

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 than 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

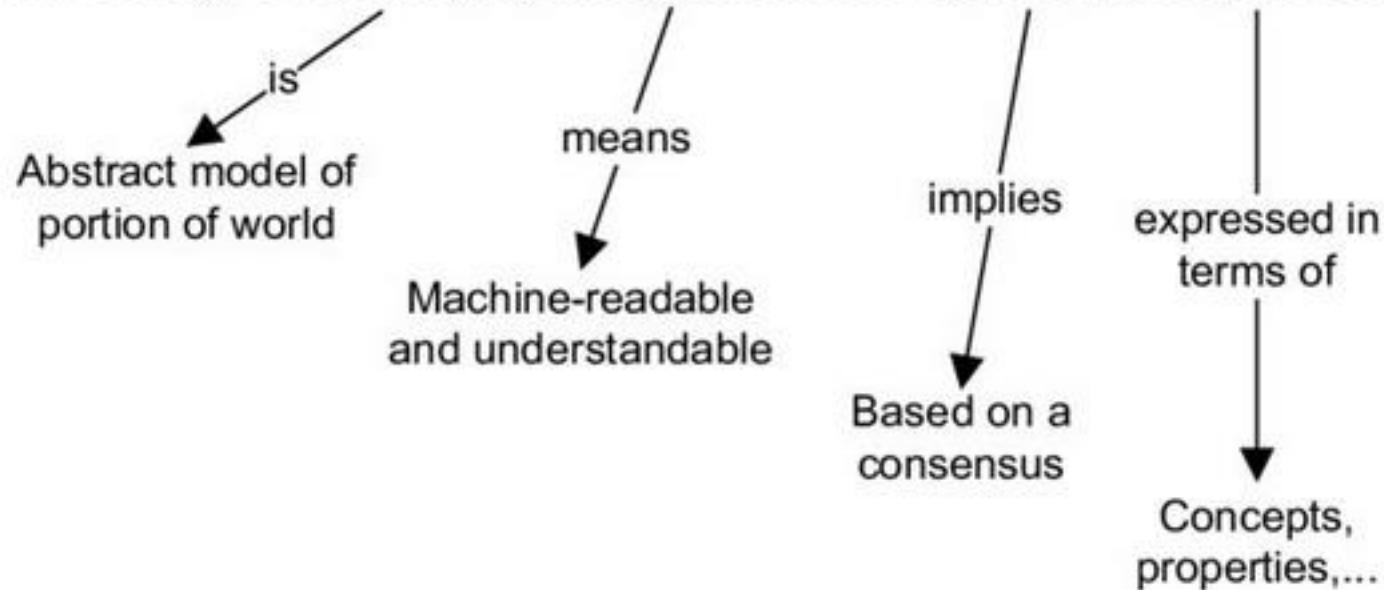


Thomas Robert "Tom" Gruber (born 1959):

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

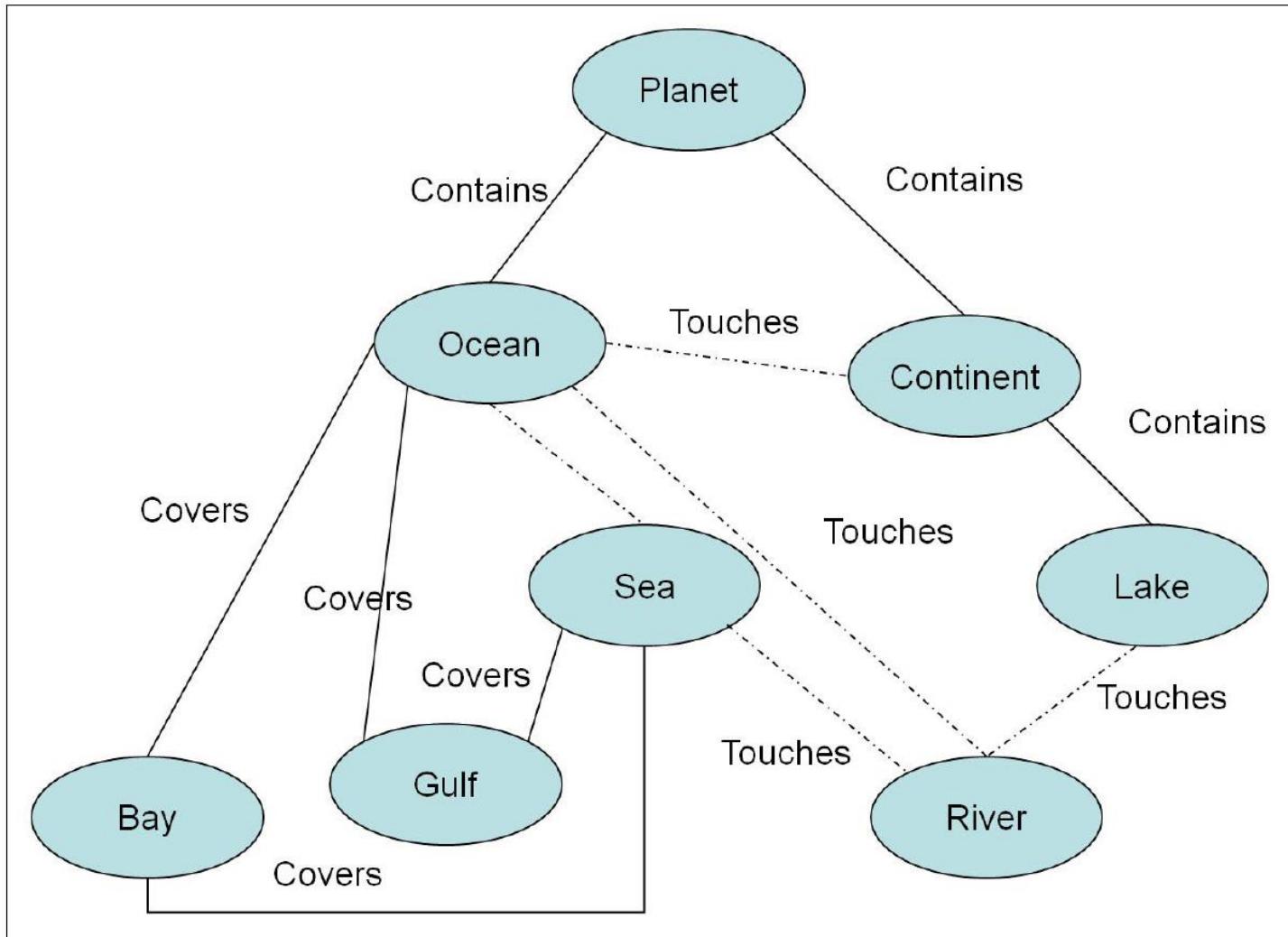
Ontologies

An ontology is a formal, explicit specification of a shared conceptualisation.



Ontology = the method of knowledge representation

Ontologies



Competency model (competency ontology)

```
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|kontakte|relacje|relacje)'
    negotiation_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'
```

YAML format

Types of ontologies

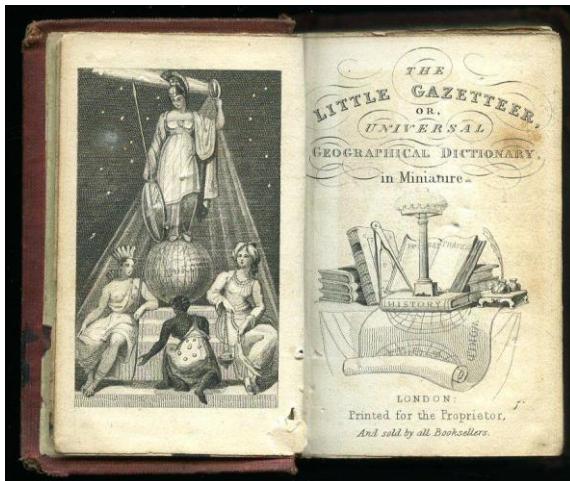
Types of ontologies

- linear
- hierarchical
- network

Linear ontologies

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

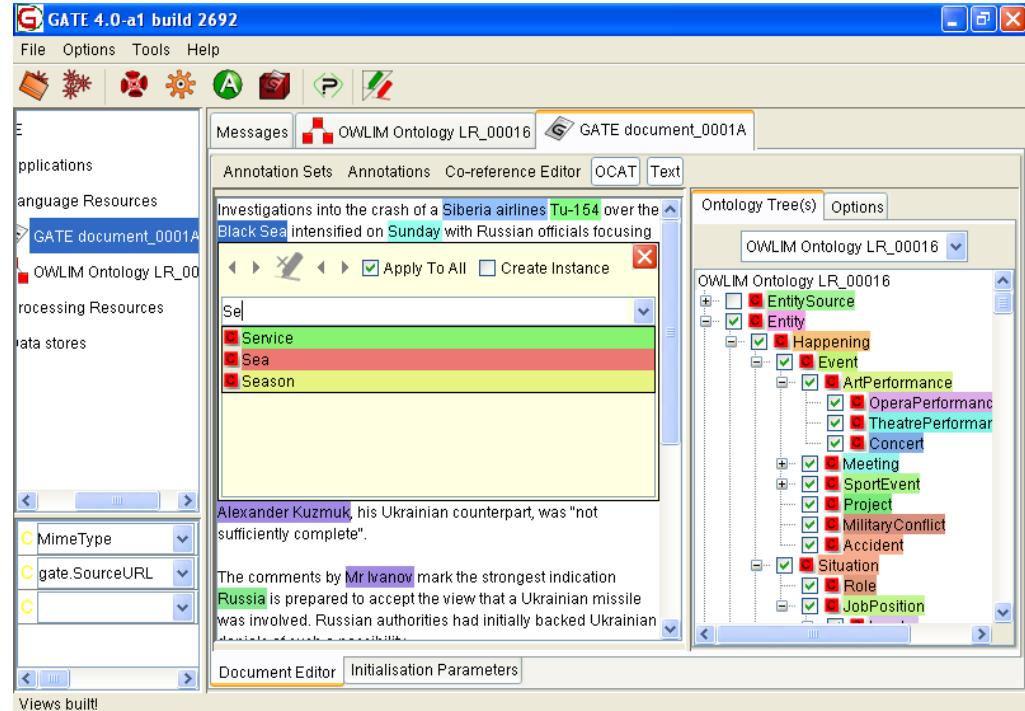
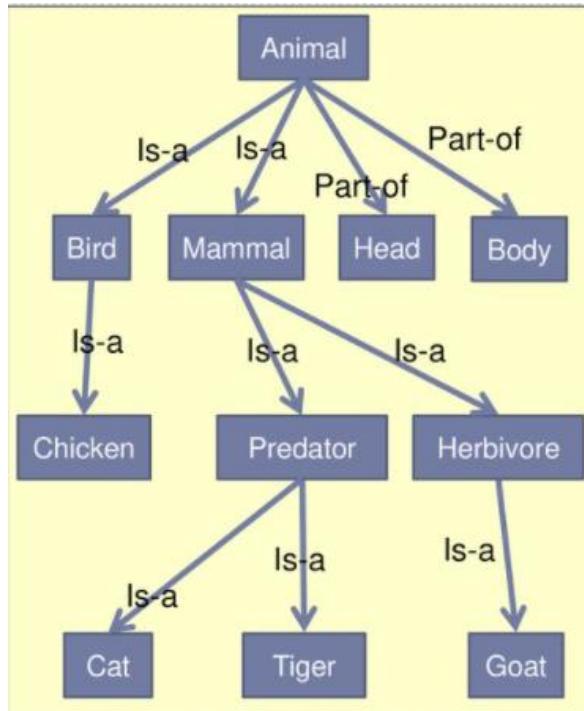
1. A geographical dictionary. 2. A directory in which the entries are arranged by geographical location. For example, a gazetteer of restaurants. can be abbreviated as gazetteer.



A screenshot of the GATE Developer software interface, specifically the Gazetteer Editor. The window shows a list of gazetteer entries in a table format. Red arrows point to the entries 'Abu Dhabi' and 'Accra' in the list. The table has columns for 'List name', 'Major', and 'Minor'. The 'Value' column lists various city names. The interface includes a toolbar at the top and a sidebar with resource features.

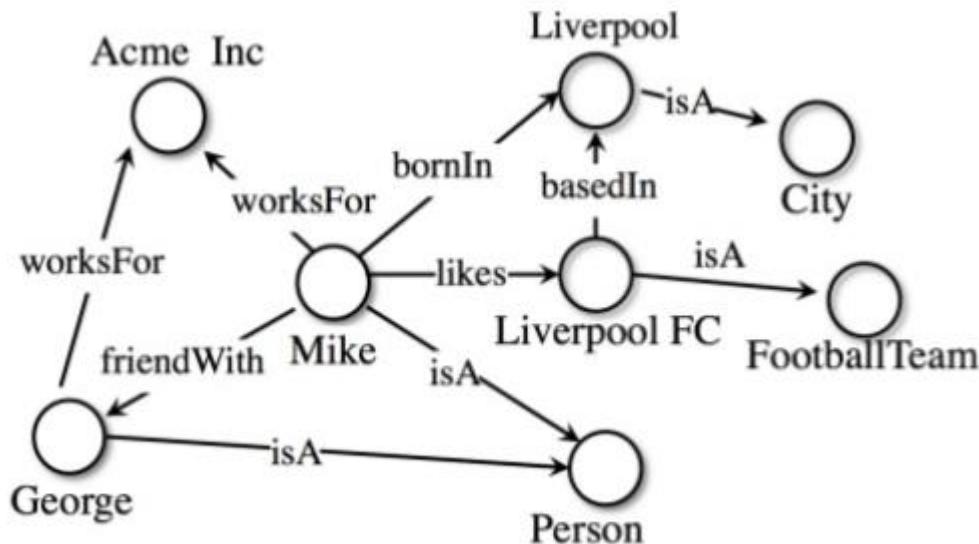
Hierarchical ontology

- 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



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](#))

The screenshot shows the ACM Computing Classification System interface. At the top, there's a navigation bar with a 'CCS' button, a large dark arrow pointing right, and a 'Assign this CCS Concept' button with a plus sign. Below the navigation bar, there's a sidebar with a 'CCS' tab selected. The sidebar contains several category sections with arrows indicating expandable content:

- General and reference
- Hardware
- Computer systems organization

Under the 'Computer systems organization' section, there's a 'Architectures' section with the following sub-categories:

- Serial architectures
- Reduced instruction set computing
- Complex instruction set computing

CCS Concept

You haven't added any CCS Concept yet.

CSO as a graph ontology

The screenshot shows the CSO Portal interface. At the top left is the CSO logo with three teal nodes connected by lines. To its right is the text "Computer Science Ontology - Portal". Below the logo is a search bar containing the text "Search across 14K topics and over 159K relationships." A button labeled "computer science" is highlighted in a light blue box. In the main content area, the heading "Annotated document:" is followed by a large block of text about computer science. This text is annotated with several orange boxes highlighting specific terms: "computer science", "computer science", "software", "geometry", "computer science", "computer programming", "Artificial intelligence", "learning", and "computer science". The text discusses the study of algorithmic processes, the range of topics from theoretical studies to practical issues, and various fields like computation, graphics, and programming. It also mentions AI, learning, and the Turing Award. Below this is a section titled "Topics found:" with a table showing extracted topics using the Syntactic module. The table has four columns: "artificial intelligence", "computer hardware", "computer programming", and "computer science". The second row contains "geometry", "learning", "programming languages", and "software". A separate section below shows topics extracted using the Semantic module, with the same four columns and rows.

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

Universal Decimal Classification
summary

Russian (Русский)

ВЕРХ ЗНАКИ ОПРЕДЕЛИТЕЛИ 0 1 2 3 4 5 6 7 8 9

АННОТАЦИЯ РУКОВОДСТВО АЛФАВИТ ЭКСПОРТ СООТВЕТСТВИЯ ПЕРЕВОДЫ

развернуть всё | свернуть всё

3 ОБЩЕСТВЕННЫЕ НАУКИ

- 303 Методы общественных наук. Методы и виды социологических исследований
- 304 Социальные вопросы. Социальная практика. Культурная жизнь. Образ жизни
- 305 Социологические исследования вопросов пола
- 308 Социология. Описательное изучение жизни общества
- + -311 Теория статистики. Статистические методы
- + -314/316 Общество
- + -32 Политика
- 33 Экономика. Народное хозяйство. Экономические науки
 - + -330 Экономические науки в целом. Политическая экономия
 - 331 Труд. Работодатели. Трудящиеся. Наука о труде. Экономика труда. Организация труда
 - 331.1 Теория и организация труда. Взаимоотношения между предприятием и персоналом
 - 331.2 Заработка плата. Оклады, премии, надбавки к заработной плате
 - 331.3 Пункты в трудовом договоре, не касающиеся заработной платы
 - 331.4 Производственная среда. Организация рабочих мест. Охрана труда
 - 331.5 Рынок труда. Занятость
 - 332 Региональная (территориальная) экономика. Земельный (аграрный) вопрос
 - 334 Формы организации и сотрудничества в экономике
 - + -336 Финансы. Государственные финансы. Финансы государственного сектора
 - + -338 Экономическое положение. Экономическая политика. Управление и планирование
 - + -339 Торговля. Международные экономические отношения. Мировое хозяйство
- + -34 Право. Юридические науки
- + -35 Государственное административное управление. Военное дело
- + -36 Обеспечение духовных и материальных жизненных потребностей. социальное развитие
- + -37 Воспитание. Обучение. Образование
- + -39 Этнография. Жизнь народа. Обычаи. Образ жизни. Фольклор

331

Труд. Работодатели. Трудящиеся. Наука о труде. Экономика труда. Организация труда

331.3

Пункты в трудовом договоре, не касающиеся заработной платы

Включая: Hours of work. Flexitime (flexitime). Shift work. Part-time working. Rest periods, breaks. Career development. In-service training. Apprenticeships. Mentoring. Retraining

The UDC Summary (UDCS) provides a selection of around 2,600 classes from the whole scheme which comprises more than 70,000 entries. Please send questions and suggestions to udcs@udcc.org

The data provided in this Summary is released under the Creative Commons Attribution Share Alike 3.0 license [[more](#)]



For complete UDC schedules see [UDC Online Hub](#)



Distances between concepts defined in ontologies

Distances for linear ontologies (gazetteers)

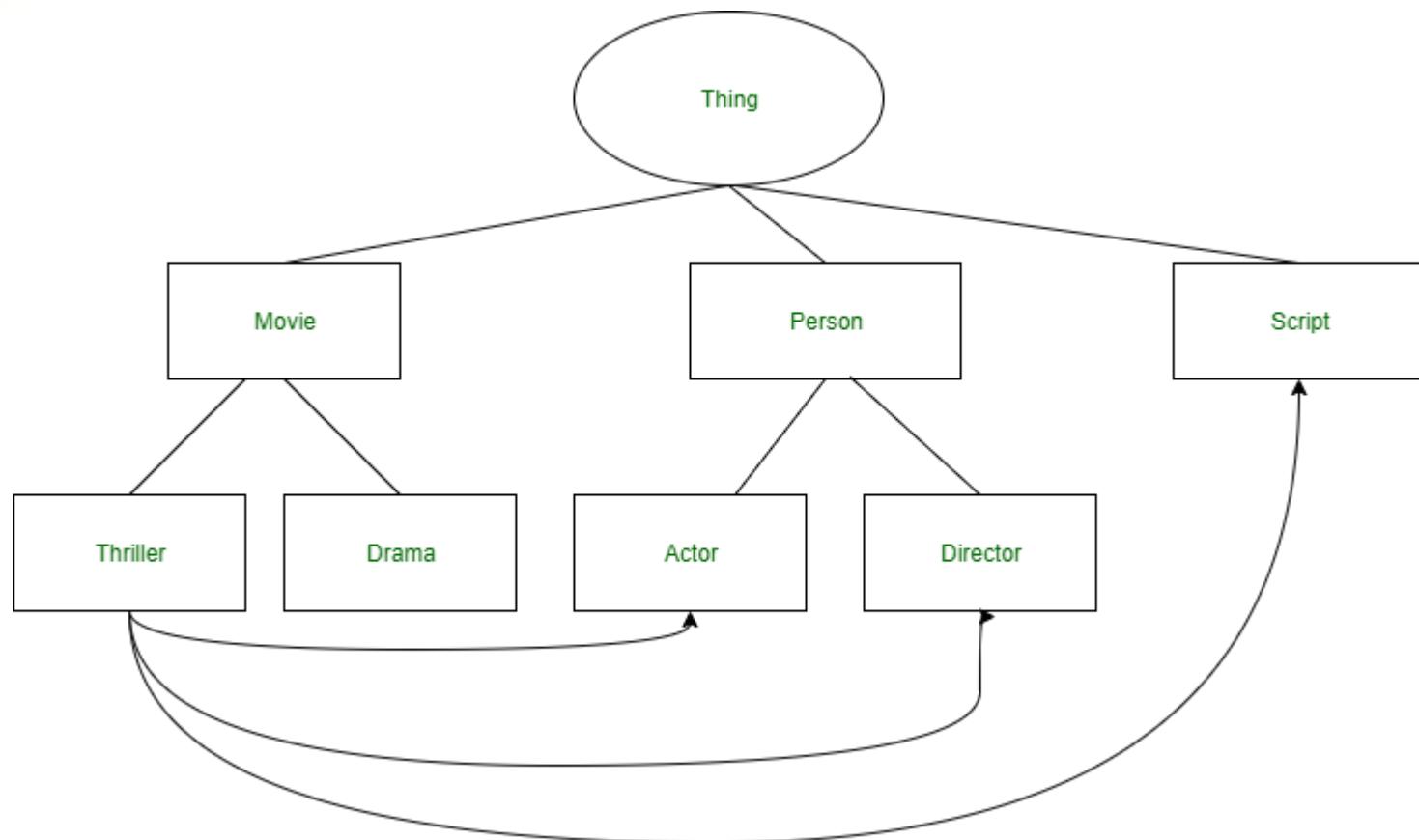


$C_i = C_j$

or

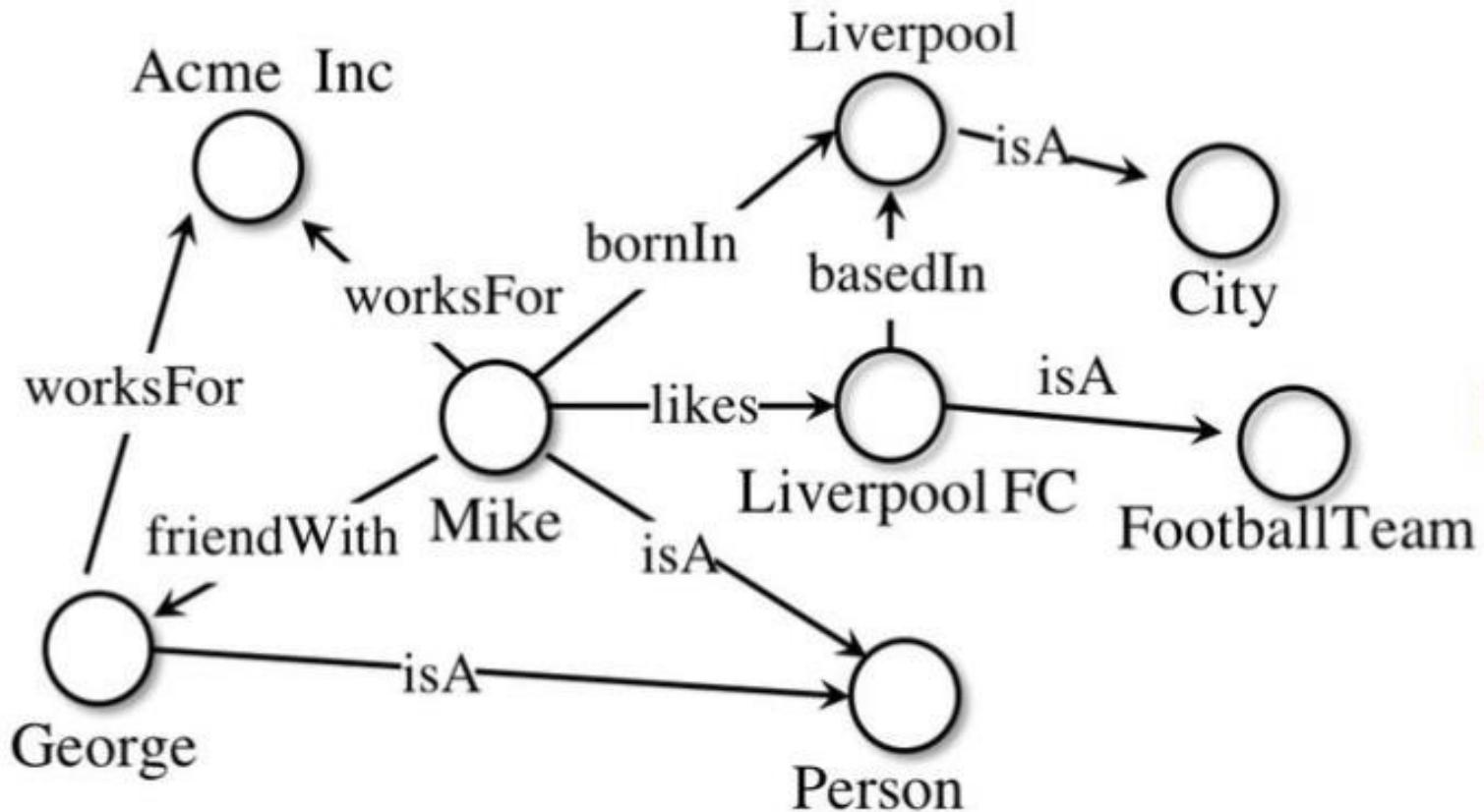
$C_i \neq C_j$

Distances for hierarchical ontologies



- based on path length,
- based on information theory

Distances for network ontologies



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

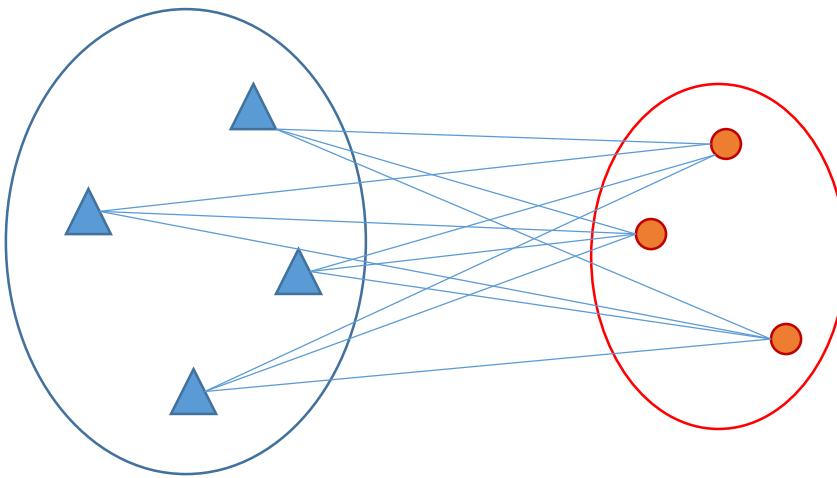
Distances between documents

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.

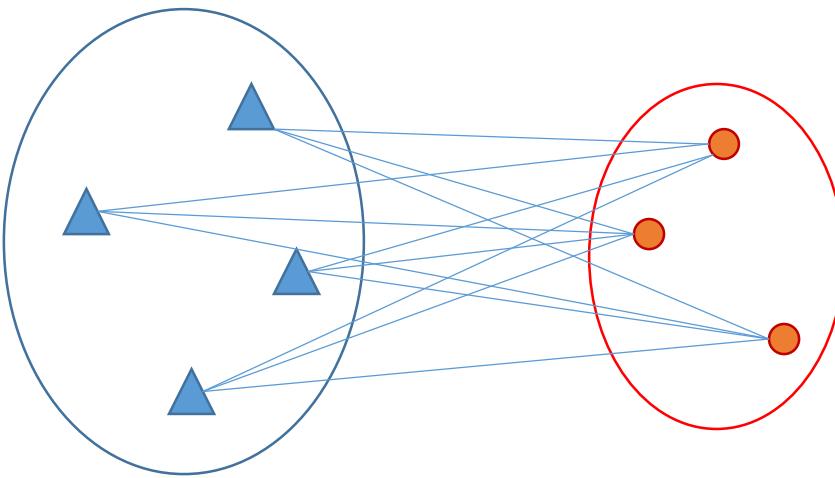
In computing, a database is an organized collection of data stored and accessed electronically from a computer system. Where databases are more complex they are often developed using formal design and modeling techniques. The **database management** system (DBMS) is the **software** that interacts with 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



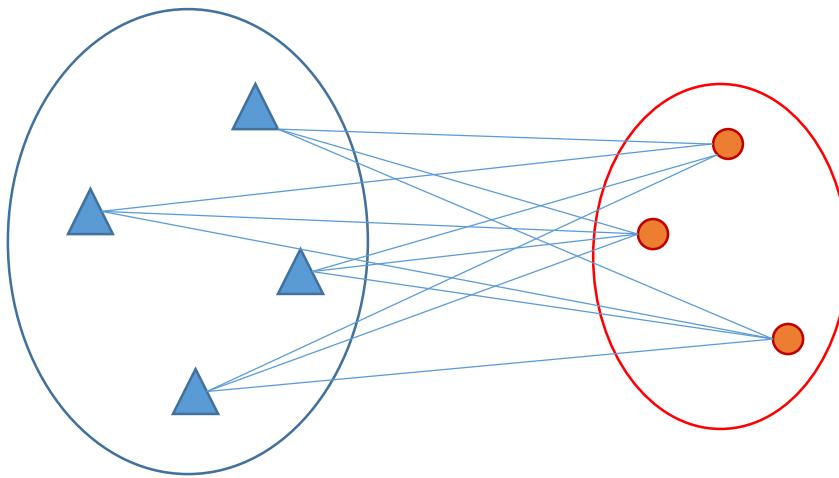
$$sim(Zb_1, Zb_2) = avg(C_i, C_j), C_i \in Zb_1, C_j \in Zb_2$$

Distances between documents



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

Distances between documents



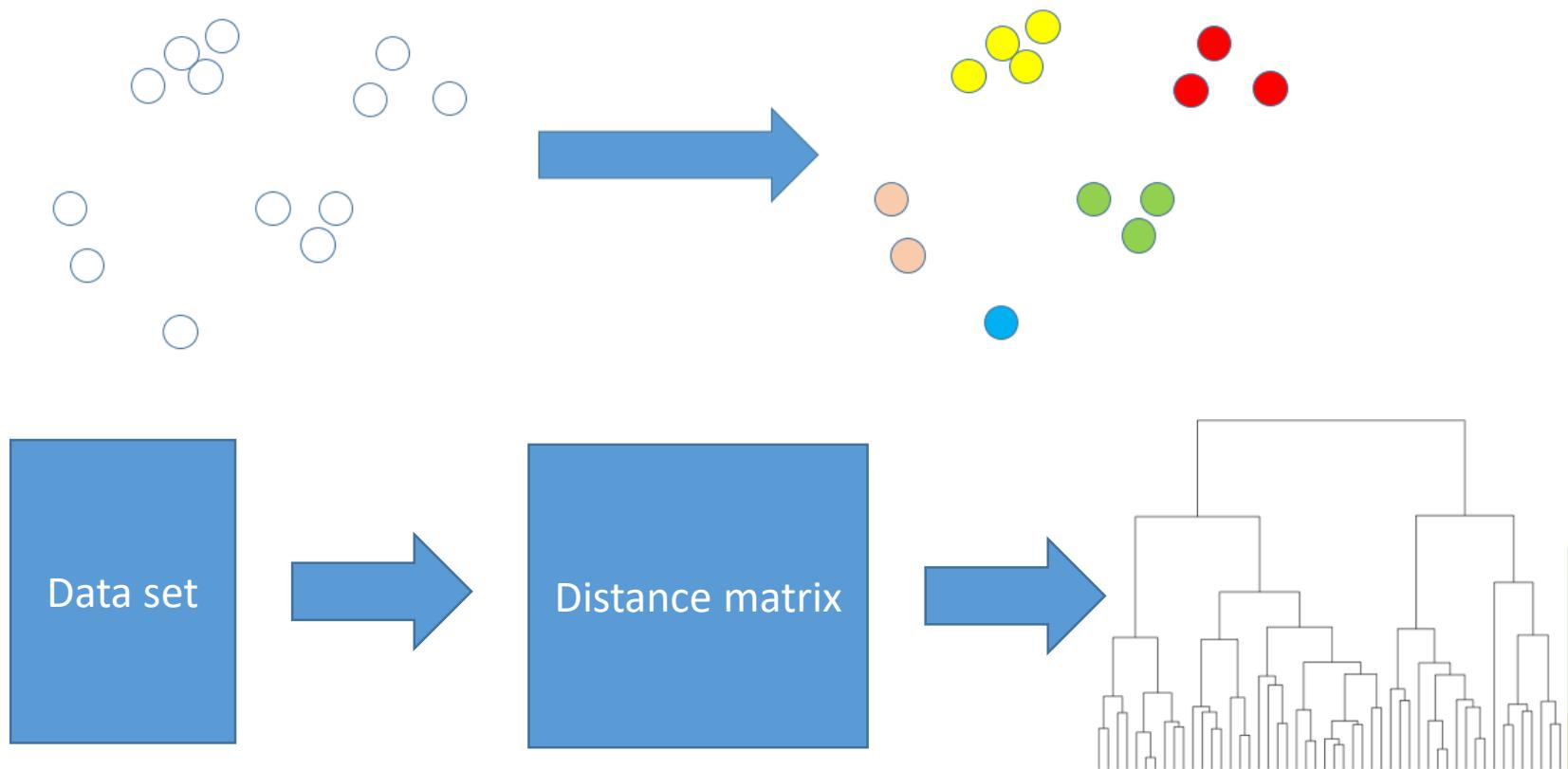
$$sim(expr_1, expr_2) = \arg \min_{c_i, c_j} \sum sim(c_i, c_j)$$

optimal alignment problem – Hungarian algorithm

Cluster analysis of annotated documents

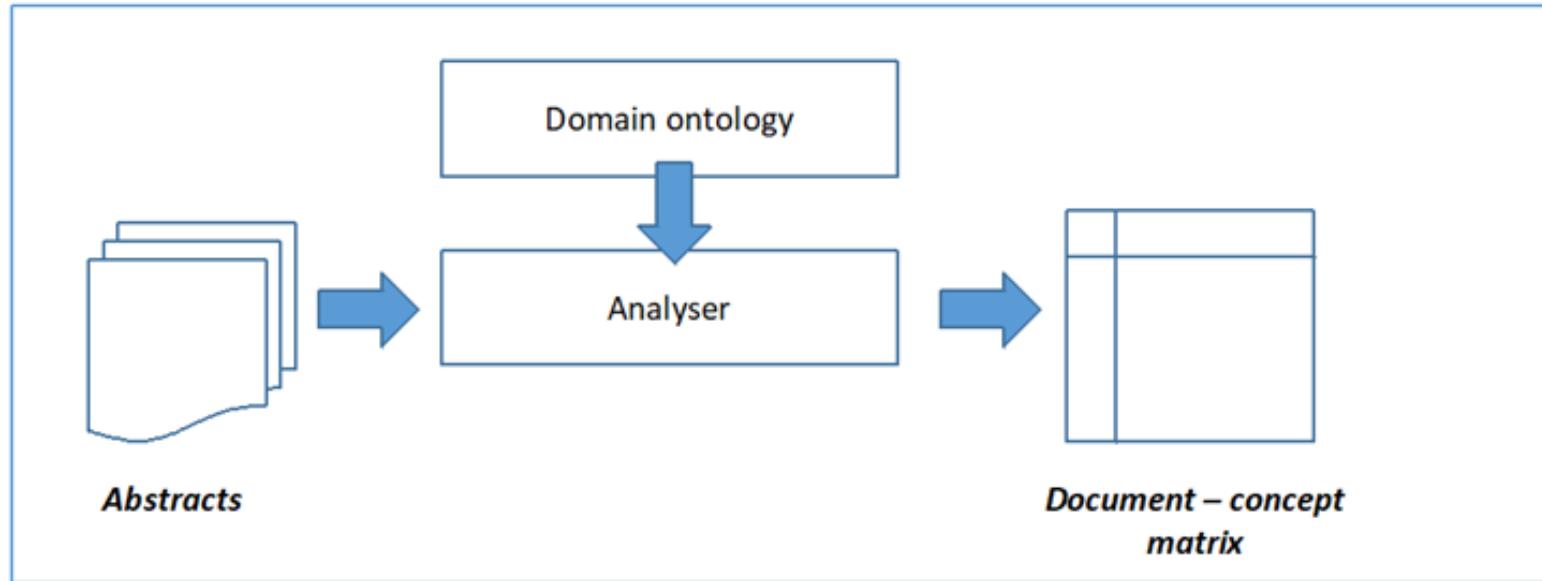
Cluster analysis of documents

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



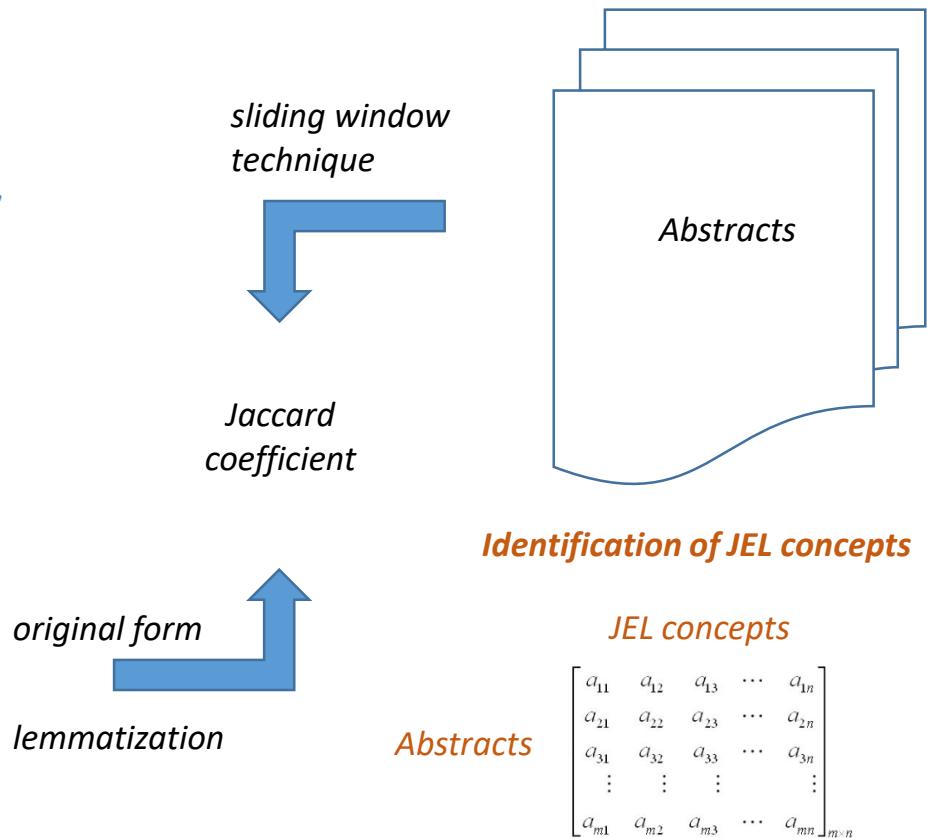
Practical aspects of cluster analysis of annotated documents (in the context of research productivity analysis)

The structure of the ontology-based analyzer

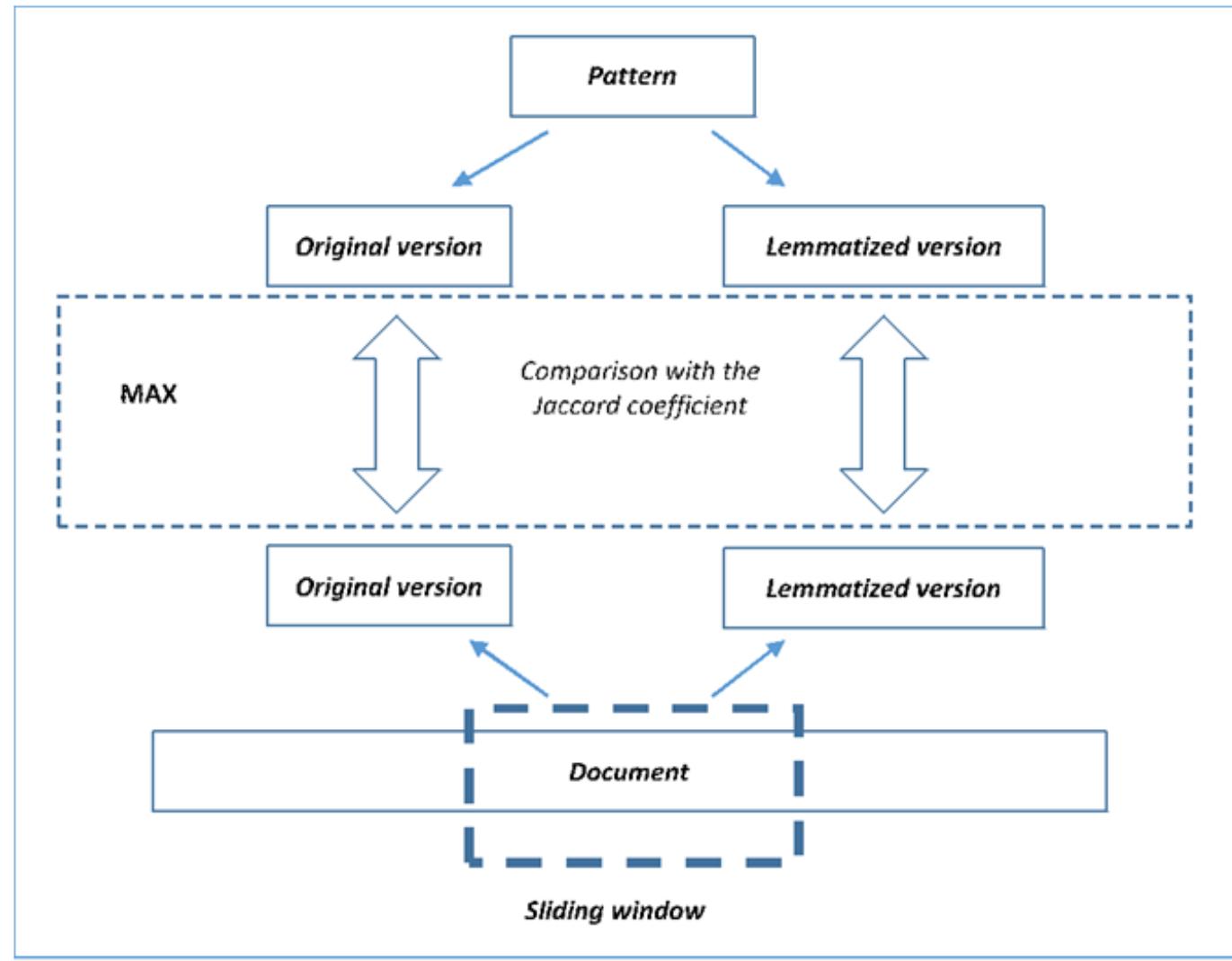


The structure of the ontology-based analyzer

```
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"
```

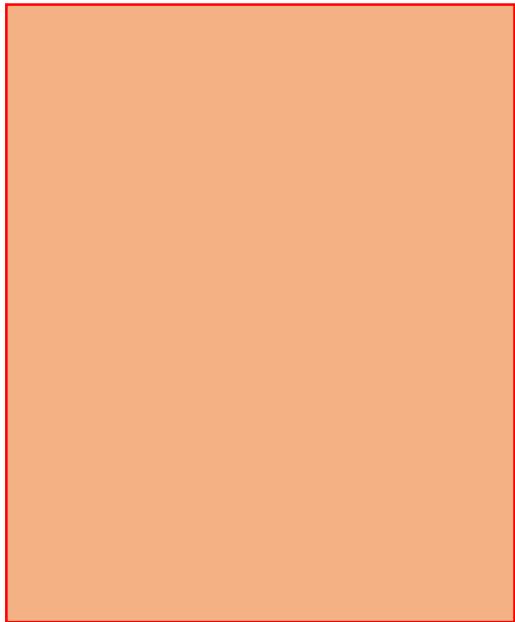


Linguistic issues in pattern analysis

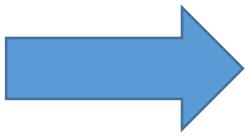


Similarity of documents

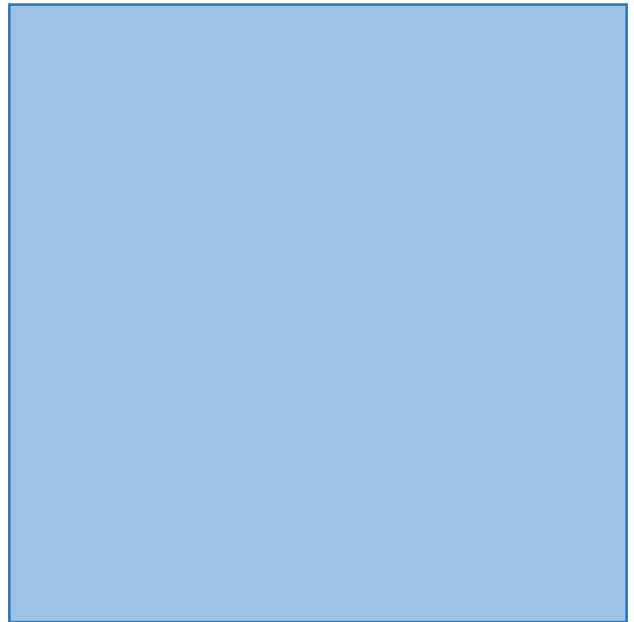
Abstracts (124460)



JEL Concepts

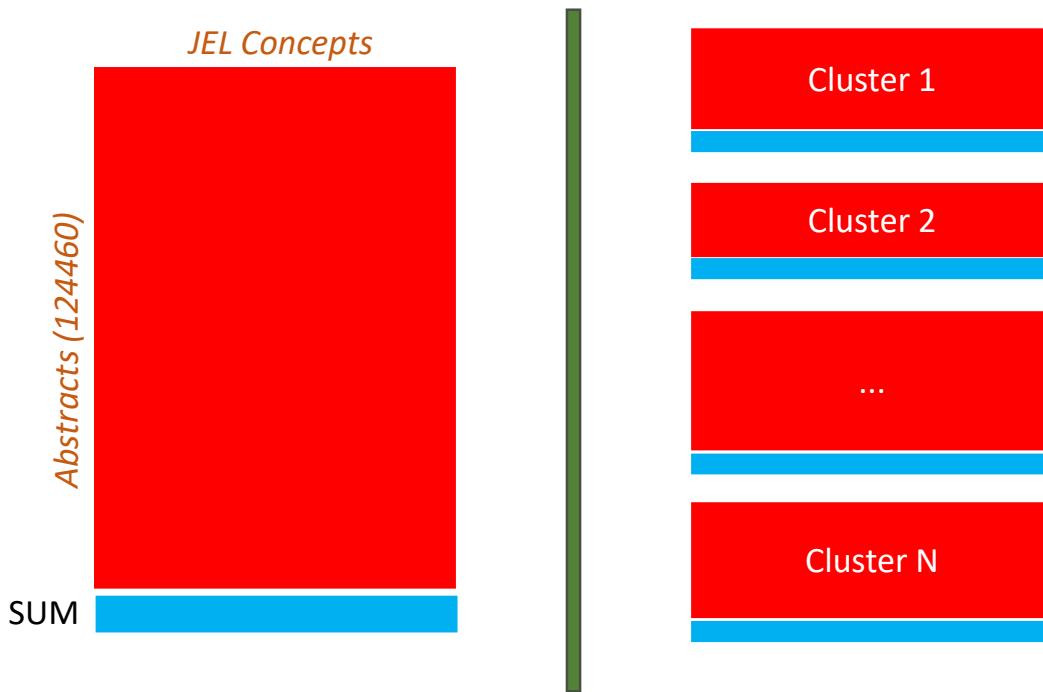


Abstracts (124460)



Abstracts (124460)

Cluster analysis of abstracts



Projects – dataset

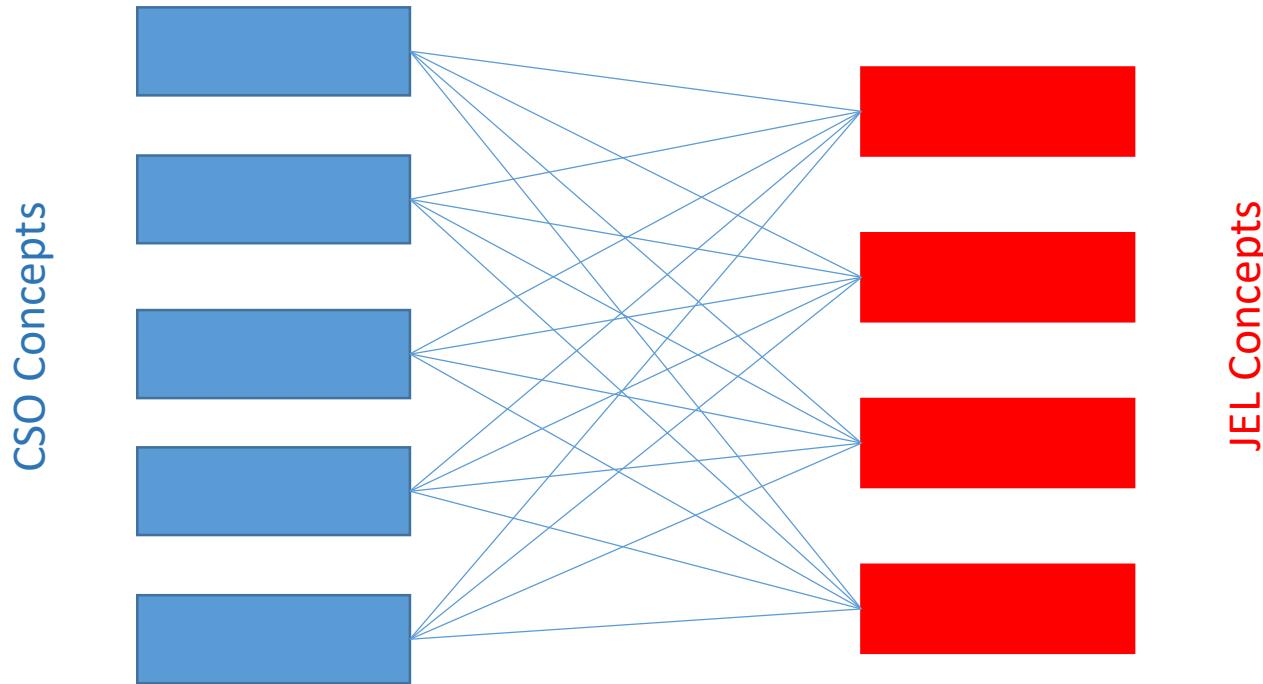
- Cordis database was used.
- Projects related to the area of economics (started in 2019 – 2021).
- Horizon and Framework Program projects were included.

Cluster analysis of projects' description



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

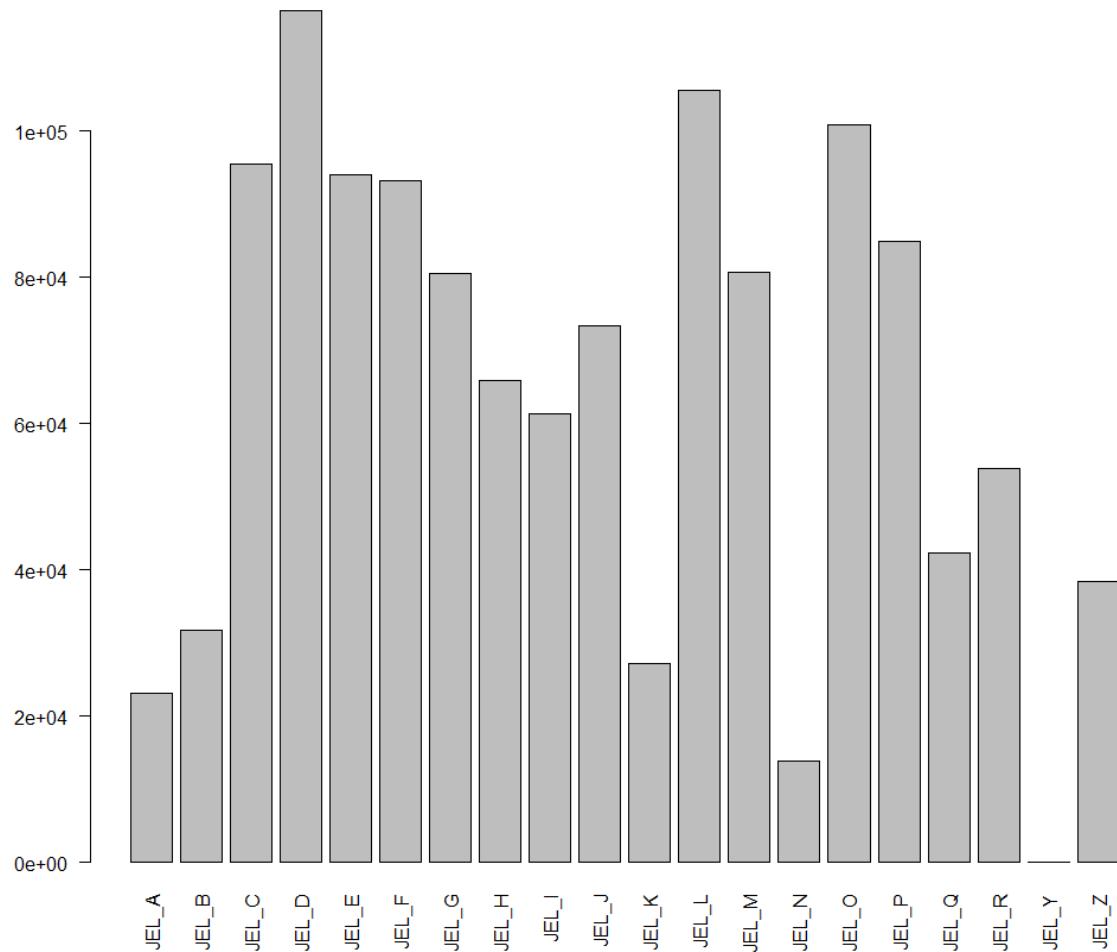
Bipartite model of interdisciplinarity



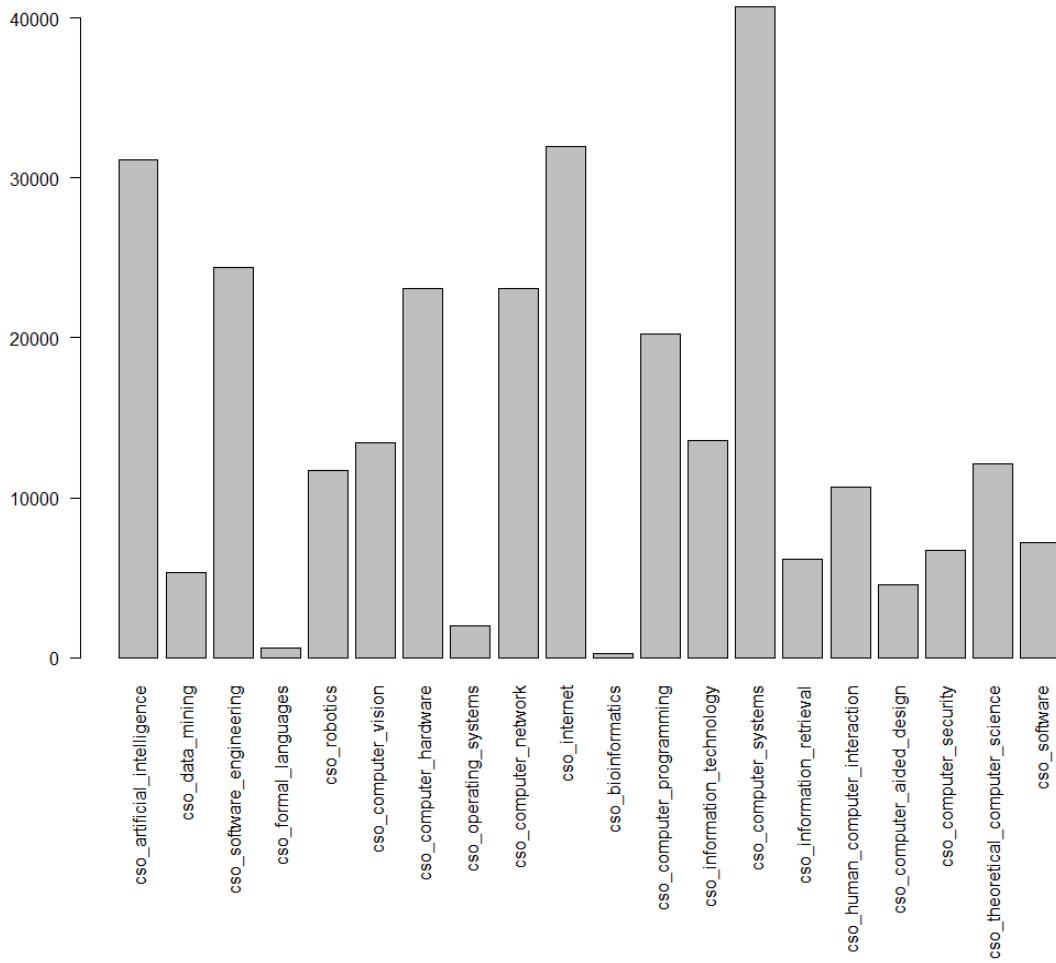
Areas of analysis:

- subarea strength
- subarea specificity
- graph model modularity

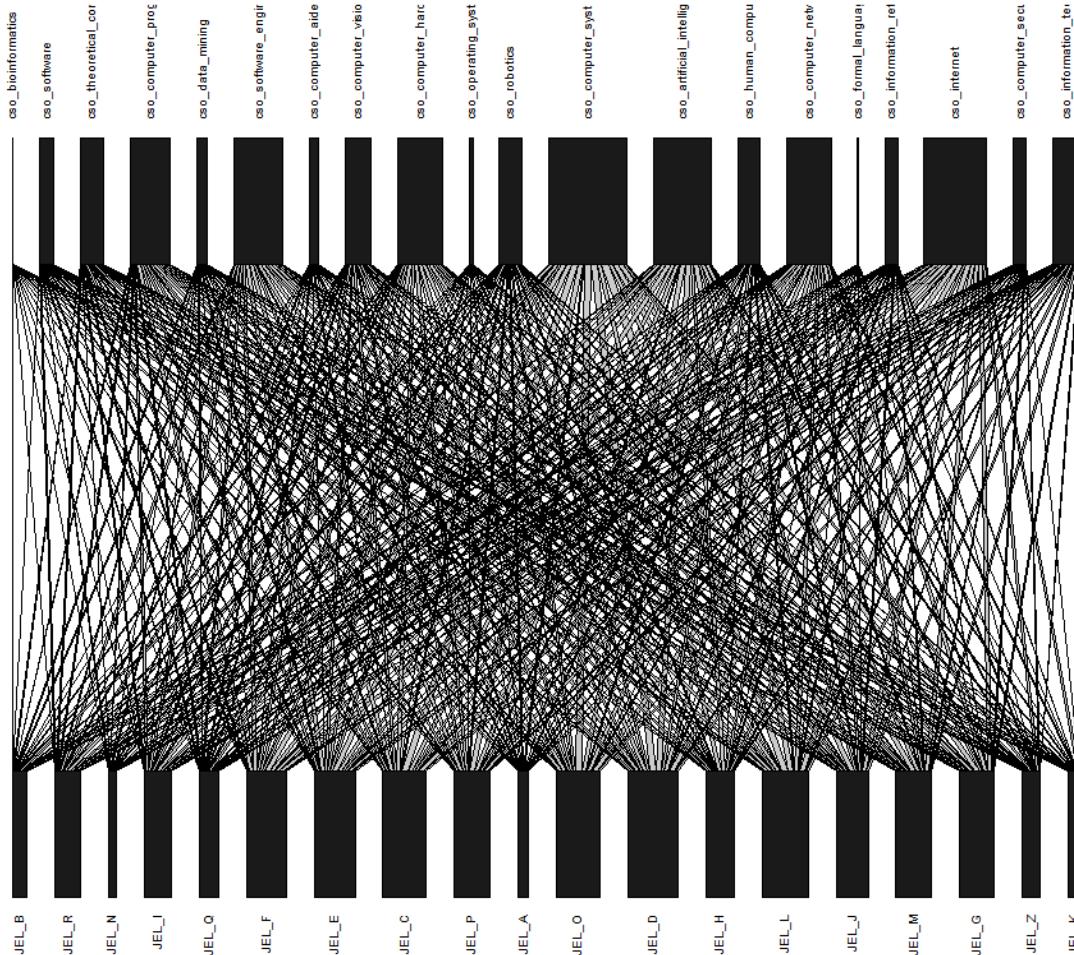
The distribution of documents over the main concepts from the JEL ontology



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



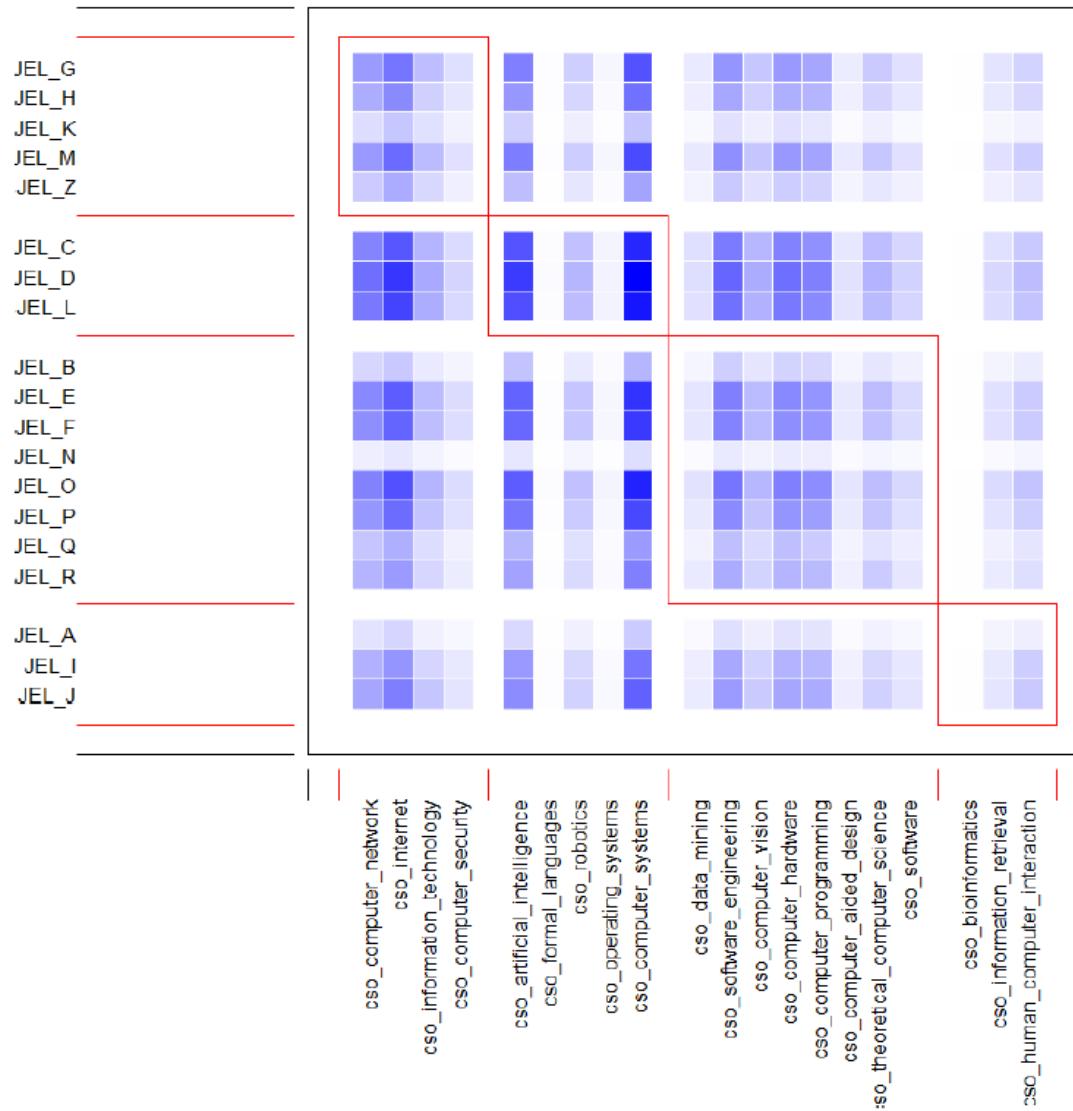
Relationships between main subareas from the JEL and CSO ontology



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 screenshot shows the homepage of the National Library of Medicine's Medical Subject Headings (MeSH) website. At the top, there is a blue header bar with the NIH logo and the text "National Library of Medicine". To the right of the logo is a search bar with the placeholder "Search NLM" and a magnifying glass icon. Below the header is a white navigation bar with a three-line menu icon. The main content area has a dark grey background. On the left, there is a logo for "Medical Subject Headings" featuring a stylized tree or leaf design above the acronym "MeSH". To the right of the logo, the text "Medical Subject Headings" is written in a large, white, sans-serif font. Below this, a horizontal menu bar contains links: "MeSH Home", "Learn About MeSH", "MeSH Browser", "Download MeSH Data", "MeSH on Demand", and "Suggestions".

[Home](#)

Welcome to Medical Subject Headings

The Medical Subject Headings (MeSH) thesaurus is a controlled and hierarchically-organized vocabulary produced by the National Library of Medicine. It is used for indexing, cataloging, and searching of biomedical and health-related information. MeSH includes the subject headings appearing in MEDLINE/PubMed, the NLM Catalog, and other NLM databases.

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

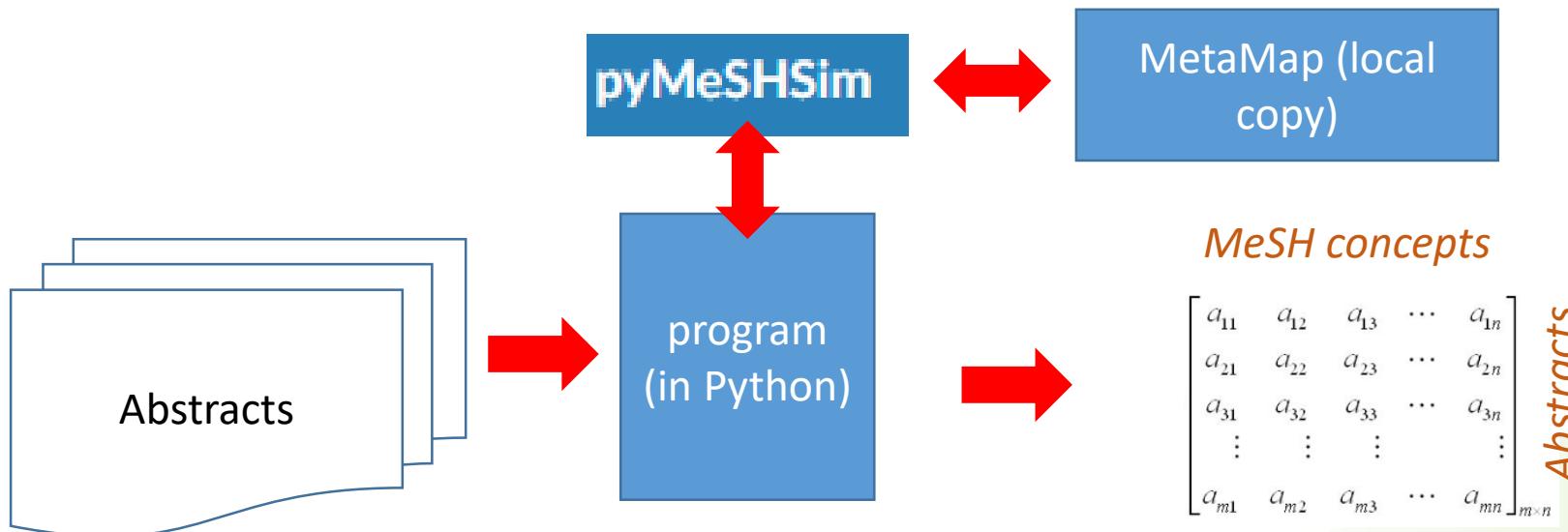


National Library of Medicine



Medical Subject Headings

MetaMap - A Tool For Recognizing UMLS Concepts in Text

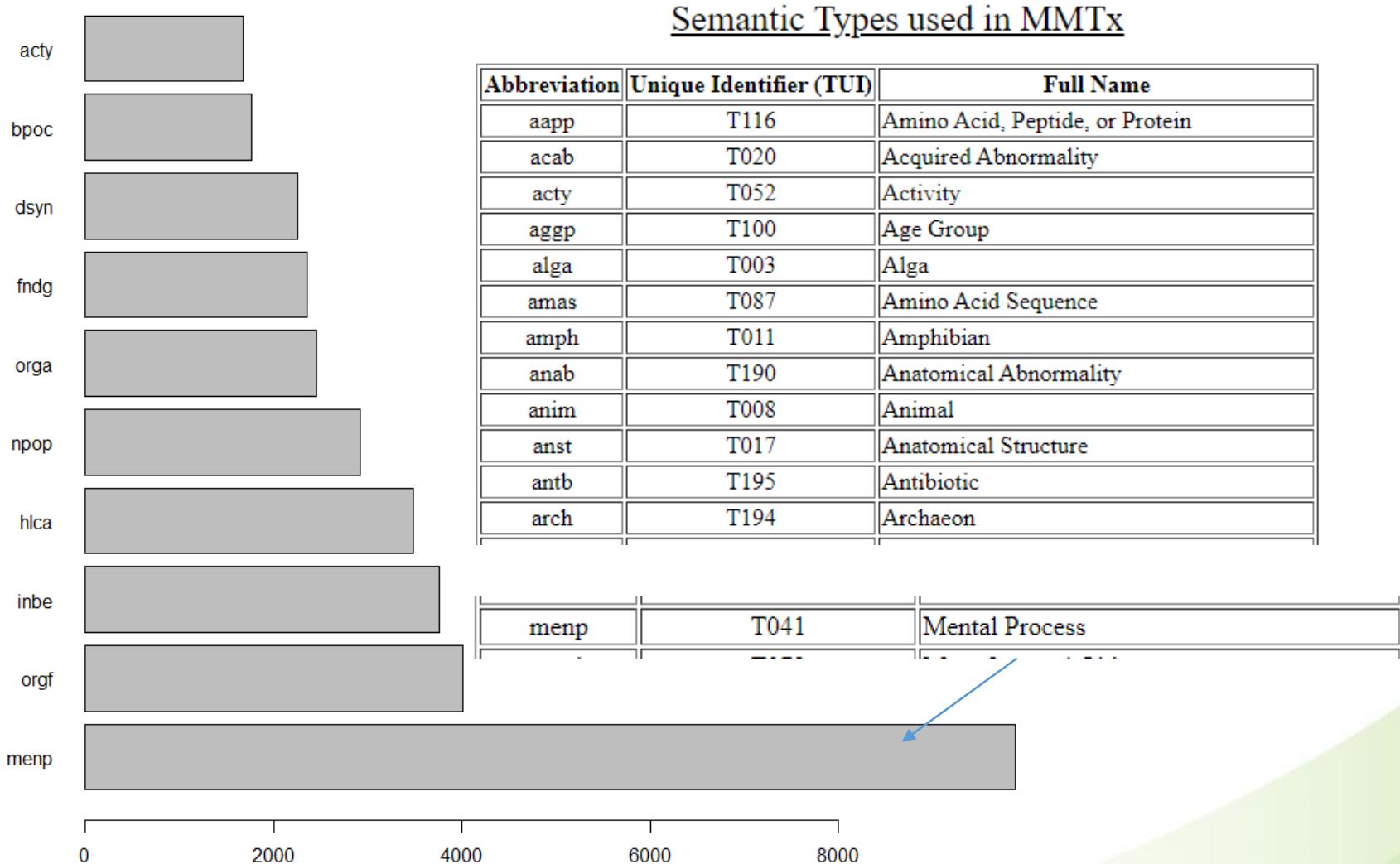


Levels of analysis

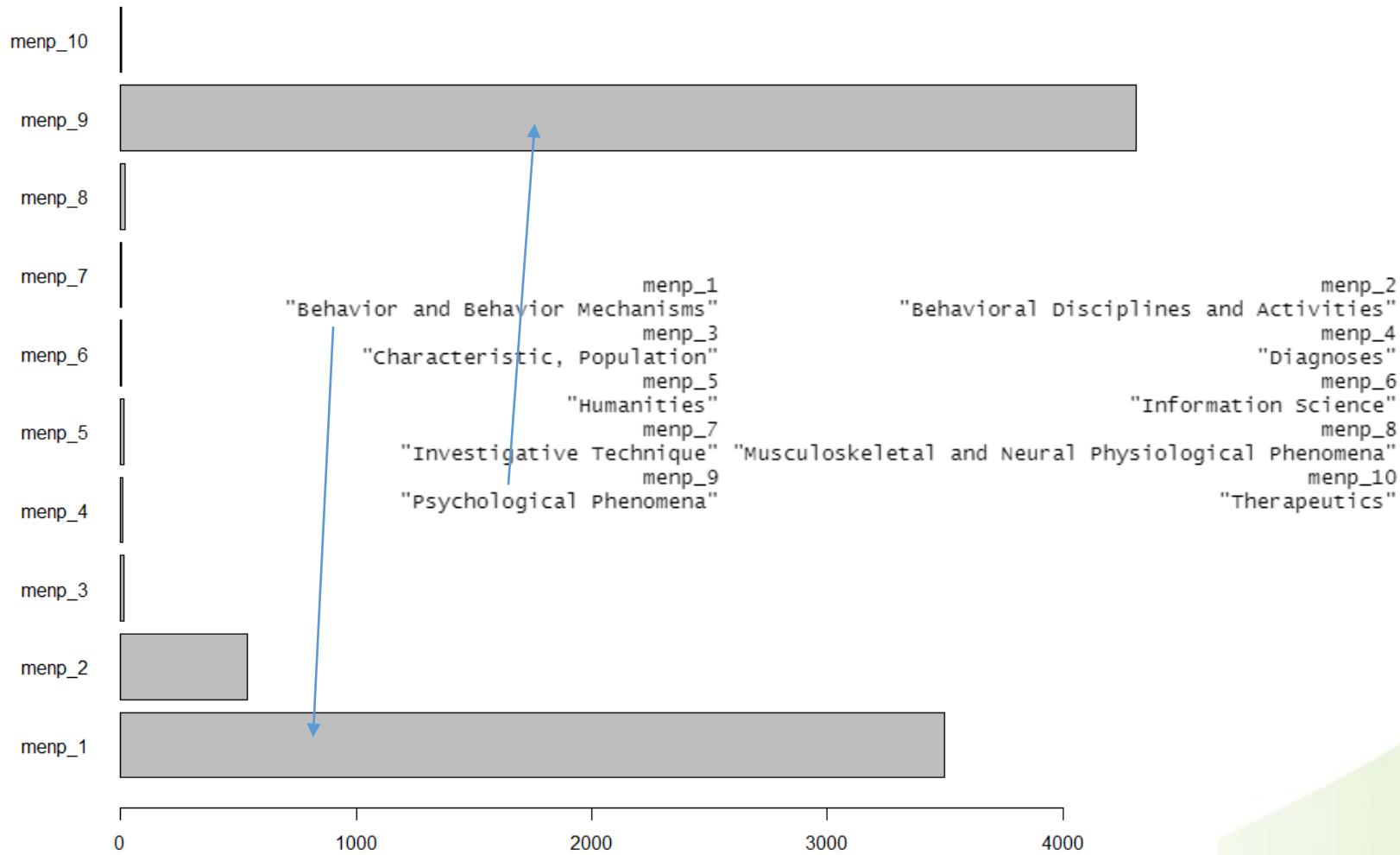
Projection to main concepts for the MeSH ontology

- Behavior and Behavior Mechanisms [F01]
- Adaptation, Psychological [F01.058]
- Attitude [F01.100]
- Behavior [F01.145]
- Child Rearing [F01.318]
- Defense Mechanisms [F01.393]
- Emotions [F01.470]
- Human Characteristics [F01.510]
- Human Development [F01.525]
- Mental Competency [F01.590]
- Motivation [F01.658]
 - Achievement [F01.658.059]
 - Aspirations, Psychological [F01.658.100]
 - Conflict, Psychological [F01.658.209]
 - Drive [F01.658.293]
 - Exploratory Behavior [F01.658.370]
 - Food Deprivation [F01.658.433]
 - Goals [F01.658.500]
 - Handling, Psychological [F01.658.556]
 - Instinct [F01.658.642]
 - Intention [F01.658.650]
 - Power, Psychological [F01.658.780]
 - Water Deprivation [F01.658.938]
- Neurobehavioral Manifestations [F01.700]
- Personality [F01.752]
- Psychology, Social [F01.829]
- Temperance [F01.914]

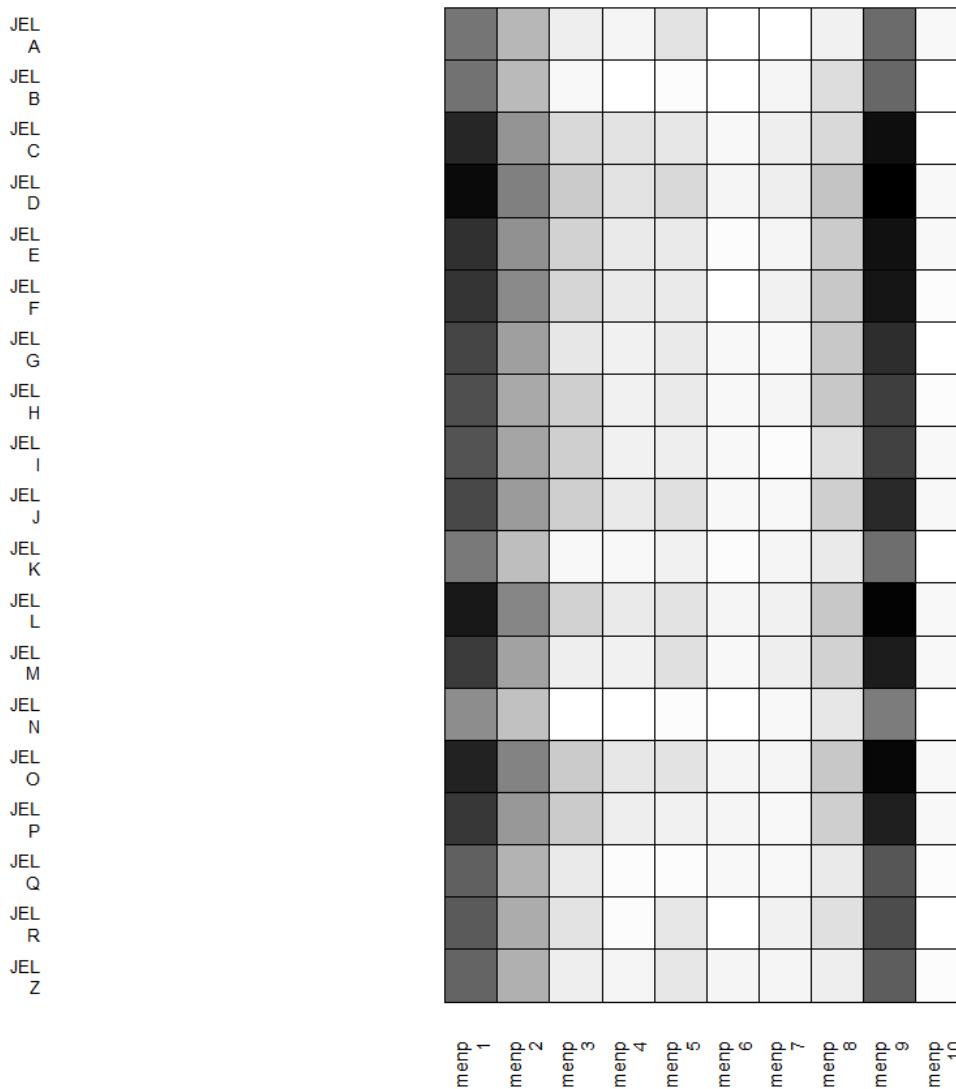
Analysis of concepts related to mental processes



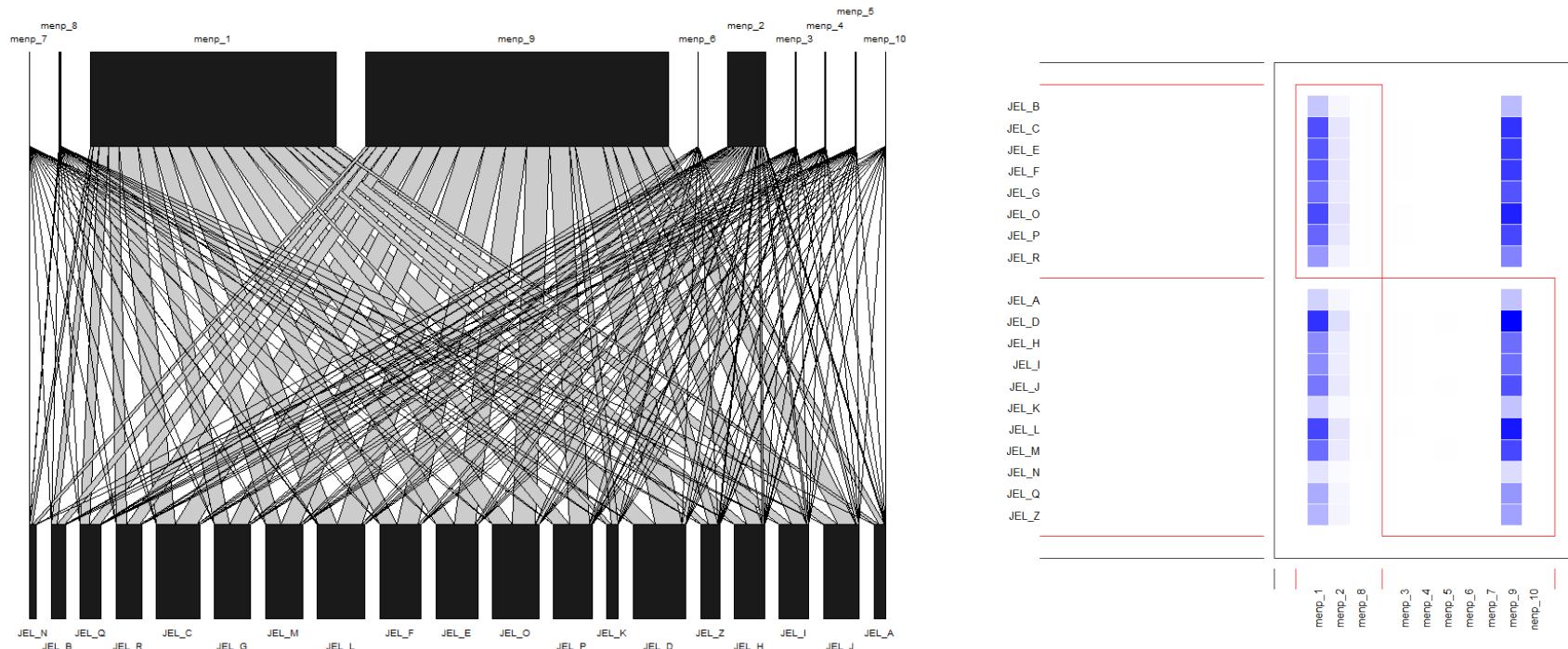
The significance of concepts related to mental processes



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