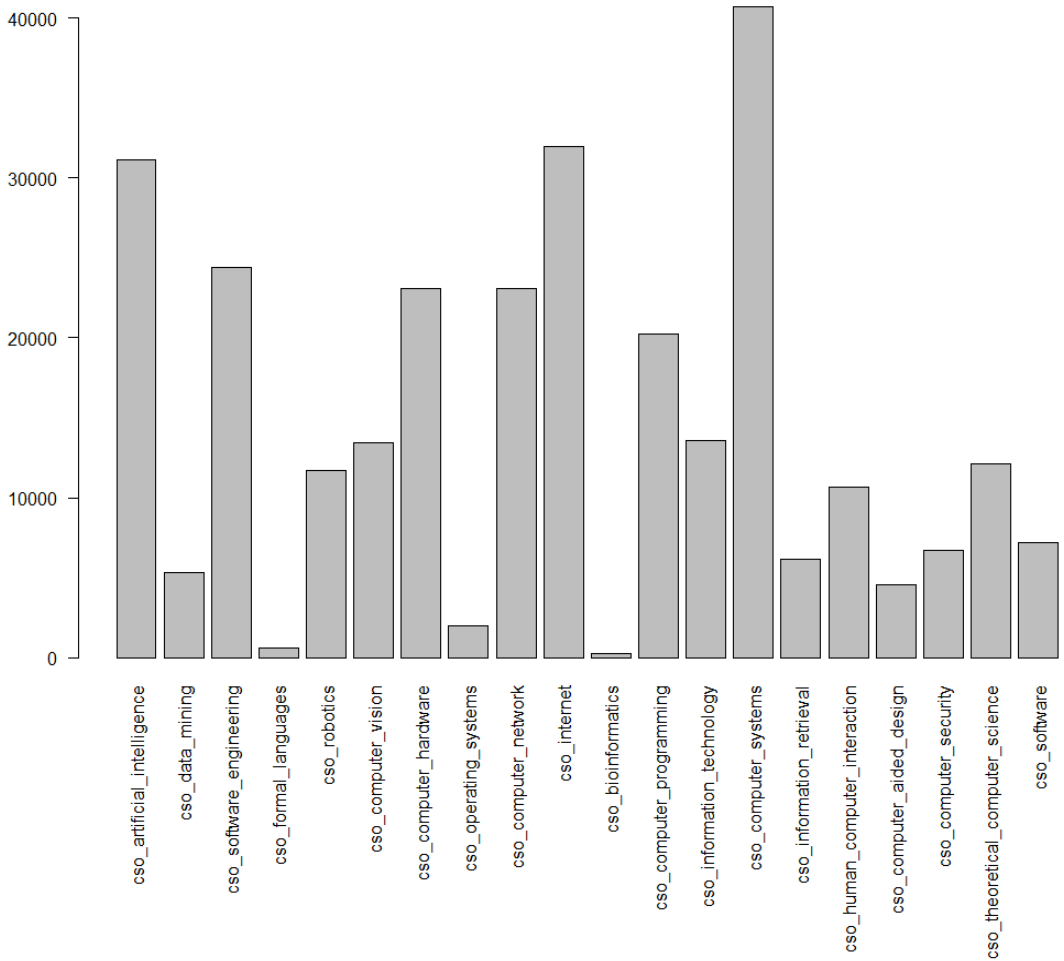
Areas of analysis:

- subarea strength
- subarea specificity
- graph model modularity

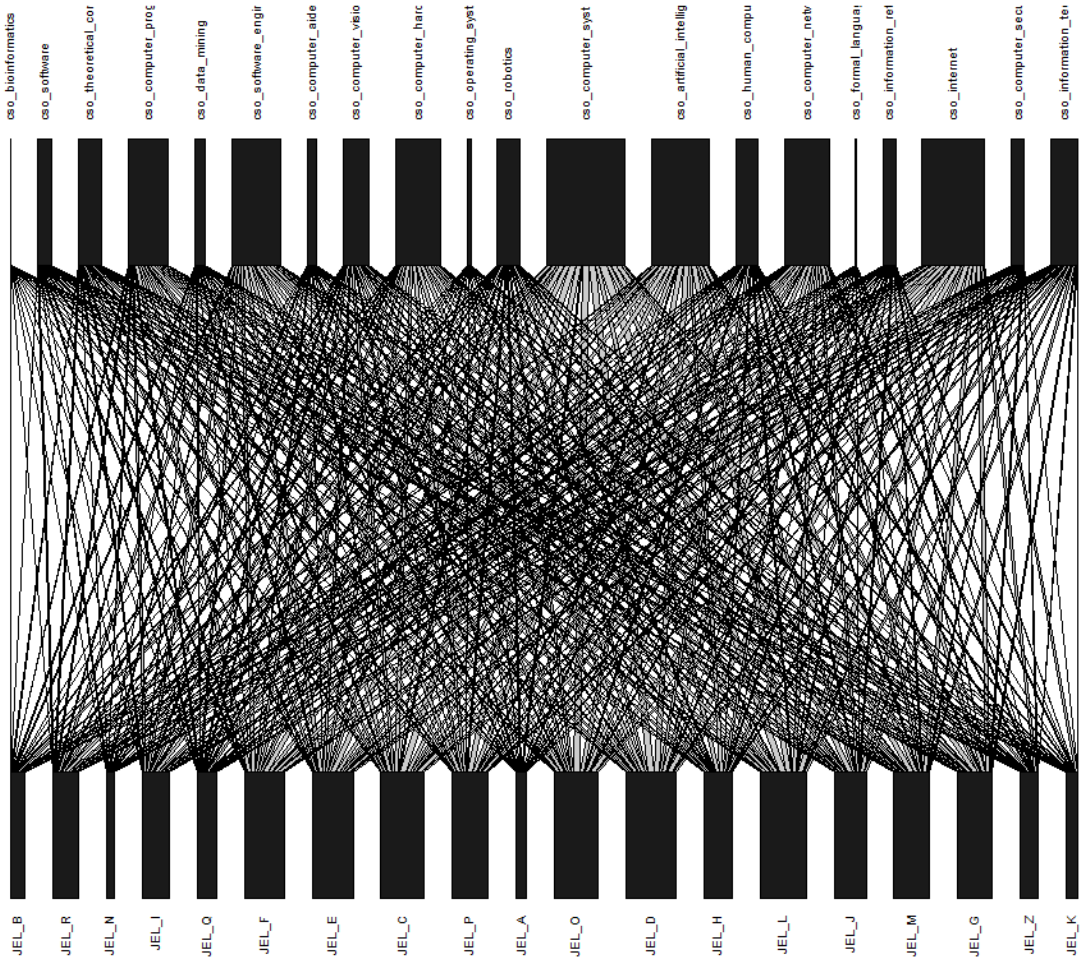
The distribution of documents over the main concepts from the JEL ontology



The distribution of documents over the main concepts from the CSO ontology



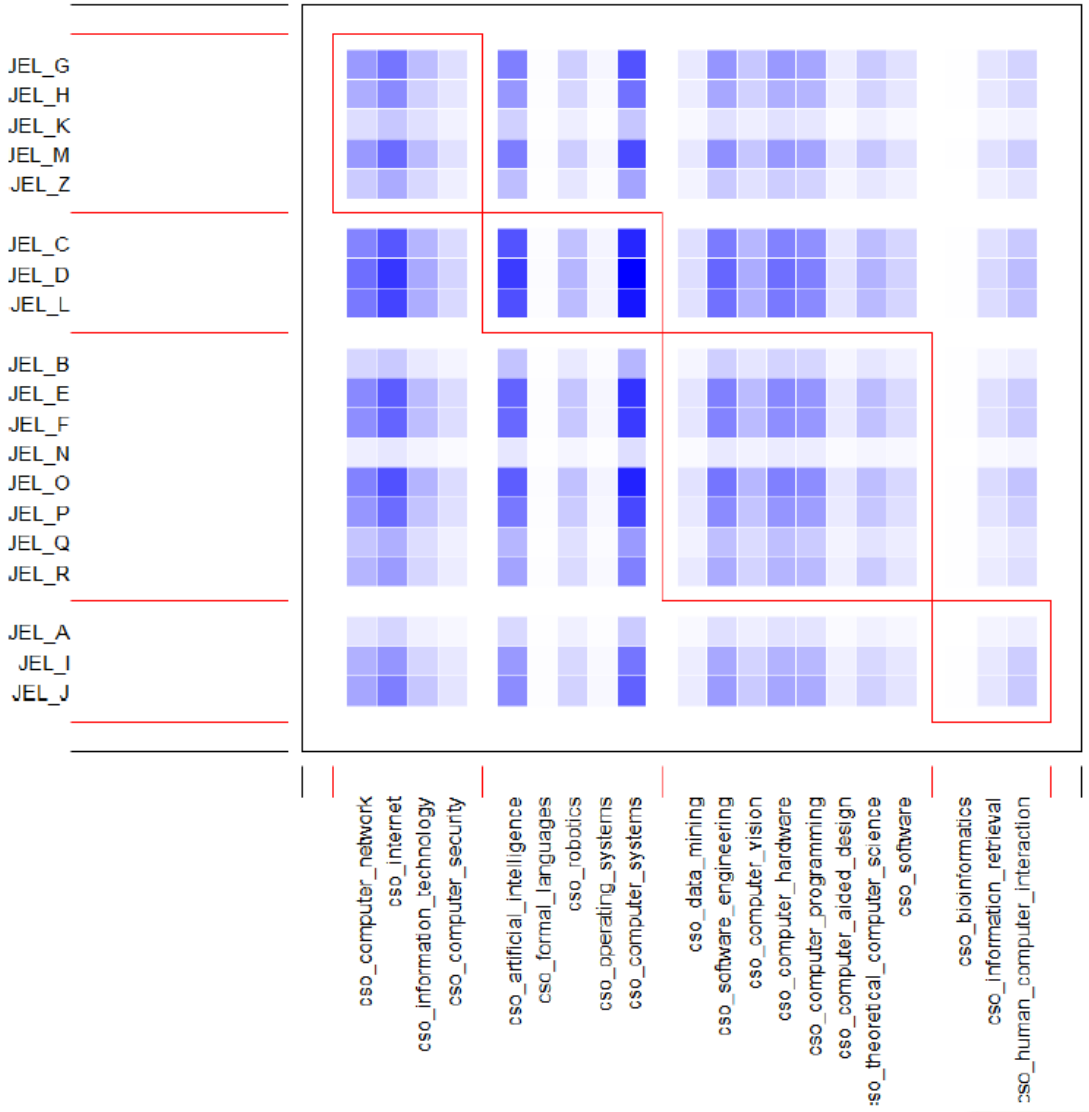
Relationships between main subareas from the JEL and CSO ontology



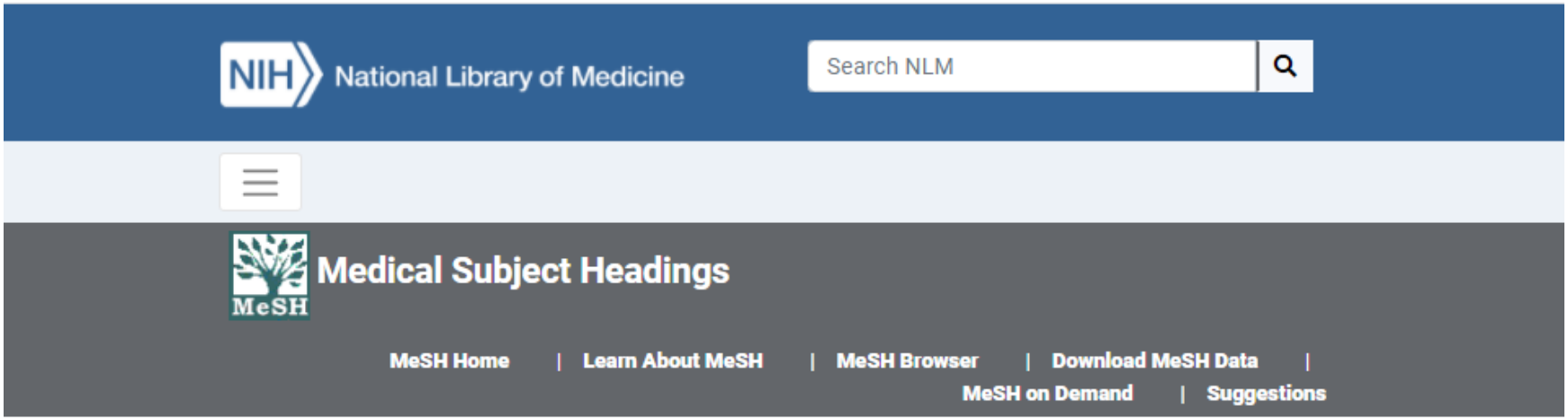
The strongest connections between main subareas from the JEL and the CSO ontology

Connection	Number of abstracts
JEL_D/cso_computer_systems	39342
JEL_L/cso_computer_systems	36160
JEL_O/cso_computer_systems	34087
JEL_C/cso_computer_systems	33515
JEL_E/cso_computer_systems	31403
JEL_D/cso_internet	31044
JEL_F/cso_computer_systems	30454
JEL_D/cso_artificial_intelligence	30178
JEL_L/cso_internet	28891
JEL_P/cso_computer_systems	28192

Communities in the JEL - CSO bipartite graph



MeSH ontology



The structure of concepts in MeSH ontology

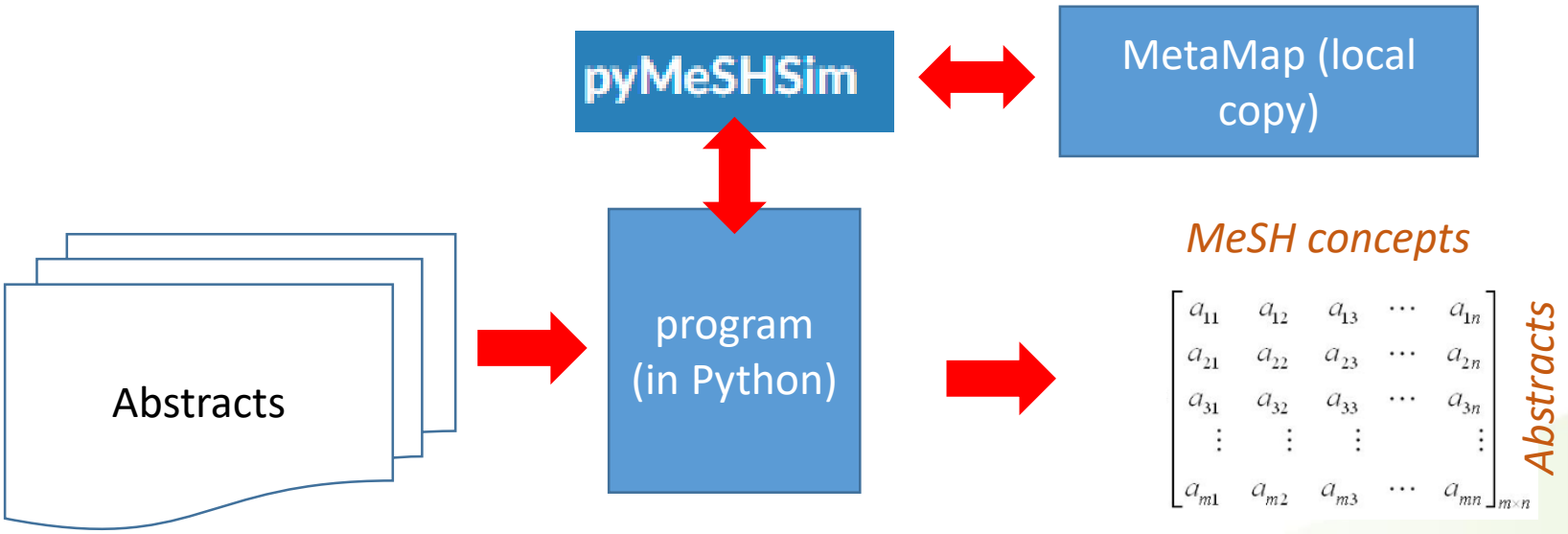
- Anatomy [A]
- Organisms [B]
- Diseases [C]
- Chemicals and Drugs [D]
- Analytical, Diagnostic and Therapeutic Techniques, and Equipment [E]
- Psychiatry and Psychology [F]
  - Behavior and Behavior Mechanisms [F01]
  - Psychological Phenomena [F02]
  - Mental Disorders [F03]
  - Behavioral Disciplines and Activities [F04]
- Phenomena and Processes [G]
- Disciplines and Occupations [H]
- Anthropology, Education, Sociology, and Social Phenomena [I]
- Technology, Industry, and Agriculture [J]
- Humanities [K]
- Information Science [L]
- Named Groups [M]
- Health Care [N]
- Publication Characteristics [V]
- Geographicals [Z]

Tools supporting ontology-based analysis with MeSH ontology

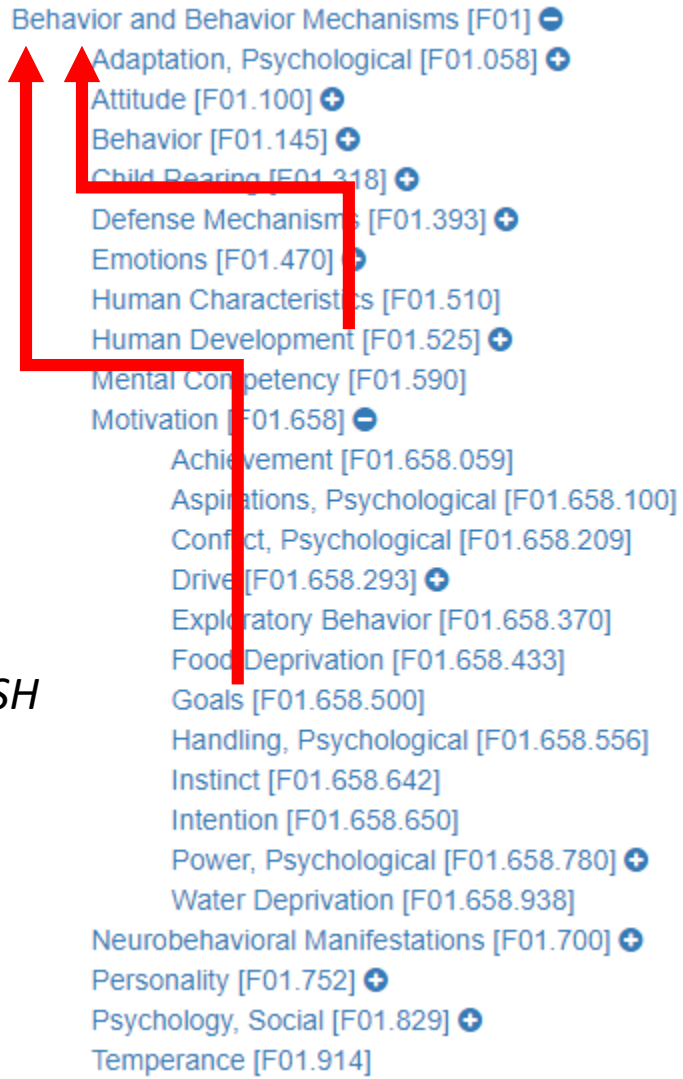
NIH National Library of Medicine

MeSH Medical Subject Headings

MetaMap - A Tool For Recognizing UMLS Concepts in Text

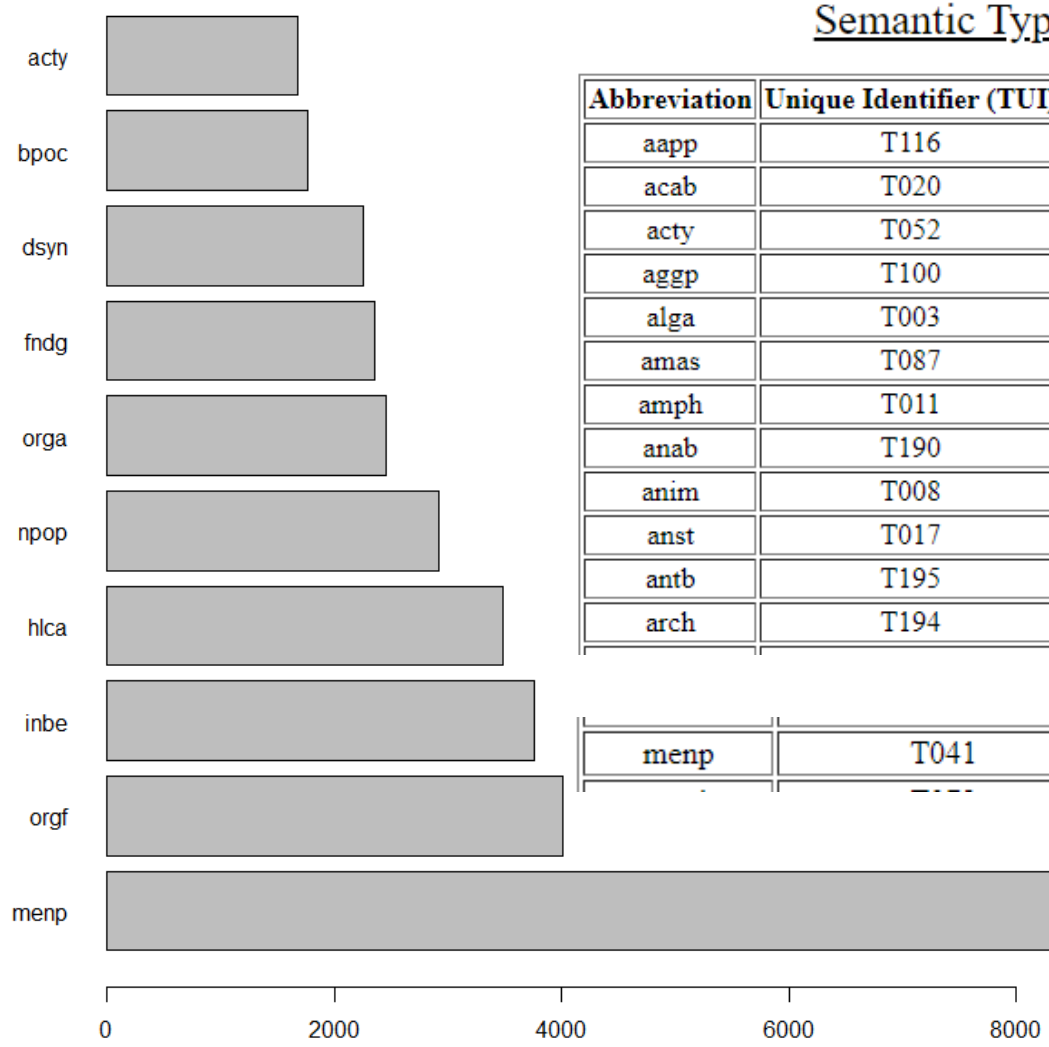


Levels of analysis



Projection to main concepts for the MeSH ontology

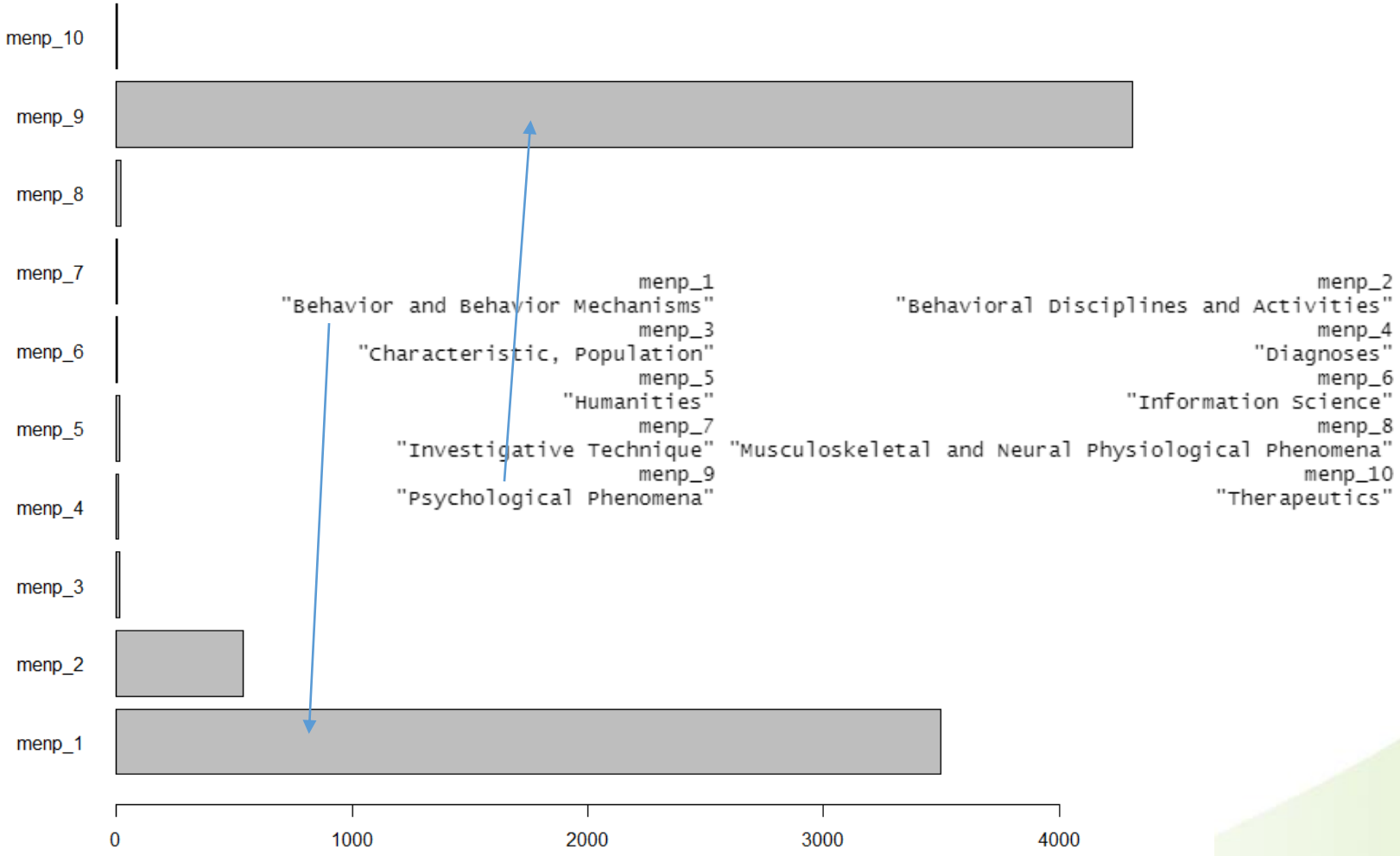
Analysis of concepts related to mental processes



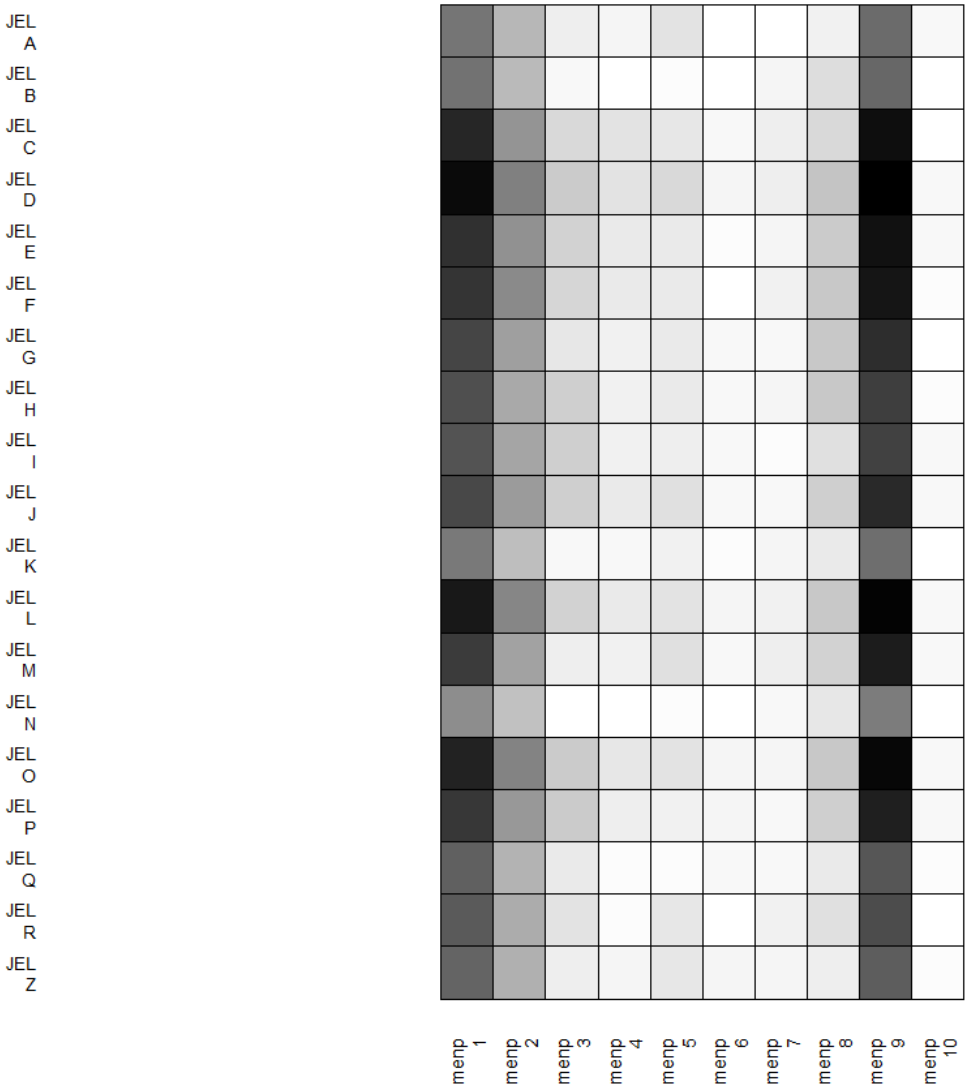
Semantic Types used in MMTx

Abbreviation	Unique Identifier (TUI)	Full Name
aapp	T116	Amino Acid, Peptide, or Protein
acab	T020	Acquired Abnormality
acty	T052	Activity
aggp	T100	Age Group
alga	T003	Alga
amas	T087	Amino Acid Sequence
amph	T011	Amphibian
anab	T190	Anatomical Abnormality
anim	T008	Animal
anst	T017	Anatomical Structure
antb	T195	Antibiotic
arch	T194	Archaeon
menp	T041	Mental Process

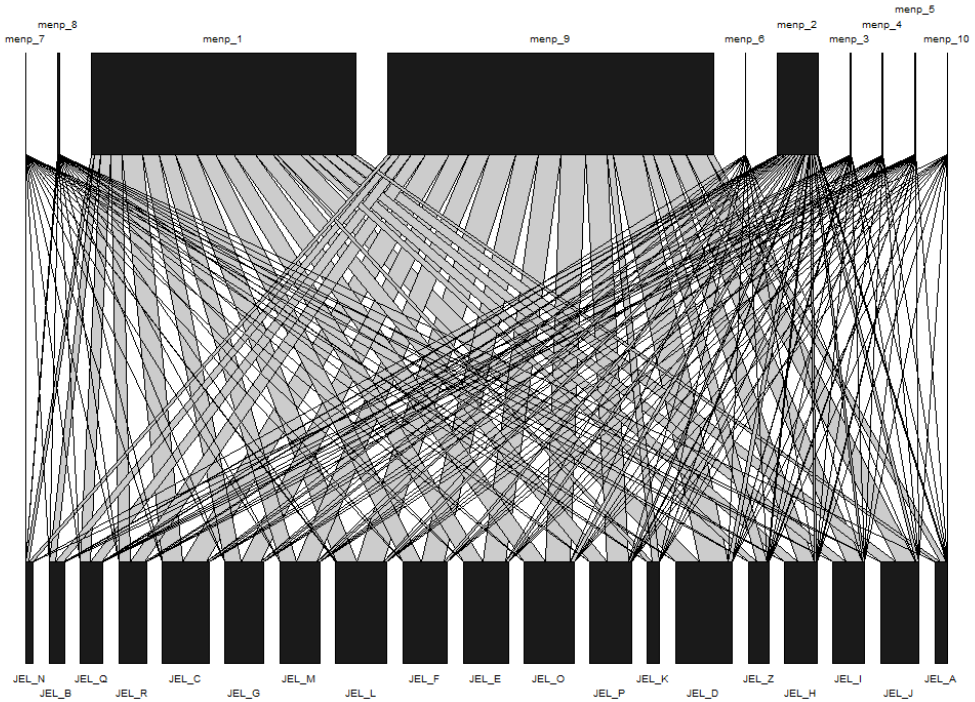
The significance of concepts related to mental processes

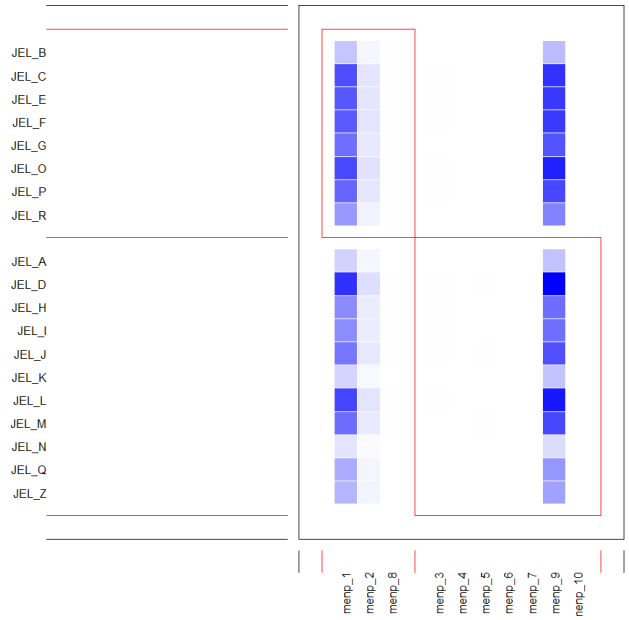


Relationships between JEL and MeSH-menp concepts



Relationships between JEL and MeSH-menp concepts





Conclusions

- Ontology-based approach improve the quality of results of exploratory text analysis
- For many areas of science domain ontologies were proposed.
- UDC system is the most popular ontology covers all disciplines of science
- Domain ontologies can be useful for analysis of multilingual documents
- R language is a good tool for implementing ontology-based solutions for exploratory text analysis

Paweł Lula, Janusz Tuchowski, Urszula Cieraszewska, Magdalena Talaga
Cracow University of Economics

Thank you for your attention!

9th International Conference "Distributed Computing and Grid Technologies in Science and Education" (GRID'2021)

5-9 July 2021