Structure mining from Clinical Free-Text Records

Adam Gabriel Dobrakowski

University of Warsaw

ad359226@students.mimuw.edu.pl

13 V 2019

Overview

Problem statement

- Data from patients' visits in health centers containing ICD-10 code, age, sex, history of diseases, doctor speciality, place, interview description, examination description, recommendation description...

Problem statement

- Data from patients' visits in health centers containing ICD-10 code, age, sex, history of diseases, doctor speciality, place, interview description, examination description, recommendation description...
- Main goal: find some similarities between visits and group them into several clusters

Problem statement

- Data from patients' visits in health centers containing ICD-10 code, age, sex, history of diseases, doctor speciality, place, interview description, examination description, recommendation description...
- Main goal: find some similarities between visits and group them into several clusters
  We would be able to:
- follow recommendations that were applied to patients with similar visits in the past to create a list of possible diagnoses
- reveal that current diagnosis is unusual
- identify subsets of visits with same diagnosis but different symptoms

ICD-10 codes

- ICD-10 – International Statistical Classification of Diseases and Related Health Problems
- code: letter + two digits, e.g. E11, F32
- the letter is a type of disease, e.g. J – diseases of the respiratory system
- Z – factors influencing health status and contact with health services

First approach

For each visit we have the history of diseases of the patient, e.g. {J02, J00, M23, S43, Z27, Z00, Z00, B86, Z00, Z71}.
We want to cluster visits based only on this history.
Tensor decomposition algorithm:

- Encode each visit to one-hot encoded vector of diseases: $X = (x_1, ..., x_d)$, d – total number of diseases
- Assume: X are the outcomes from mixed binary distributions, $Y \in \{1, ..., k\}$ – a hidden variable, $\mathbb{P}(Y = j) = \omega_j$, $\mathbb{E}(x_i = 1 | Y = j) = \mu_{ij}$.
  To make clustering compute:
  $$\mathbb{P}(Y = j|X) \propto \omega_j \prod_{i=1}^{d} \mu_{ij}^{x_i}(1 - \mu_{ij})^{1-x_i}.$$
- Estimate μ_{ij} and ω_j based on first three moments and tensor decomposition.

Tensor decomposition algorithm

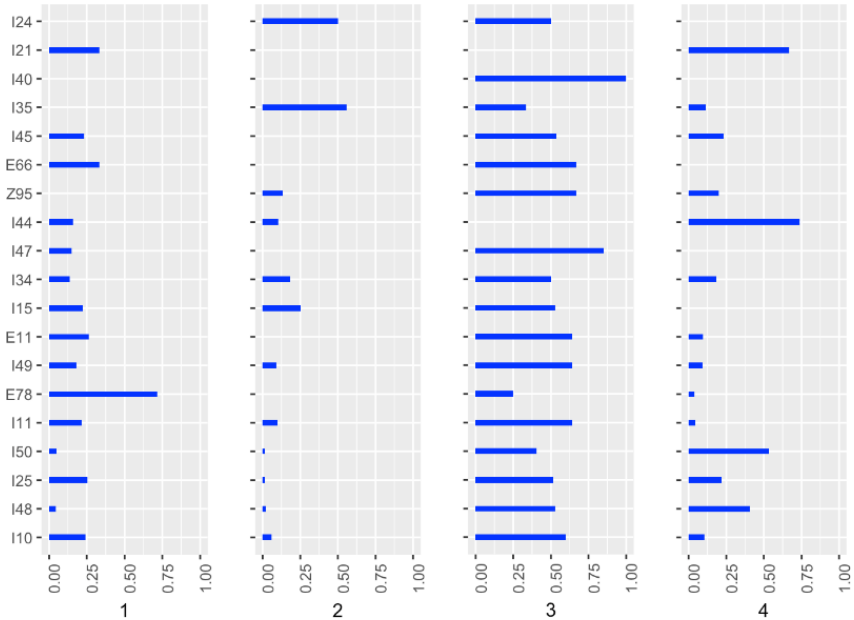


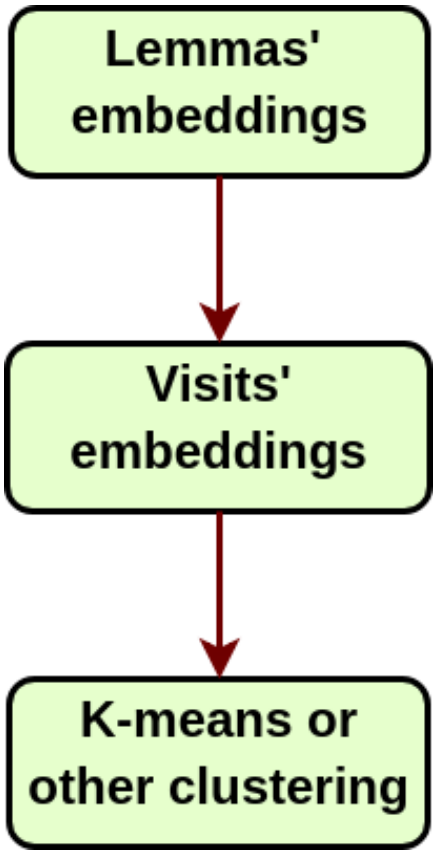
Figure: Diseases distribution into clusters.

Problems

- Unbalanced clusters
- Number of clusters in results is different than assumed

Second idea

Usage of interview and examination description.



Text preprocessing

Text preprocessing is done by our collaborators from IPI PAN.

```
"Nos-dsn w prawo ,drożność nosa upośledzona¶Gardło-bez
zmian zapalnych,migdałki podniebienne bez retencji
przerosniete,języczek powiększony¶Krtań-szpara głośni
wolna ,struny głosowe symetrycznie ruchome ¶Otoskopowo
UP-bł.bębenkowa z refleksem¶                    UL-
bł.bębenkowa z refleksem¶szyja palpacyjnie bez zmian"
```

	lemma	type
1	bez zmian zapalnych	obj
2	bez zmian	wynik
3	bez	NEG
4	bębenkowy	lokAnat
5	drożność nosa	bad
6	gardło	anat
7	głośny	cecha
8	krtań	anat
9	migdałek podniebienny	anat
10	nos	anat
11	otoskopowo	bad
12	prawy	later
13	refleks	fiz
14	ruchomy	cecha
15	struna głosowa	anat
16	szpara	anat-frag
17	szyja	anat
18	up	anat
19	upośledzony	wCechy
20	wieczorem	pora
21	wolny	wCechy
22	nic	NEG
23	ujemny	wCechy
24	powiększony	cecha
25	powiększyć	cecha
26	ul	anat
27	ul	jedn

GloVe algorithm

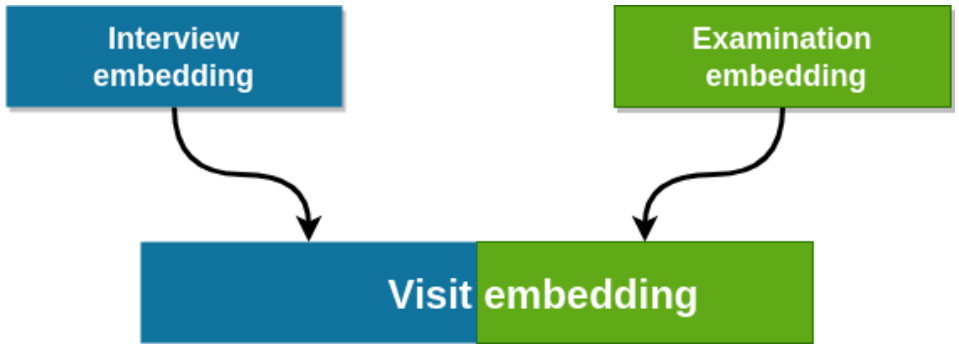
- Build X – Term Cooccurrence Matrix (TCM)
- Minimize: $J = \sum_{i,j=1}^{v} f(X_{ij})(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - log(X_{ij}))^2$, where $b_i,\ \tilde{b}_j \in \mathbb{R},\ w_i, \tilde{w}_i \in \mathbb{R}^n$,

$$f(x) = \begin{cases} (x/x_{max})^{\alpha} \ if \ x < x_{max}, \\ 1 \qquad\qquad otherwise. \end{cases}$$

- Return $w_i + \tilde{w}_i$

Visit embeddings

- embedding of description as an average of terms' embeddings
- arithmetic mean or mean weighted by TF-IDF
- make interview embedding and examination embedding separately



Clustering

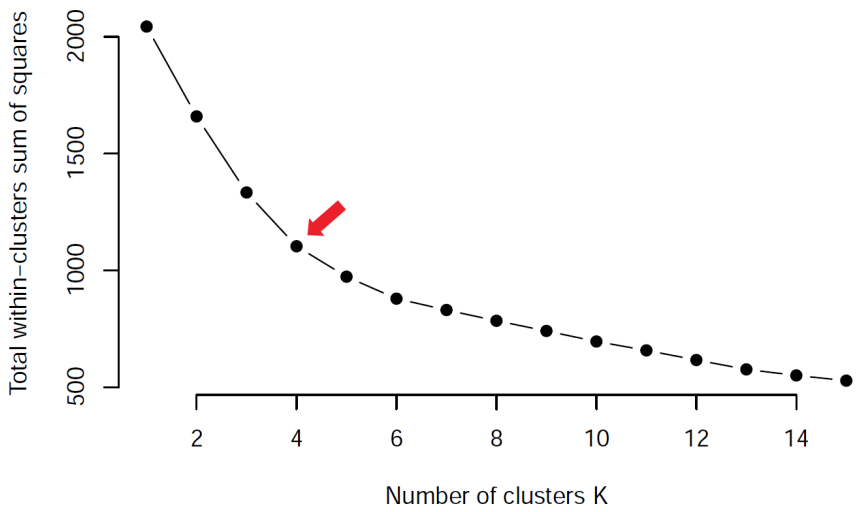
- K-means
- Hierarchical clustering

Comparision of clusterings by adjusted Rand index.

Rand index: $R = \frac{a+b}{a+b+c+d}$

Adjusted Rand index: corrected-for-chance version of the Rand index.

Determining optimal number of clusters

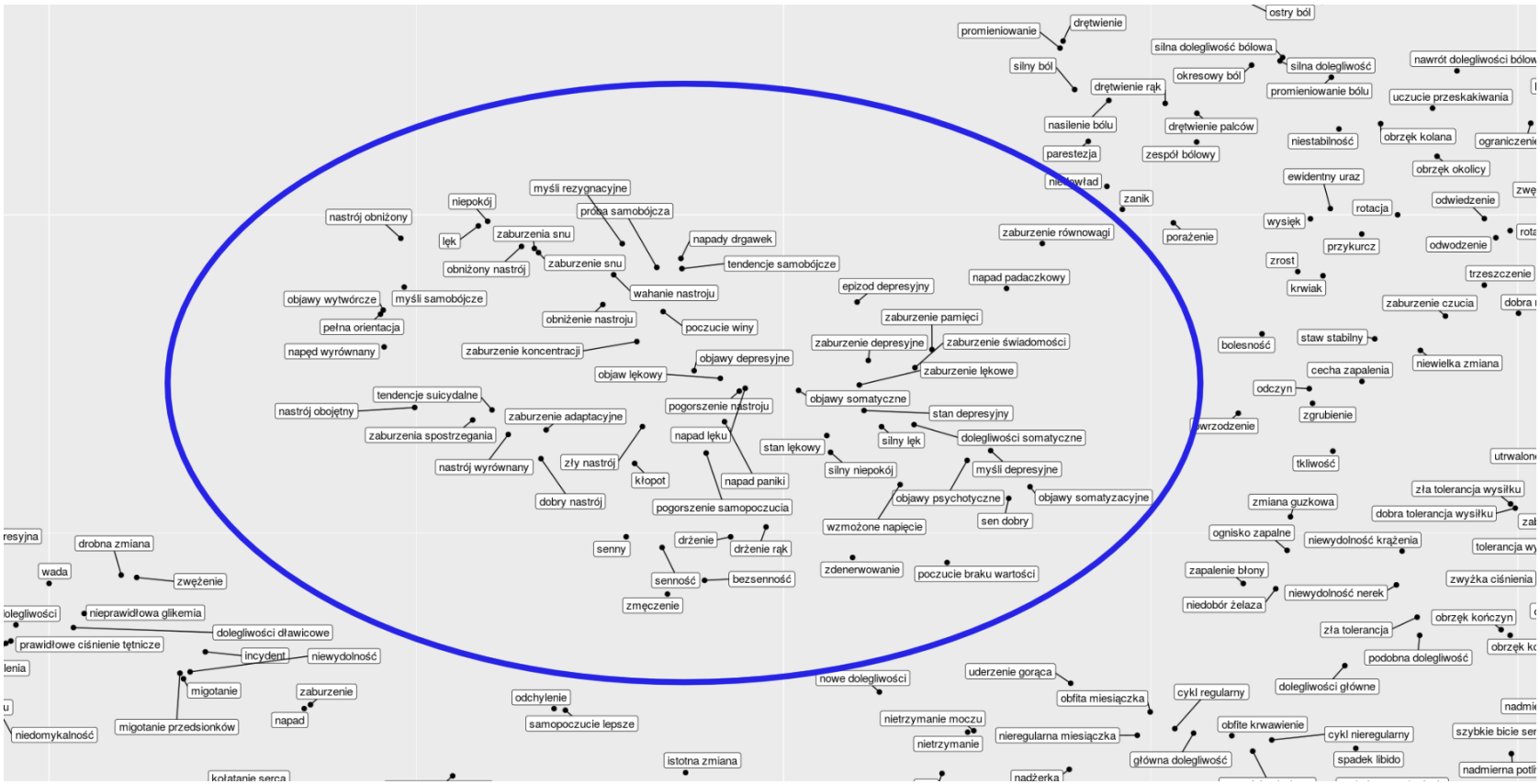


Elbow method: for each k plot sum of squares of differences between elements in clusters and clusters' centers.

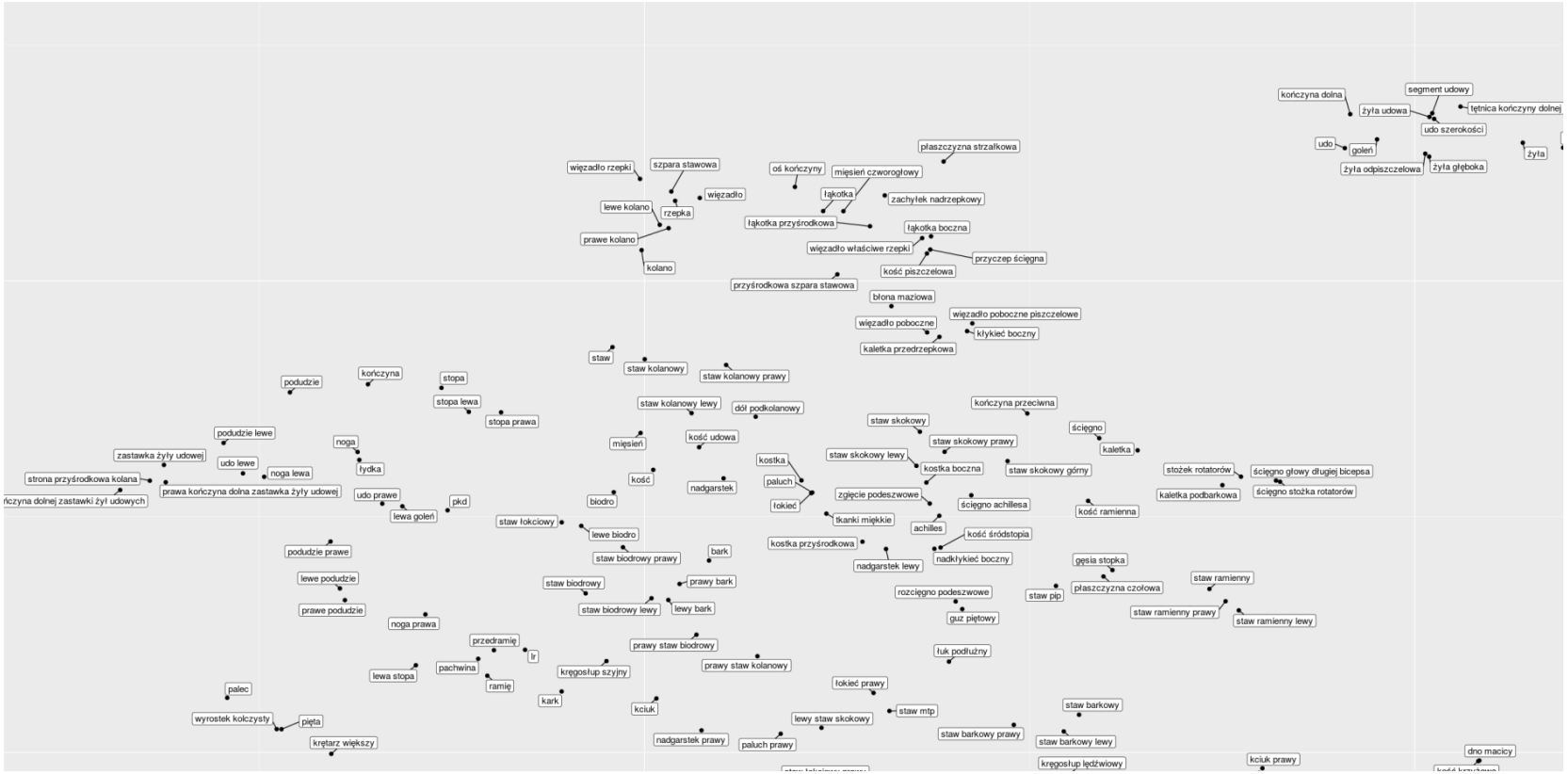
Term embeddings

- 3816 interview terms, 3559 examination terms
- visualisation using t-SNE or PCA
- 3 different embeddings (for interview, examination and recommendation)
- 137 categories
- top categories: symptom, anatomic, feature, process, disease

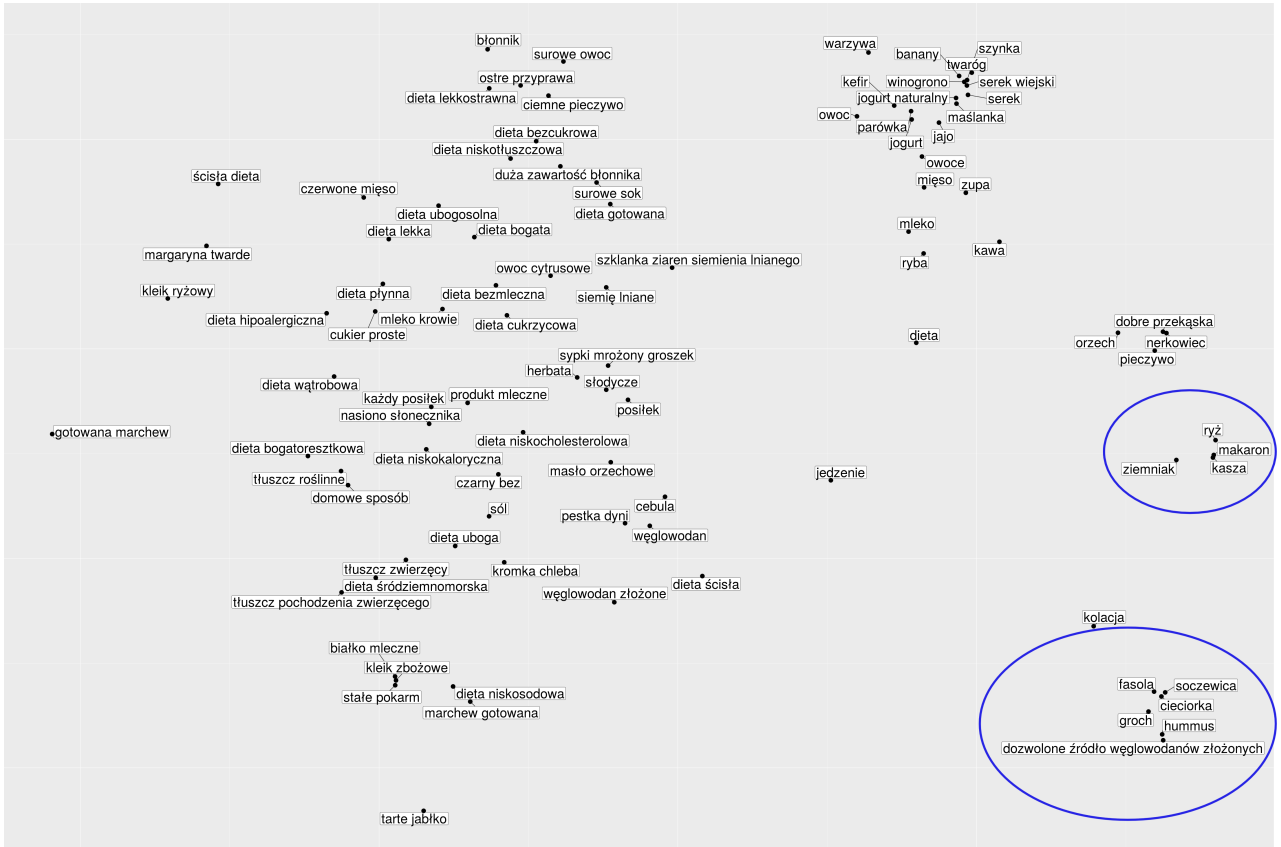
Symptoms embedding fragment (from interview embeddings)



Anatomic embedding fragment (from examination embeddings)



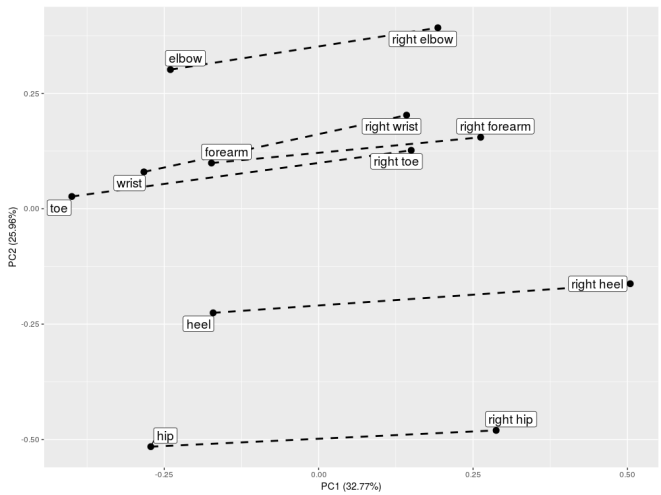
Diet embedding fragment (from recommendation embeddings)

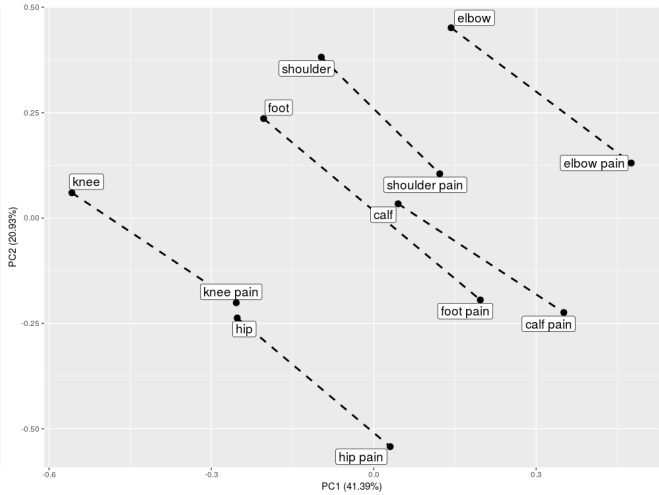


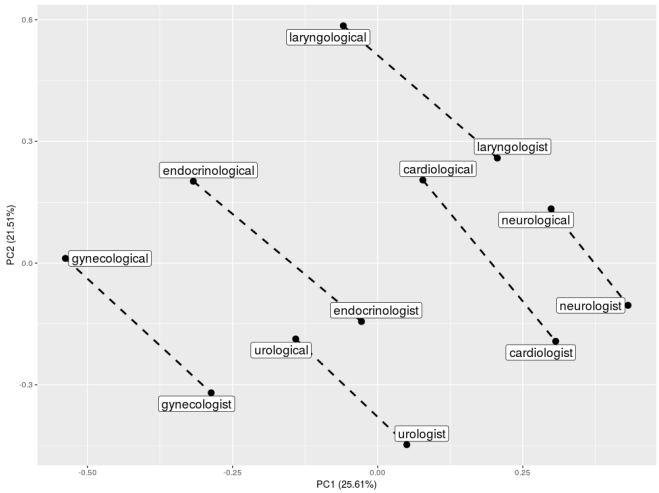
Term analogy task

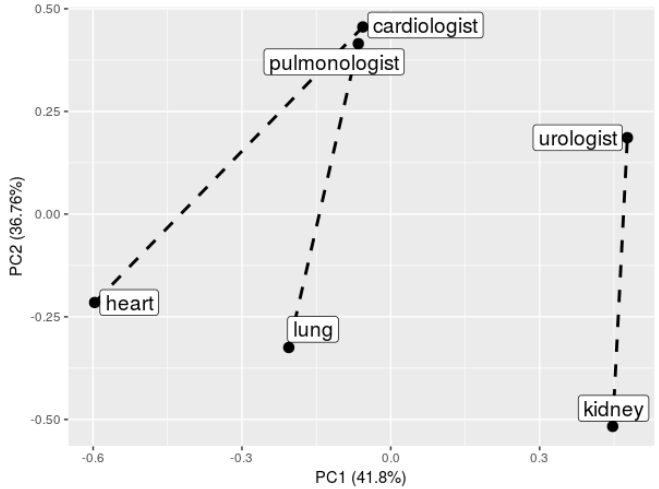
Type of relationship	# Pairs	Term Pair 1		Term Pair 2	
Body part – Pain	22	eye	eye pain	foot	foot pain
Speciality – Adjective	7	dermatologist	dermatological	neurologist	neurological
Body part – Right side	34	hand	right hand	knee	right knee
Body part – Left side	32	thumb	left thumb	heel	left heel
Spec. – Consultation	11	surgeon	surgical consult.	gynecologist	g. consult.
Speciality - Body part	9	cardiologist	heart	oculist	eye
Man - Woman	9	patient (male)	patient (female)	brother	sister

$$\text{eye} - \text{eye pain} + \text{foot pain} \approx \text{foot}$$


elbow
right elbow
forearm
right wrist
right forearm
right toe
wrist
toe
right heel
heel
hip
right hip
PC1 (32.77%)
PC2 (25.96%)


elbow
shoulder
foot
elbow pain
shoulder pain
knee
calf
knee pain
foot pain
calf pain
hip
hip pain
PC1 (41.39%)
PC2 (20.93%)


laryngological
laryngologist
endocrinological
cardiological
neurological
gynecological
endocrinologist
neurologist
urological
cardiologist
gynecologist
urologist
PC1 (25.61%)
PC2 (21.51%)


cardiologist
pulmonologist
urologist
heart
lung
kidney
PC1 (41.8%)
PC2 (36.76%)

Dim. / Context	1	3	5
10	0.1293	0.2189	0.2827
15	0.1701	0.3081	0.4123
20	**0.1702**	0.3749	0.4662
25	0.1667	0.4120	0.5220
30	0.1674	0.4675	0.5755
40	0.1460	**0.5017**	0.6070
50	0.1518	0.4966	**0.6190**
100	0.0435	0.4231	0.5483
200	0.0261	0.3058	0.4410

Tabela: Mean accuracy of correct answers on term analogy tasks.

Embedding similarity

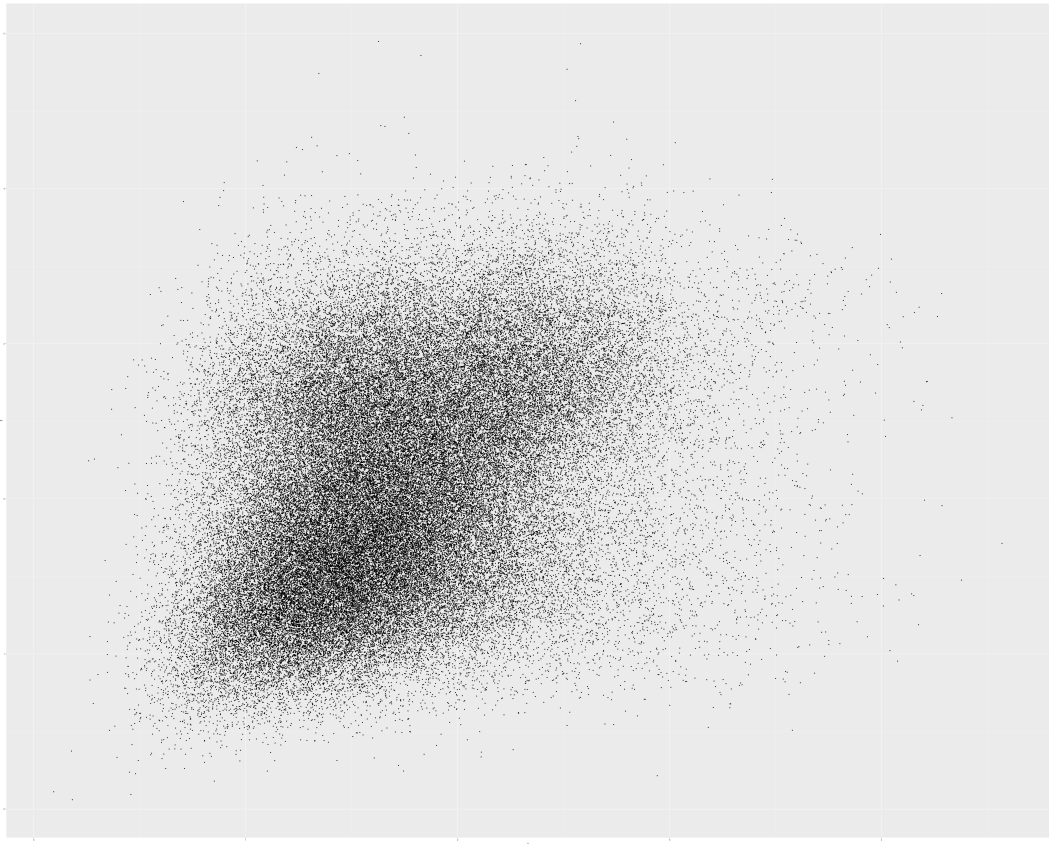
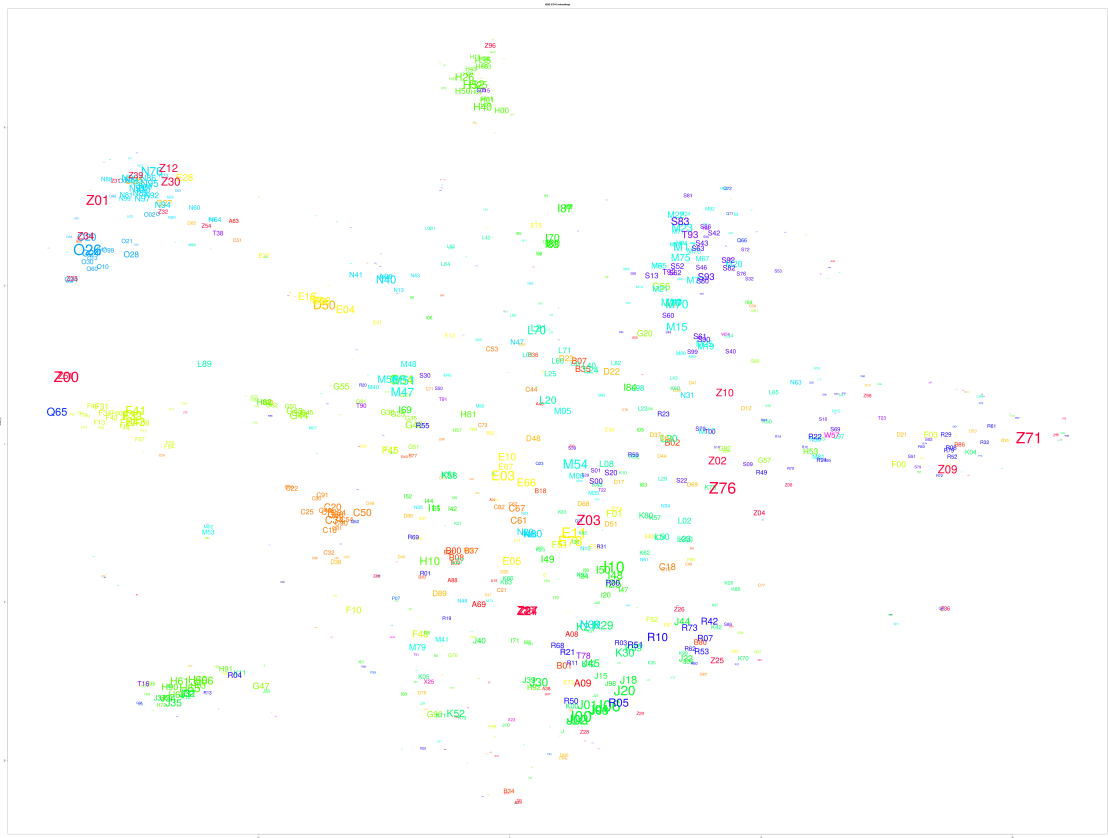


Figure: Interview embedding distance vs. examination embedding distance for anatomic words.

ICD-10 embeddings

ICD-10 code embeddings: arithmetic mean of all visits' embeddings



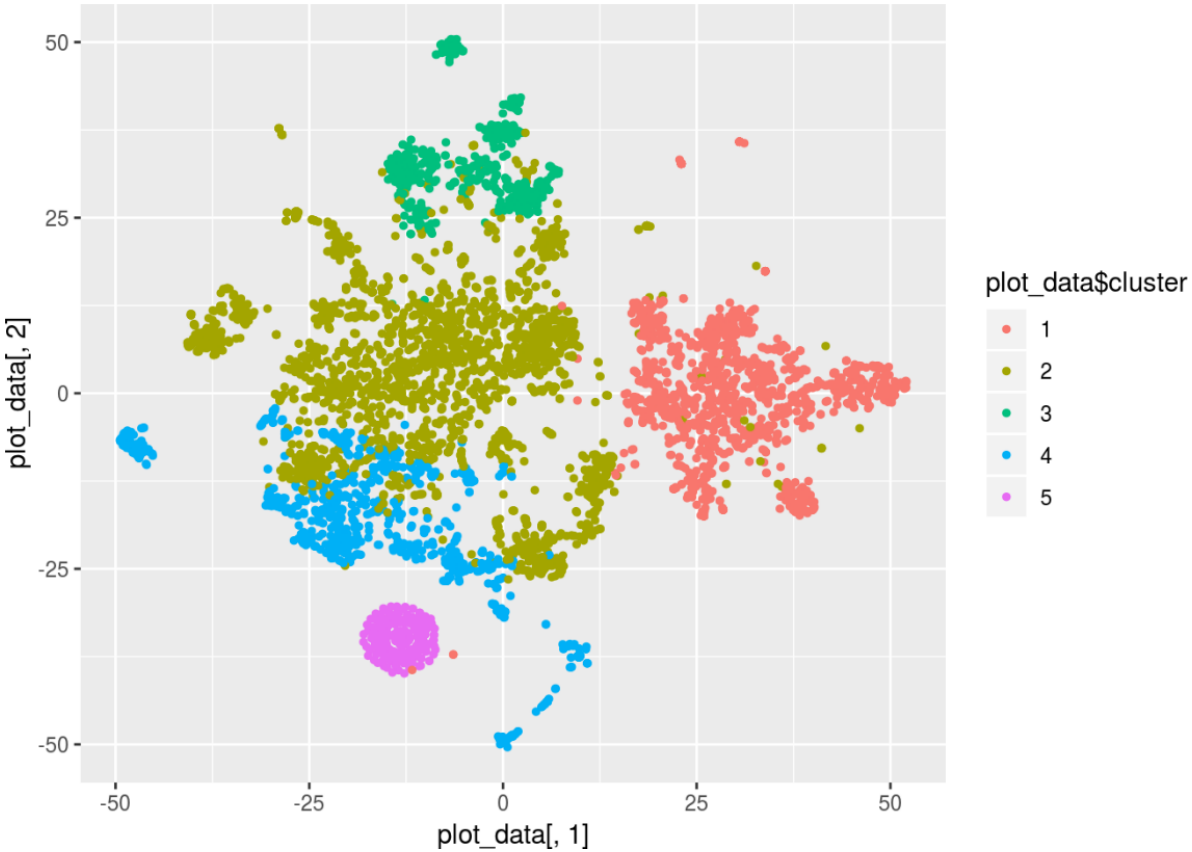
Visits segmentation

- Separate clusterings for each doctor speciality: internal medicine, family medicine, pediatrics, gynecology etc.
- Adjusted Rand index between Tensor decomposition and embedding-based clustering: 0.06 - 0.09

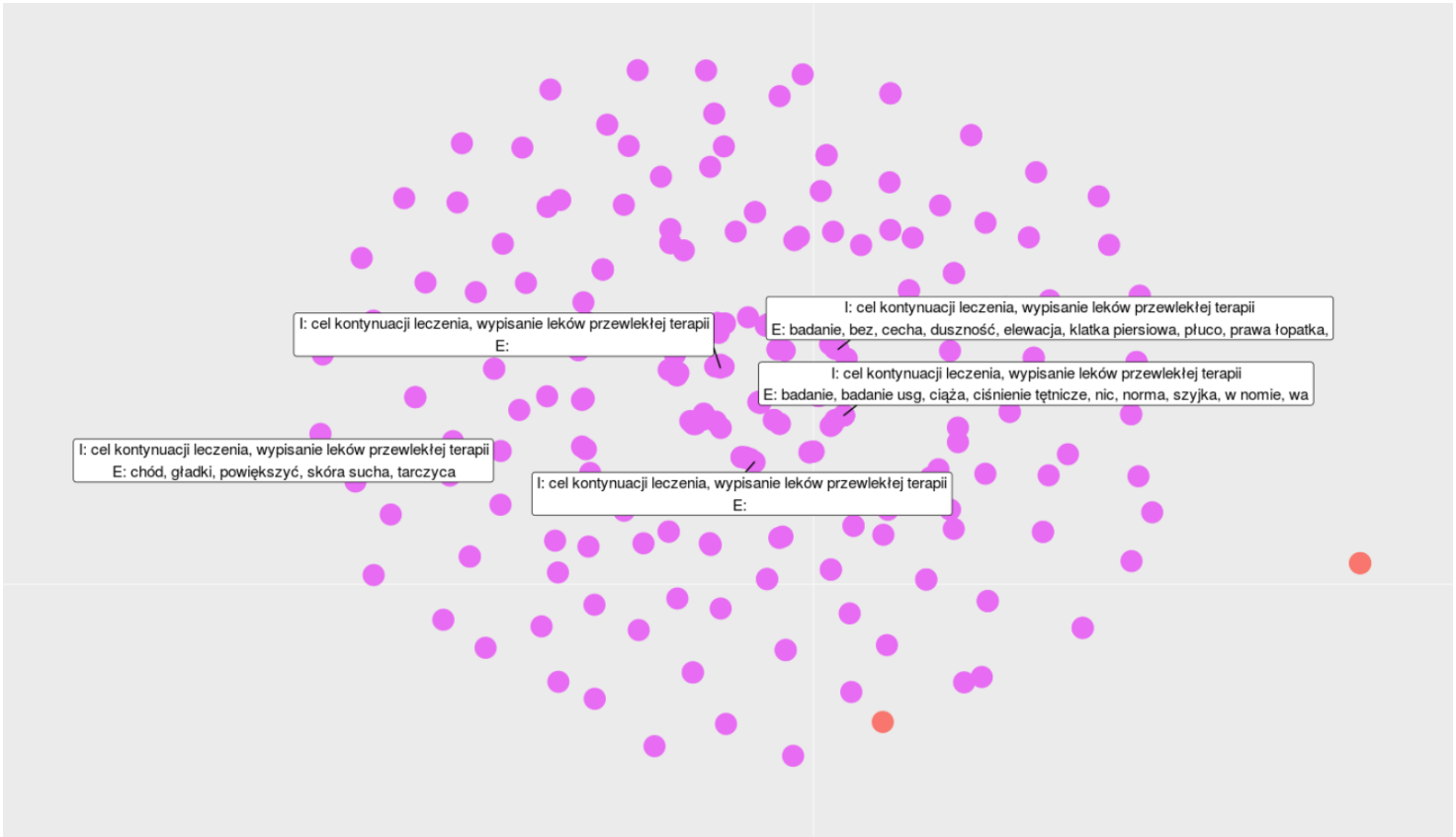
Domain	# clusters	# visits	clusters' size	K-means - hclust
Cardiology	6	1201	428, 193, 134, 303, 27, 116	0.87
Dermatology and venereology	6	1204	455, 89, 176, 30, 391, 63	0.64
Endocrinology	5	1510	389, 412, 208, 183, 318	0.8
Family medicine	6	11230	3108, 2353, 601, 4518, 255, 395	0.69
Gynecology	4	3456	1311, 1318, 384, 443	0.8
Internal medicine	5	6419	1915, 1173, 1930, 1146, 255	0.76
Orthopedics	4	1869	360, 1257, 102, 150	0.19
Pediatrics	5	4742	1751, 658, 666, 715, 952	0.46
Psychiatry	5	1012	441, 184, 179, 133, 75	0.81

Internal medicine segmentation

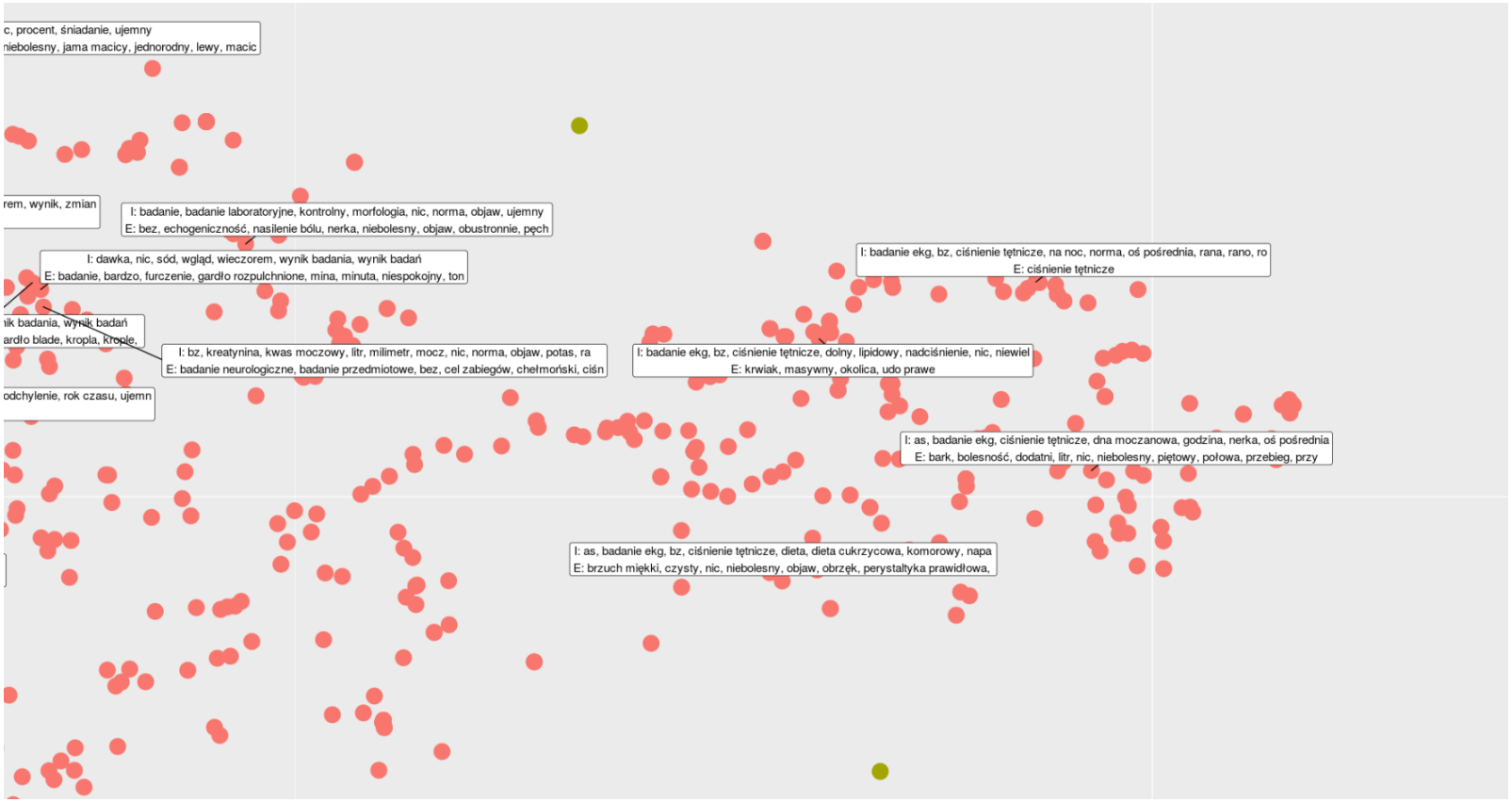
5435 visits



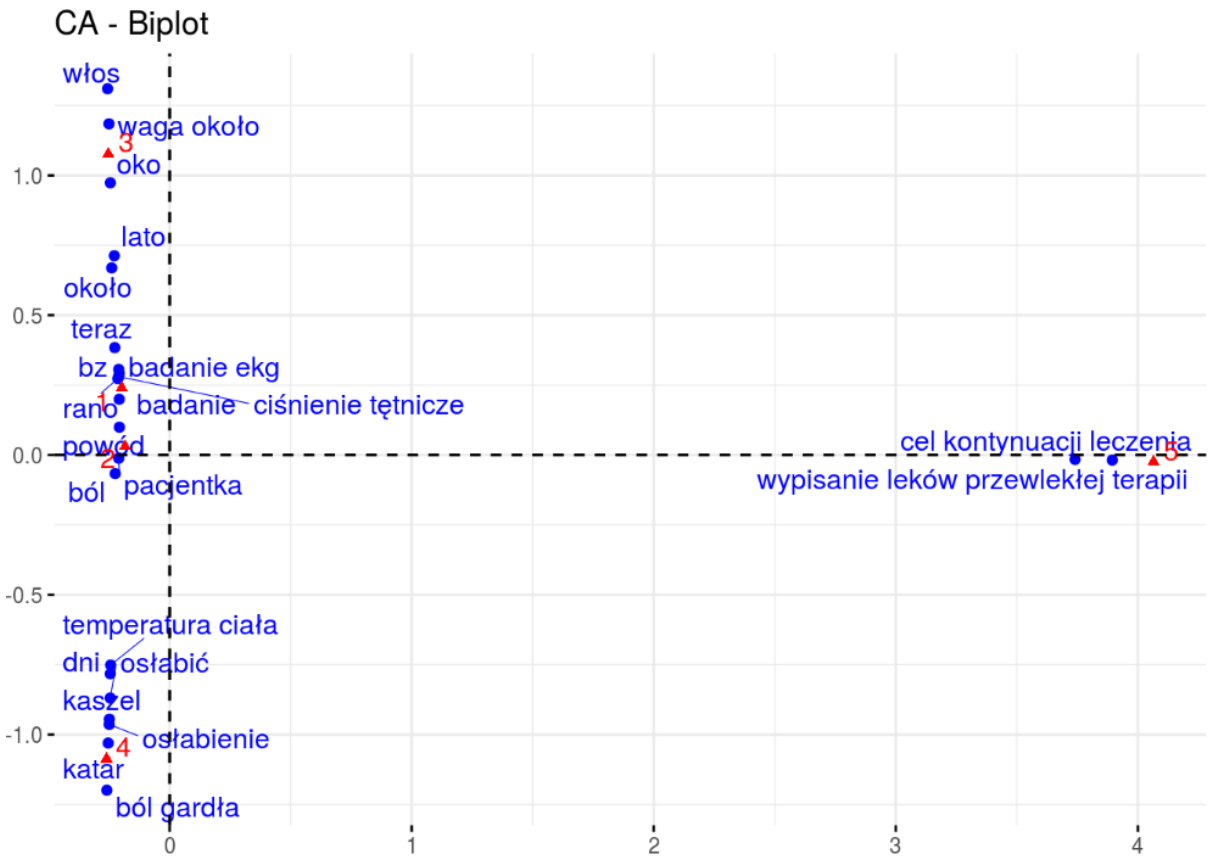
Internal medicine segmentation



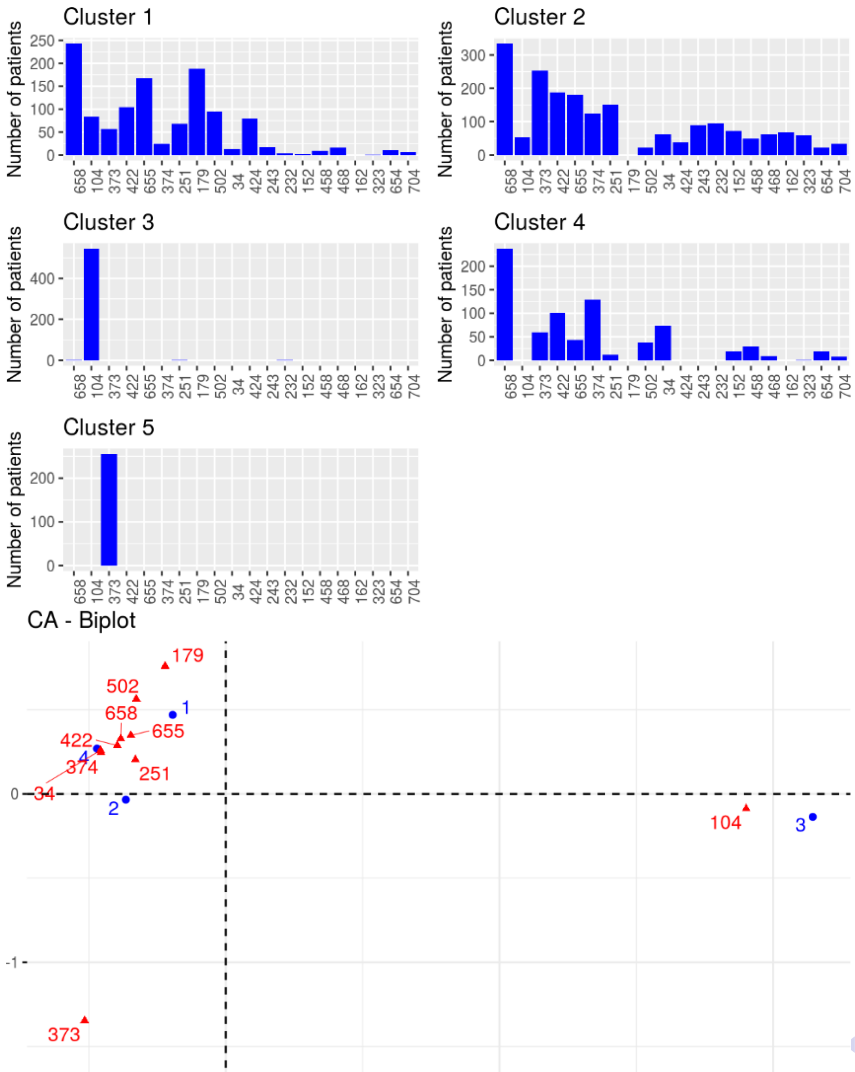
Internal medicine segmentation



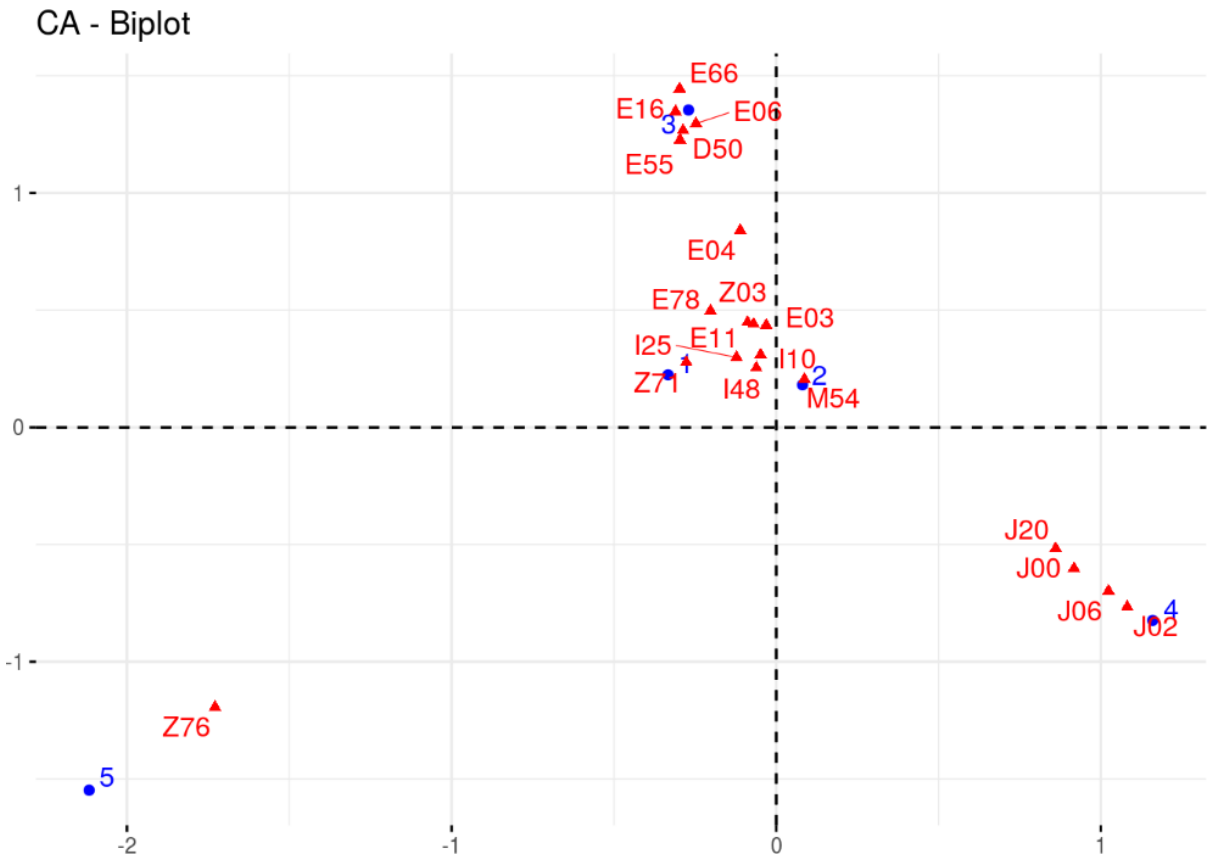
Characteristic terms for clusters



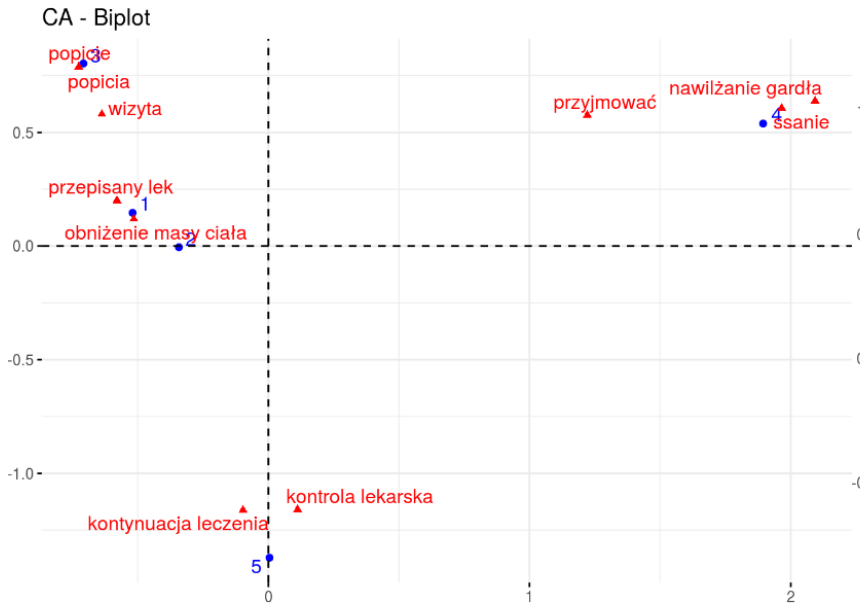
Cluster analyse – doctor ID

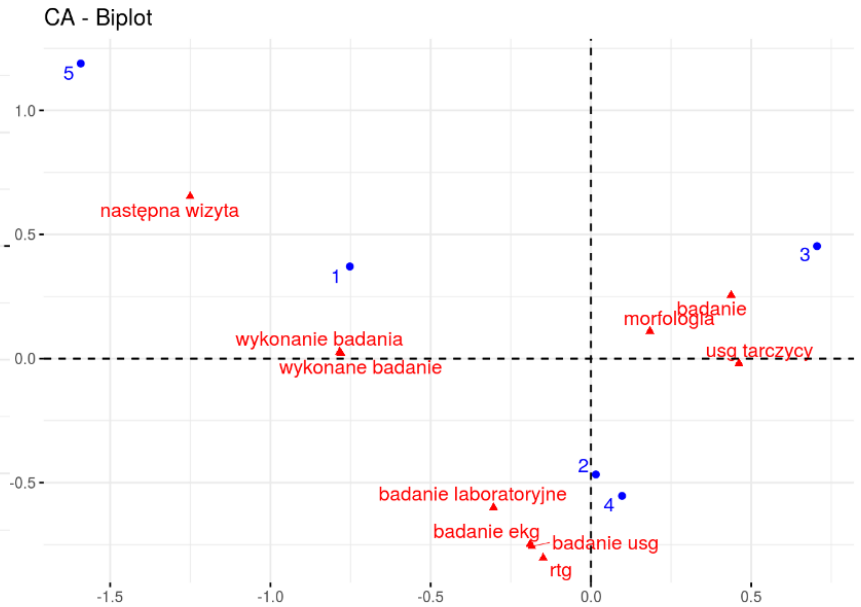


Codes ICD-10


CA - Biplot
E66
E16
3
E06
D50
E55
E04
E78
Z03
E03
I25
E11
I10
Z71
1
I48
2
M54
J20
J00
J06
4
J02
Z76
5
1
0
-1
-2
-1
0
1

Recommendations





Recommendations for cardiology

cluster	5 most common recommendations
1	control (23.4%), medicament (20.6%), therapy (16.4%), treat (15.4%), holter ekg (7.5%)
2	next visit (59.6%), echo of the heart (8.8%), control (4.1%), visit (3.1%), therapy (1.6%)
3	meal (18.7%), sugar-free diet (17.2%), physical activity (14.2%), programmable effort (14.2%), medicament (12.7%)
4	control (25.4%), medicament (20.8%), control of blood pressure (12.9%), therapy (11.9%), treat (11.9%)
5	helpline (100%), cardiomedical (100%), therapy (74.1%), treat (74.1%), body weight reduction (66.7%)
6	helpline (90.5%), cardiomedical (90.5%), therapy (82.8%), treat (82.8%), body weight reduction (61.2%)

cluster	5 most common recommendations in a given cluster
1	holter ekg (7.5%), continuation of treatment (7%), ekg examination (5.6%), everyday walk (5.1%), abidance of recommended diet (5.1%)
2	echo of the heart (8.8%), visit (3.1%), b12 (1%), ekg examination (0.5%), everyday walk (0.5%)
3	meal (18.7%), sugar-free diet (17.2%), physical activity (14.2%), programmable effort (14.2%), regular measurement of blood pressure (12.7%)
4	control of blood pressure (12.9%), echo of the heart (11.2%), holter ekg (9.2%), repeat control (7.6%), urgent consultation (6.6%)
5	helpline (100%), cardiomedical (100%), body weight reduction (66.7%), performing the examination (48.1%), isotonic effort (44.4%)

Conclusions

- a new method for segmentation of visits in health centers based on descriptions written by doctors
- creating medical terms' embeddings and embeddings of ICD-10 codes
- validation by a specific term analogy task
- segmentation based on visits' embeddings

Applications

- segmentation can be used to assign new visits to already derived clusters
- based on description of an interview or a description of patient examination we can identify similar visits and show corresponding recommendations

References

J. Pennington, R. Socher, C. D. Manning (2014)

GloVe: Global Vectors for Word Representation

https://www.aclweb.org/anthology/D14-1162

L. van der Maaten, G. Hinton (2008)

Visualizing Data using t-SNE

http://jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

Acknowledgements

This work was financially supported by the grant of Polish Centre for Research and Development POIR.01.01.01-00-0328/17.

Questions, comments...

The End