ZRK Zintegrowany Rejestr Kwalifikacji

Clustering and Visualisation of Qualifications – Preliminary Results

Warsaw, December 13, 2019

Marcin Będkowski, Leopold Będkowski

European Funds
Knowledge Education Development

Republic of Poland

ZRK Zintegrowany Rejestr Kwalifikacji

European Union
European Social Fund

Integrated Qualificaton Register

- ✓over 10 000 qualifications (10 004)
- ✓ca. 700 contain full descriptions of LO’s
- ✓ca. 500 contain short descriptions of LO’s

By the end of the year:

- ✓several dozens of new market qualifications (over 200 in queue for next year)
- ✓ca. 5000 descriptions for HE qualifications
- ✓ca. 215 descriptions for VET qualifications (new curriculum)

Our Context: Qualifications Register modernization

- ✓ Improving searching and browsing usability
  (semantic search, filtering options, categorization and/or tagging of content, context browsing tools)
- ✓ Developing automatic reporting and additional queries
  (qualifications comparison, generating lists of qualifications based on selected criteria, e.g. containing phrases, simliar to)
- ✓ Designing web applications:
  - ✓ „Compass”;
  - ✓ „Learning pathways”;
  - ✓ „Virtual assistant”.

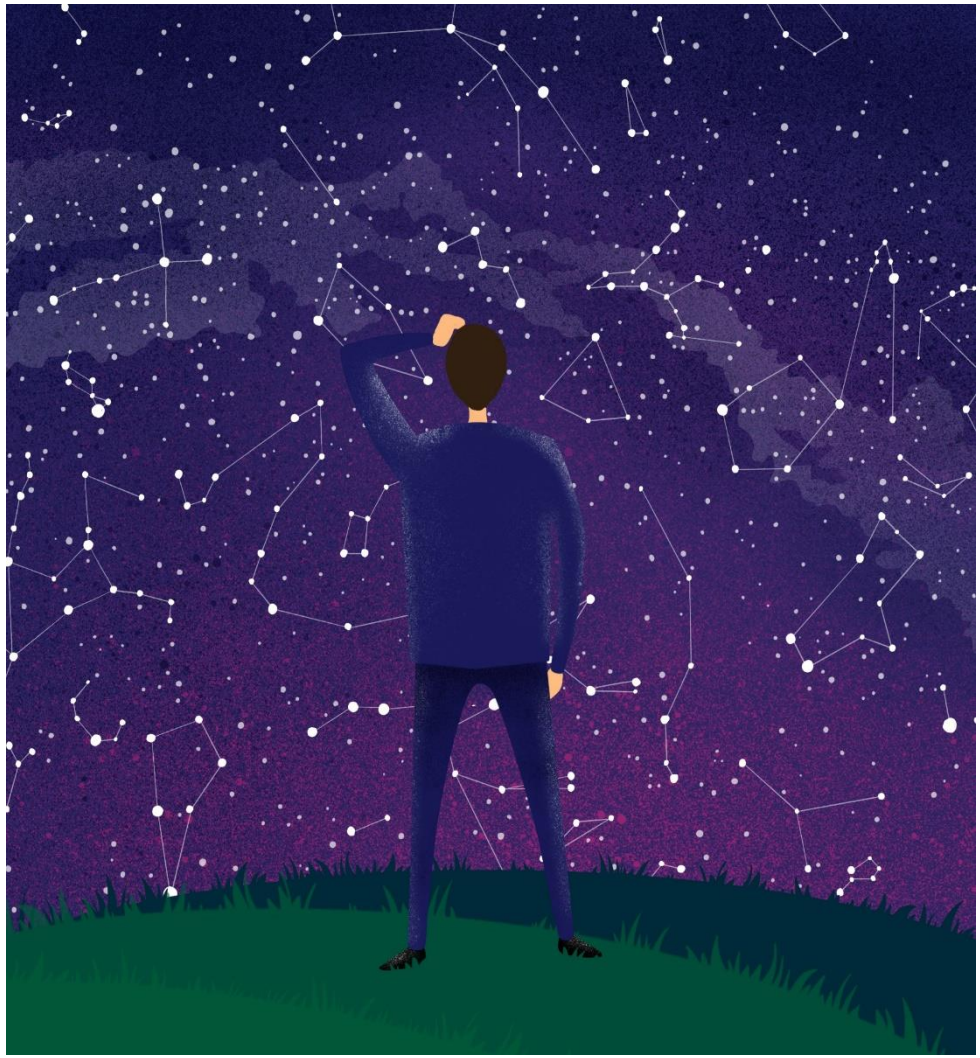
Planned web applications

- ’Compass’ – enables identification of qualification or qualifications groups in the IQR based on sucessive approximations using predefined criteria and categories
- ’Learning pathways’ – visual presentation of relations between qualifications (preceeds, similar-to, is-part-of etc.) based on information extraction
- ’Virtual assistant’ (feasibility study for now) – conversational user interface offering access to the IQR and a range of services: information search, FAQ and possibly services such as CV-qualification matching, diagnosing / counseling tests

The „WHY?”: Similar trends in international context

Policy perspectives:

- ✓Accessibility and transparency of qualifications system
- ✓Credit accumulation and transfer and builiding learning pathways (awarding bodies and learners)
- ✓Preventing proliferation of similar qualifications
- ✓…



The „WHAT?”: long list

- ✓**Assessing similarity of objects**
- ✓Determining and representing relations between qualifications
- ✓**Grouping / clustering of qualifications**
- ✓Classifying and linking to existing taxonomies/classifications
- ✓Supporting decision process and qualification design/description
- ✓Supporting levelling proces

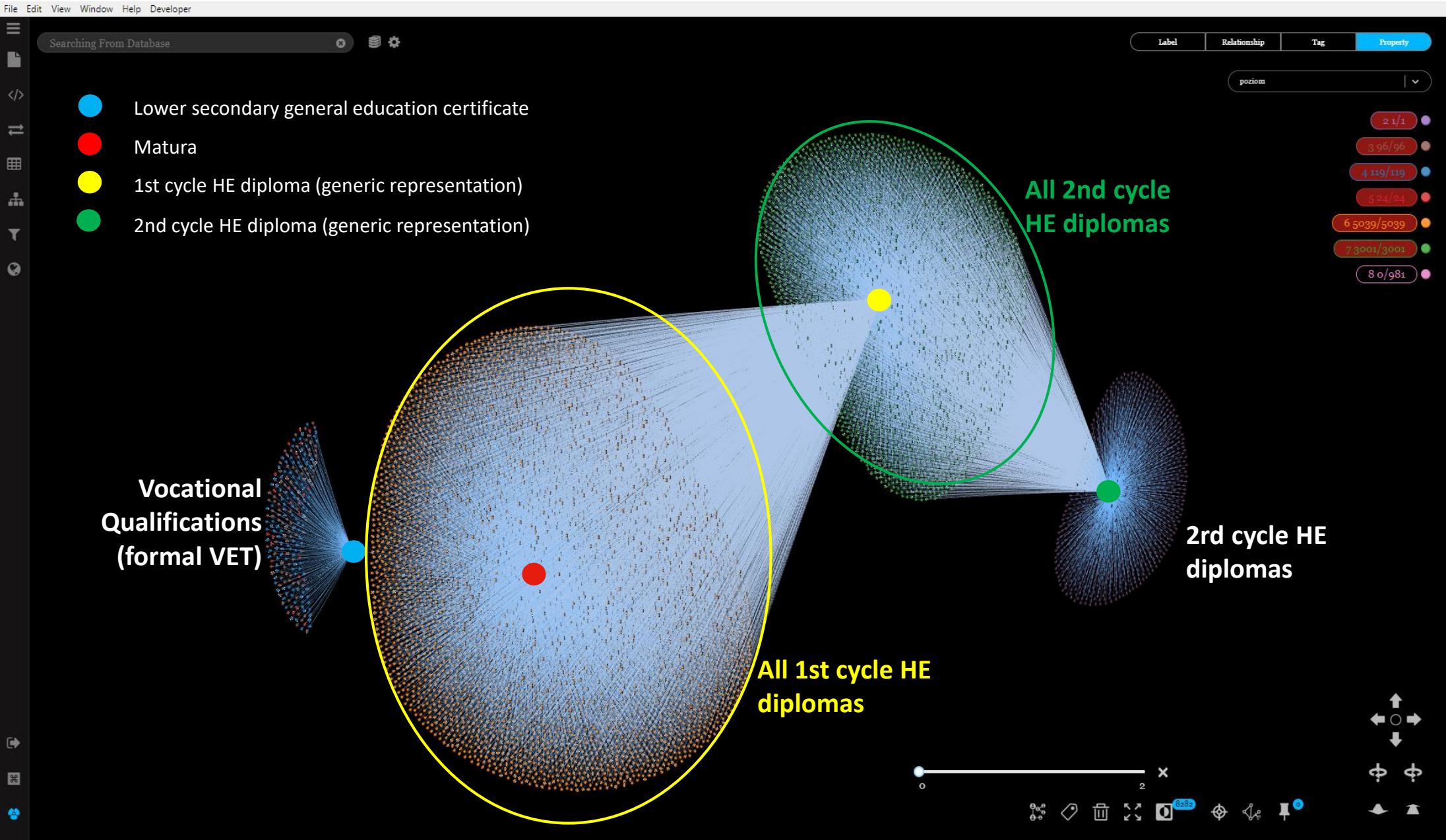
The „HOW?”: Similar challenges?

How many people do we need to compare, group, tag **10 000** qualifications of different structure and content?

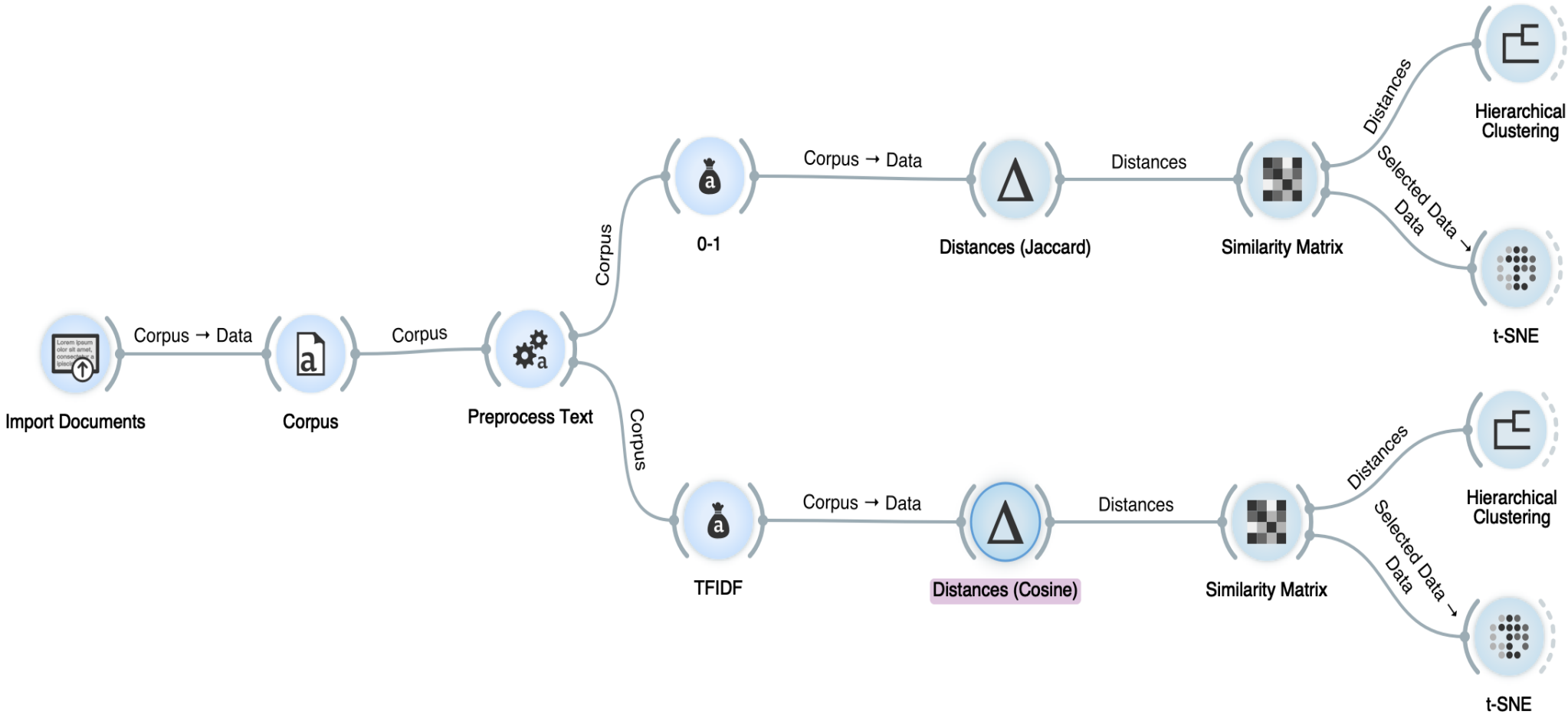
$$(X + Y) \times N$$

- ✓X analytics
- ✓Y qualifications experts
- ✓for N months

Representation of qualifications in the register that have formal prerequistes



Exemplary pipelines in Orange



Two baseline approaches

	Approach no. 1	Approach no. 2
Basis for comparison	Learning outcomes	Synthetic description
Features	lemmatized nouns	lemmatized n-grams
No. of features	ca. 3300	ca. 4000
Feature weighting	0–1	TFIDF
Measure of similarity	jaccard	cosine

Natural Language Processing – basic terms

- ✓ lemmatization (determining base forms of words)
- ✓ jaccard index (of similarity)
- ✓ TFIDF
- ✓ n-grams
- ✓ cosine similarity

Example of data preprocessing: 'Atomization' of learning outcomes (LO) – difficult task for Polish

Using NLP tools, we atomized the LO's and extracted and lemmatized relevant words:

- (The learner) **describes** and **explains** the construction of **hammers** and **nails**

➔

- **Describes** the construction of the **hammer** +
  Explains the construction of the **hammer** +
  Describes the construction of the **nails** +
  Explains the construction of the **nails**

➔

- (**describe**, **hammer**, **explain**, **nail**, construction)

Jaccard index

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

TFIDF

$$tfidf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \cdot \log \frac{|D|}{|d : t_i \in d|}$$

Cosine similarity

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

New baseline approach

	Approach no. 3
Basis for comparison	Name + synthetic description + learning outcomes + [elements from partial vocational qualifications]
Features	lemmatized n-grams
No. of features	ca. 20 000 – 44 000 [sic!]
Feature weighting	TFIDF
Measure of similarity	cosine

… but let us focus on the results

Case 1. Dental assistant qualification [VET qualification]: automated extraction of keywords

Key phrases (with TFIDF values):

0.191 lekarz dentysta [dentist]

0.191 dentysta [dentist]

0.183 stomatologiczny [dental]

0.167 dentystyczny [dental]

0.158 gabinet dentystyczny [dental surgery – place]

0.155 gabinet [surgery – place]

Case 1. Dental assistant qualification: most similar qualifications based on calculation of cosine similarity

The most similar qualifications (with cosine similarity):

0.8077 Dental hygienist (325102)

0.7298 Assisting the dentist and keeping the surgery ready for use

0.6719 Paramedic (325601)

0.5551 Massage technician (325402)

0.5492 Orthoptician (325906)

0.5481 Dental technician (321402)

Case 2. Design of websites qualification [market qualification]: automated extraction of keywords

Key phrases (with TFIDF values):

0.369 witryna [website]

0.217 witryna internetowy [website]

0.196 serwer [server]

0.180 zarządzanie treść [content management]

0.180 system zarządzanie treść [content management system]

0.180 przeprowadzać test [perform test]

Case 2. Design of websites qualification: most similar qualifications based on calculation of cosine similarity

The most similar qualifications (with cosine similarity):

0.2997 Programming, development and administration of websites and databases since September1, 2017

0.2456 Web application and database development and administration

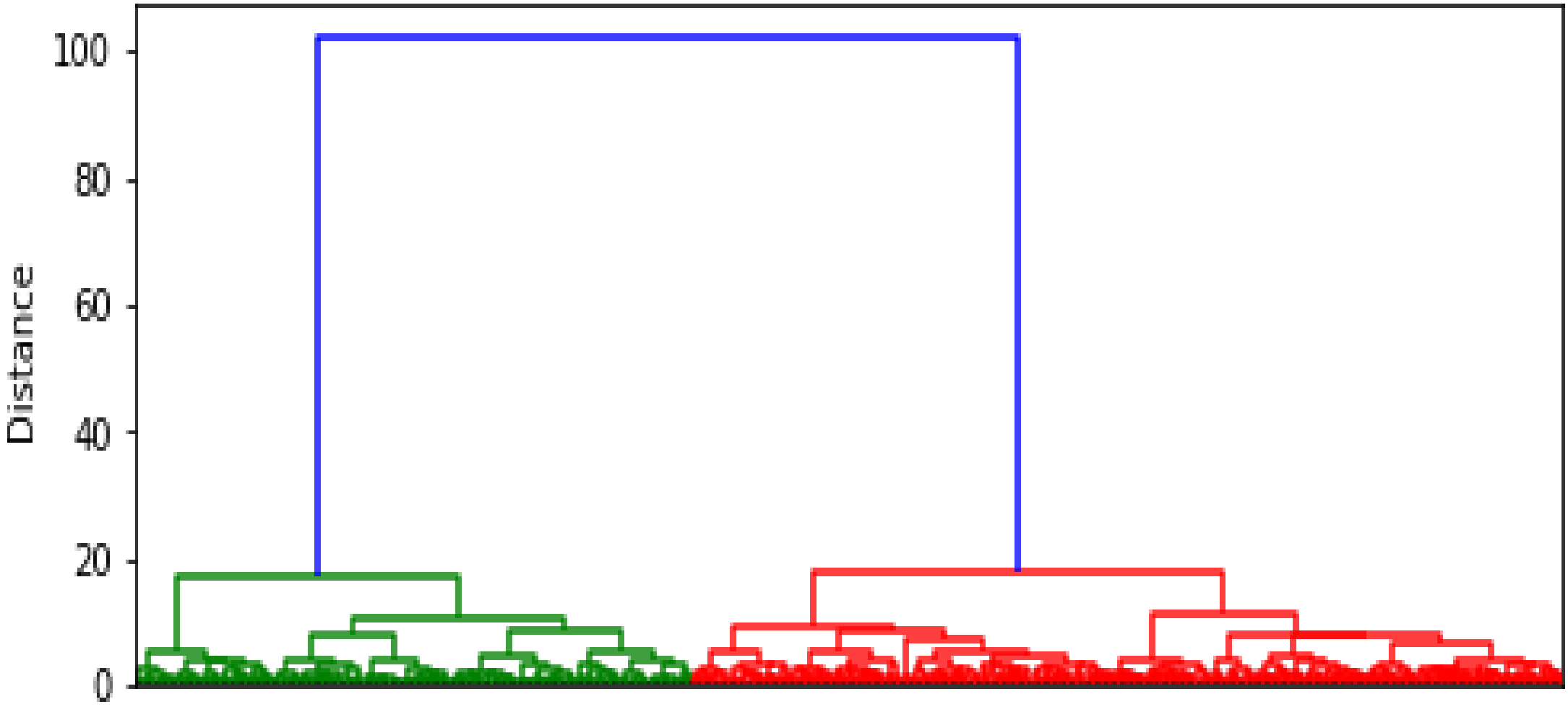
0.2229 Computer graphics design

0.2072 IT technician (351203) since September 1, 2017

0.1875 Multimedia projects since September 1, 2017

0.1712 IT technician (351203)

Case 3. Hierarchical clustering dendrogram



Case 3. Hierarchical clustering dendrogram – example of 4th level cluster

- ✓Printer
- ✓Printing process technician
- ✓Bookbinding process technician (311936)
- ✓Bookbinder (732301)
- ✓Digital computer graphics technician
- ✓Graphics and digital printing technician September 1, 2017
- ✓Photographer
- ✓Photo-technical engineering
- ✓Photographer since September 1, 2017

Case 3. Hierarchical clustering dendrogram – example of 6th level cluster

- ✓Printer
- ✓Printing process technician
- ✓Bookbinding process technician
- ✓Bookbinder
- ✓Digital computer graphics technician
- ✓Graphics and digital printing technician, since Sep
- ✓Photo-technical engineering
- ✓Photographer
- ✓Photographer, since September 1, 2017

Printing

Bookbinding

Graphics and digital printing

Photography

Case 4. Hierarchical clustering dendrogram – example of 6th level cluster with human labelling

- ✓ Beekeeping
- ✓ Conducting agricultural production
- ✓ Running an agritourism farm
- ✓ Organisation and supervision of agricultural and beekeepi
- ✓ Organisation and supervision of agricultural production
- ✓ Animal husbandry, breeding and insemination
- ✓ Organisation of horse breeding
- ✓ Animal husbandry and insemination (since 1 September 2
- ✓ Gardener's technician
- ✓ Gardener
- ✓ Beekeeper
- ✓ Beekeeper technician
- ✓ Agricultural technician
- ✓ Agri-business technician
- ✓ Farmer

Agricultural production and beekeeping

Animal husbandry, breeding and insemination

Gardening

Beekeeping

Farming

… why baseline?

Case 5. Problematic grouping of market qualifications. Lowest level of hierarchy

Earlier result:

- ✓Working with a child using Maria Montessori's method
- ✓Installing and maintaining connections and indoor installations in fiber optic technology

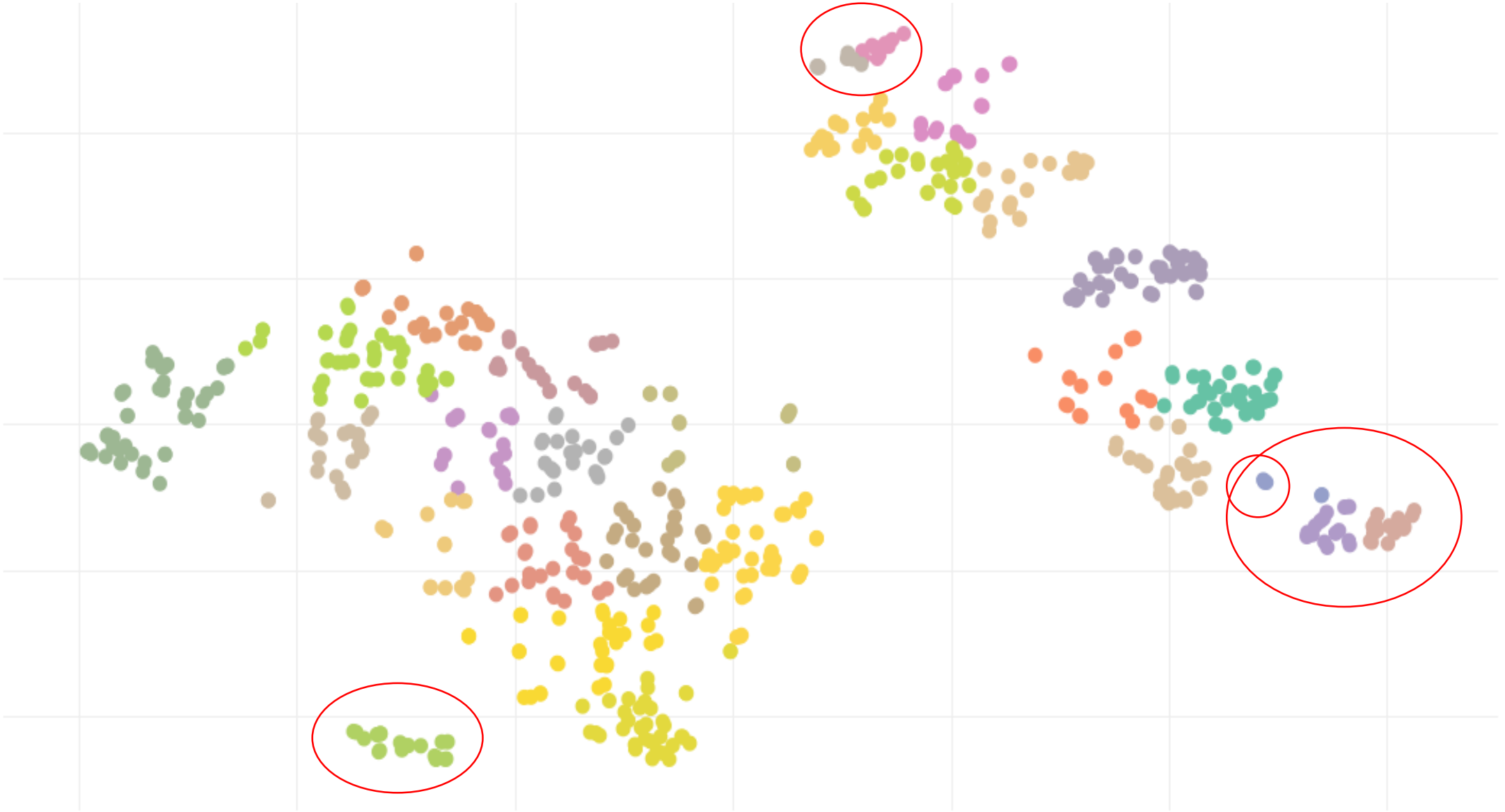
Current result:

- ✓Working with a child using Maria Montessori's method
- ✓Conducting therapy for children and adolescents

Case 5. Problematic grouping of market qualifications

- ✓Recovering data from HDDs
- ✓Working with a child using Maria Montessori's method
- ✓Conducting training by means of activation methods
- ✓Programming and operation of the 3D printing process
- ✓Installing and maintaining connections and indoor installations in fiber optic technology
- ✓Conducting therapy for children and adolescents
- ✓Repair, maintenance and modernisation of bicycles
- ✓Installation and maintenance of autonomous detectors: carbon monoxide, smoke, heat and gas

Case 6. T-SNE visualisation in 2D space with K-Means clustering (colours) (demo)



https://lbedk.shinyapps.io/t-sne/

Another problematic examples

They do not form consistent groups:

- ✓Florist
- ✓Medical electronics and informatics technicians
- ✓Cosmetician

Observations

- ✓ Regarding the data:
  - ✓ Merging qualifications improves results (keywords saturation)
  - ✓ Example of market qualifications:
    - ✓ elimination of common learning outcomes from vocational qualifications?
    - ✓ exclusion of partial qualifications from vocational education?
- ✓ Regarding features:
  - ✓ 20.000 features is (still) too much
    - ✓ but the results are relatively good (and promising)
    - ✓ a good baseline approach that can be improved in the next steps
- ✓ Regarding hierarchical grouping:
  - ✓ we need to examine the stability of clusters
  - ✓ Ward is not the optimal method?

We see our results as a proof of concept for:

- ✓ automated comparison of qualifications and explicable/interpretable degree of similarity
- ✓ automated extraction of key phrases for qualifications
- ✓ grouping / clustering of qualifications – independent from existing classifications

Work in progress

- ✓ grouping methods pilot study and clustering application – testing other approaches
  - ✓ knowledge-based measures using WordNet
  - ✓ vector language models (word2vec, fasttext, ELMo, USE…)
  - ✓ ARTM (topic modeling)
  - ✓ model ensembling
- ✓ collecting data concerning occupations, job offers, etc. for the purpose of model training and data augmentation
- ✓ consultations with experts, evaluation of results
- ✓ feasibility study on chatbot
- ✓ three applications supporting register users

ZRK Zintegrowany Rejestr Kwalifikacji

Thank You!

m.bedkowski@ibe.edu.pl

Educational Research Institute
IQS Project Office
Górczewska 8, 01-180 Warsaw, Poland
phone: +48 22 24 17 100, +48 22 24 17 111
e-mail: rejestr@ibe.edu.pl

http://rejestr.kwalifikacje.gov.pl | http://www.ibe.edu.pl

European Funds Knowledge Education Development
Republic of Poland
ZRK Zintegrowany Rejestr Kwalifikacji
European Union European Social Fund