Wspomaganie przesiewania dokumentów w przeglądach systematycznych z użyciem uczenia maszynowego i eksploracji tekstu

Piotr Przybyła
piotr.przybyla@manchester.ac.uk

National Centre for Text Mining, University of Manchester

4 czerwca 2018

Plan prezentacji

MANCHESTER 1824
The University of Manchester

Motywacja

Podstawa

- Uczenie aktywne
- System RobotAnalyst
- Ewaluacja

Rozszerzenia

- Grupowanie
- Przyspieszanie uczenia
- Analiza błędów

Zakończenie

Motywacja

Przeglądy systematyczne

Podstawowe narzędzie w naukach biomedycznych:

- Przegląd stanu wiedzy na temat danego zagadnienia,
- Systematyczne: jasno określone i udokumentowane czynności:
  1. Zdefiniowanie pytania i kryteriów włączania,
  2. Wyszukiwanie literatury,
  3. **Przesiewanie wyników na podstawie streszczeń**,
  4. Analiza pełnej treści artykułów,
  5. Agregacja i synteza wyników.
- Standaryzacja i propagowanie: *Cochrane Collaboration*

Czy witamina C pomaga zapobiegać przeziębieniom?
`https://www.ncbi.nlm.nih.gov/pubmed/17636648`

Przesiewanie wyników

- Wejście: od 1.000 do 100.000 pozycji literatury (tytuł, abstrakt, metadane),
- Metoda: ocena spełniania kryteriów włączenia,
- Wyjście: przypisane etykiety: Włączony/Wyłączony/Niepewny
- Czas pracy: dla 5.000 pozycji 83-125 godzin pracy eksperta,
- Koszt: $\sim$£13,000 za jeden przegląd.

Projekty

- *Supporting Evidence-based Public Health Interventions using Text Mining*
  - University of Manchester (NaCTeM), University of Liverpool i National Institute of Health and Care Excellence (NICE)
  - Zastosowanie: wytyczne zdrowotne dla NHS
  - Grant UK Medical Research Council MR/L01078X/1
- *SLiM: Pilot study of the utility of text mining and machine learning tools to accelerate systematic review and meta-analysis of findings of in vivo research*
  - NaCTeM, University of Edinburgh, University College London, Imperial College London
  - Zastosowanie: przeglądy systematyczne badań *in vivo*
  - Grant UK Medical Research Council MR/N015665/1

$\rightarrow$ *RobotAnalyst*

Wspomaganie

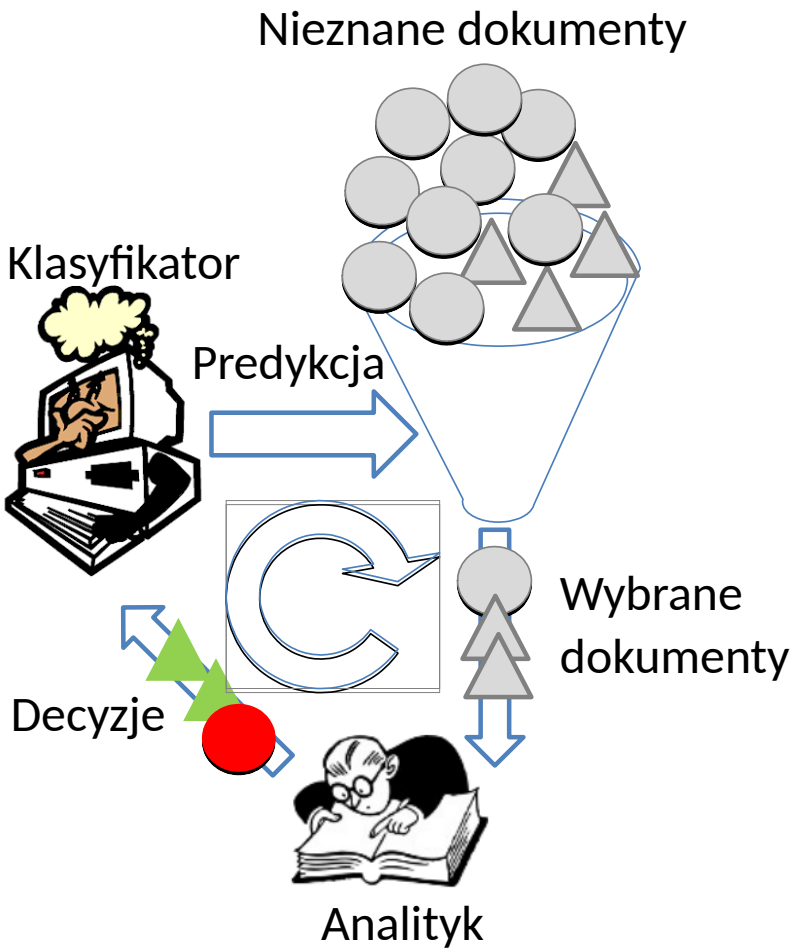
Zastosowanie technik ML i analizy tekstu w kierunku:

1. Uczenia nienadzorowanego:
   - wykrywanie pojęć wielowyrazowych,
   - grupowanie tematycznego,
   - wyszukiwanie podobnych dokumentów.
2. Uczenia nadzorowanego:
   - 2.1 zbieranie decyzji użytkownika,
   - 2.2 budowa modelu klasyfikującego włączone/wyłączone dokumenty,
   - 2.3 stosowanie modelu do nieoznakowanych dokumentów.

Włączenie analityka w proces: czarna skrzynka niemile widziana.

Podstawa

Uczenie aktywne



1. Wybór początkowej grupy dokumentów do oznakowania,
2. Wskazanie decyzji przez analityka,
3. Dodanie dokumentów do zbioru treningowego,
4. Budowa modelu,
5. Zastosowanie modelu do nieznanych dokumentów i wybór do następnej iteracji,
6. Powrót do (2).

Wybór przypadków

Ma na celu wspomaganie:

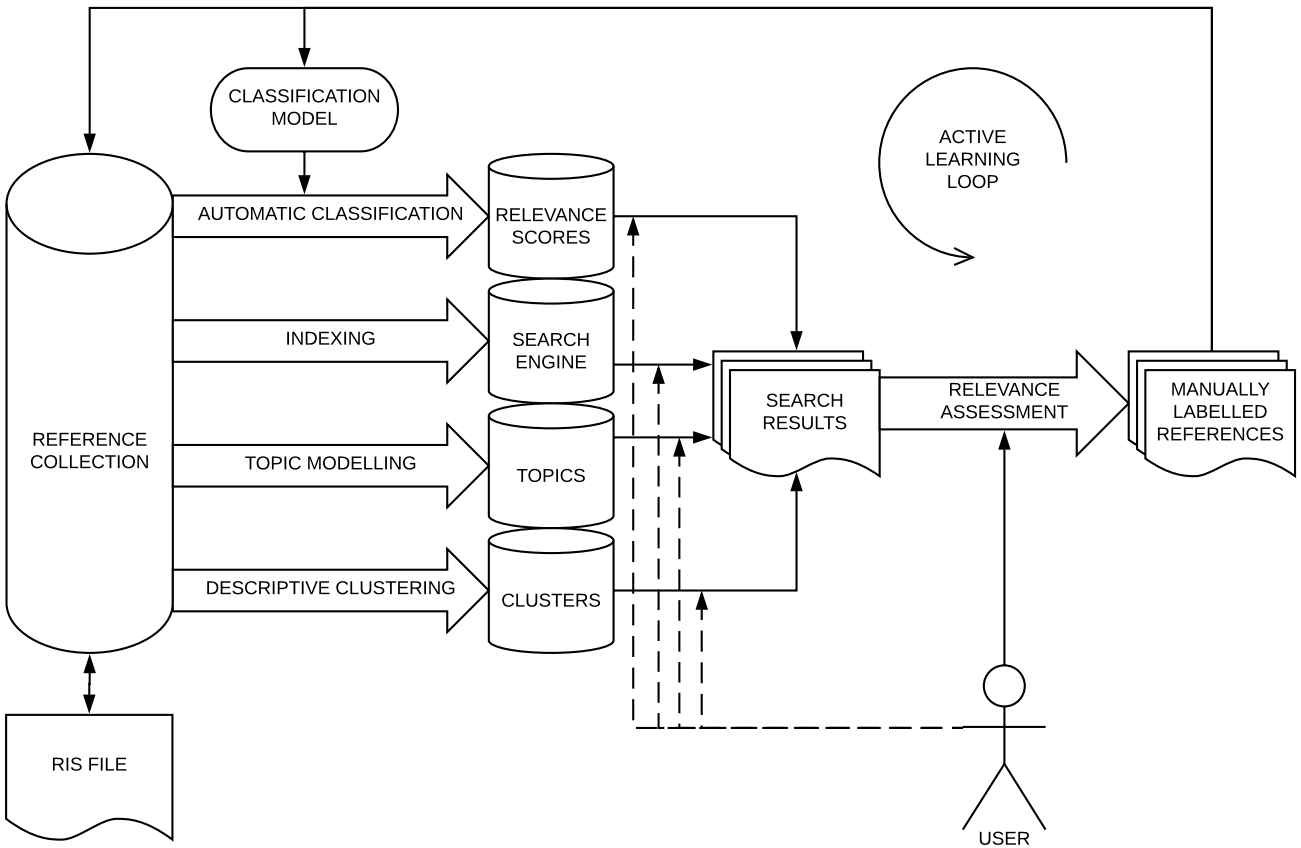
- klasyfikatora poprzez wybór *dobrego* zbioru uczącego (równoważenie!),
- analityka poprzez wybór *interesujących* dokumentów.

Dwie strategie:

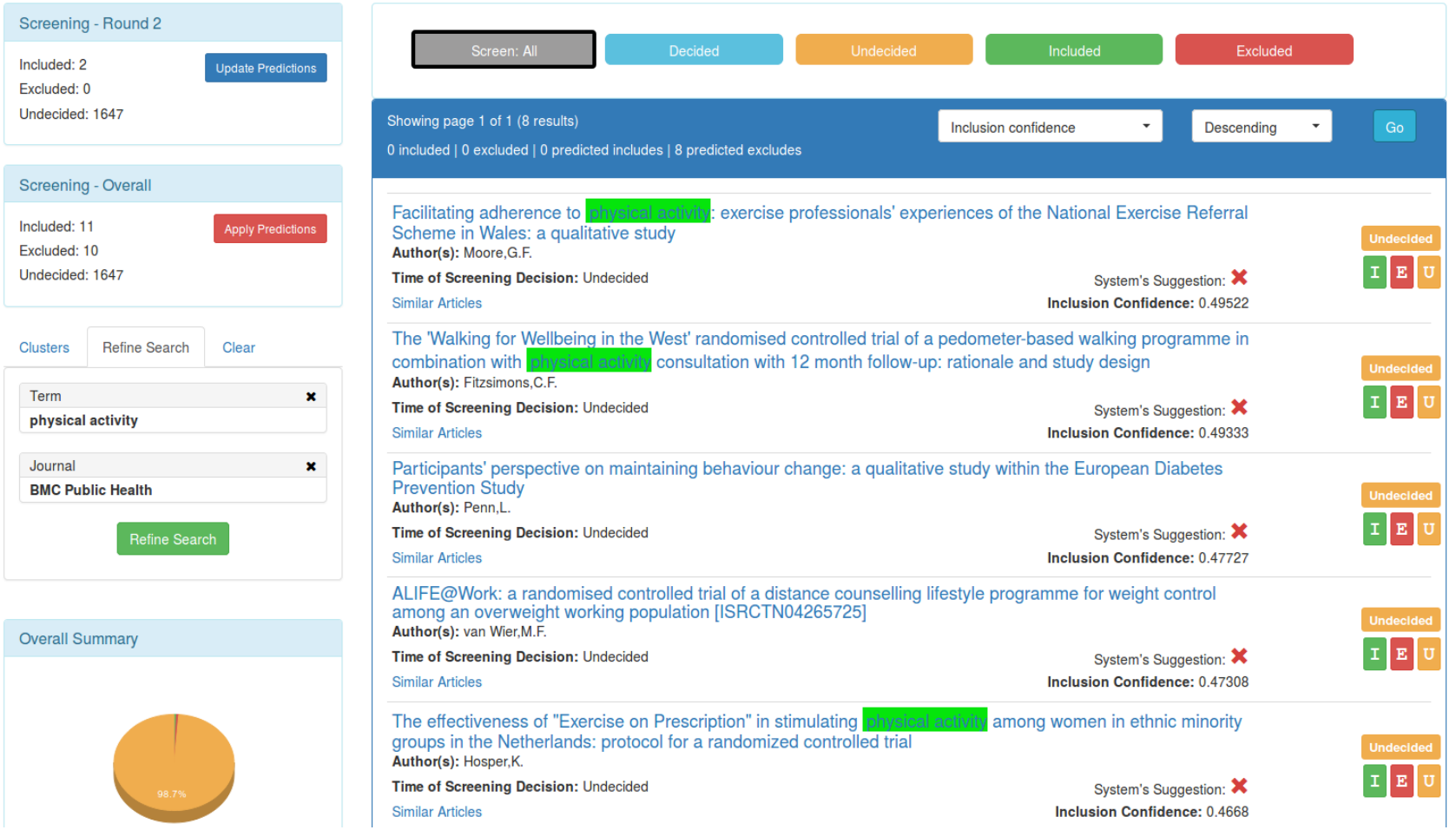
- *uncertainty-based*: przypadki najbliżej granicy decyzyjnej,
- *relevancy-based*: przypadki najbardziej obiecujące [Miwa et al., 2014].

Etykietowanie pozostałych dokumentów: według modelu lub negatywnie, odpowiednio.

SystemRobotAnalyst



Collection: BC test



Klasyfikacja

1. Lematyzacja tytułów i abstraktów (GENIA tagger),
2. Budowa modelu 300 tematów LDA (MALLET),
3. Cechy:
   - L2-normalizowany wektor TF/IDF dla lematów słów z tytułu,
   - analogicznie dla abstraktu,
   - współczynniki tematów z LDA,
4. Klasyfikacja modelem liniowym SVM ze stratą L2 i regularyzacją L2 (liblinear),
5. Konwersja wartości do przedziału (0,1) poprzez nałożenie funkcji logistycznej.

Miary wydajności

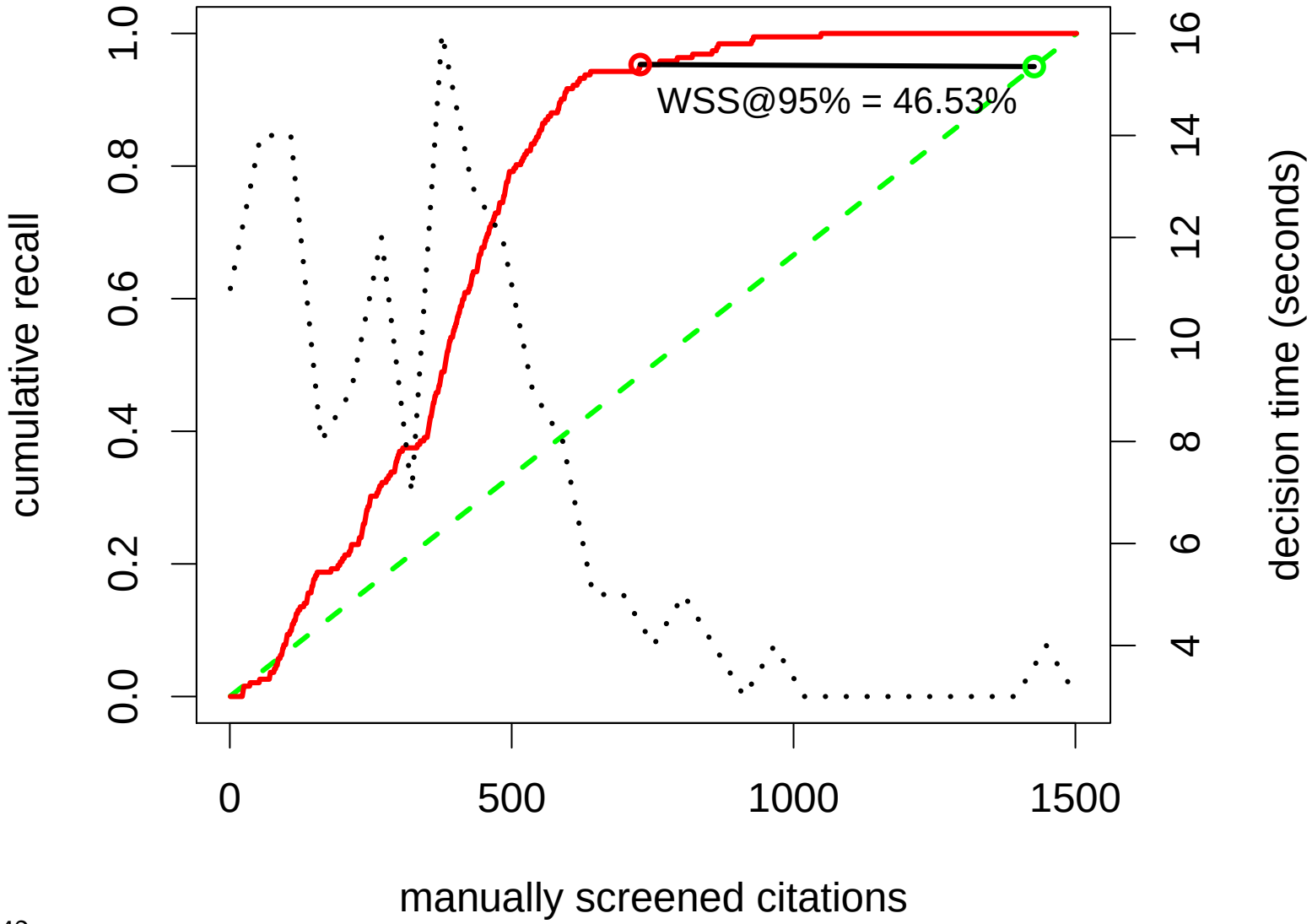
Cel: minimalizacja nakładu pracy przy założonym poziomie czułości (95%).

$$recall(i) = \frac{TP(i)}{TP(i) + FN(i)},$$

$$i_{\mathrm{R95}} = \min\{i \in \{1, \ldots, N\} : recall(i) \geqslant 0.95\}$$

$$WSS@95\% = 0.95 - \frac{TP(i_{\mathrm{R95}}) + FP(i_{\mathrm{R95}})}{N} = 0.95 - \frac{i_{\mathrm{R95}}}{N},$$

Przykład


WSS@95% = 46.53%
cumulative recall
0.0
0.2
0.4
0.6
0.8
1.0
decision time (seconds)
4
6
8
10
12
14
16
0
500
1000
1500
manually screened citations

Kolekcje dokumentów

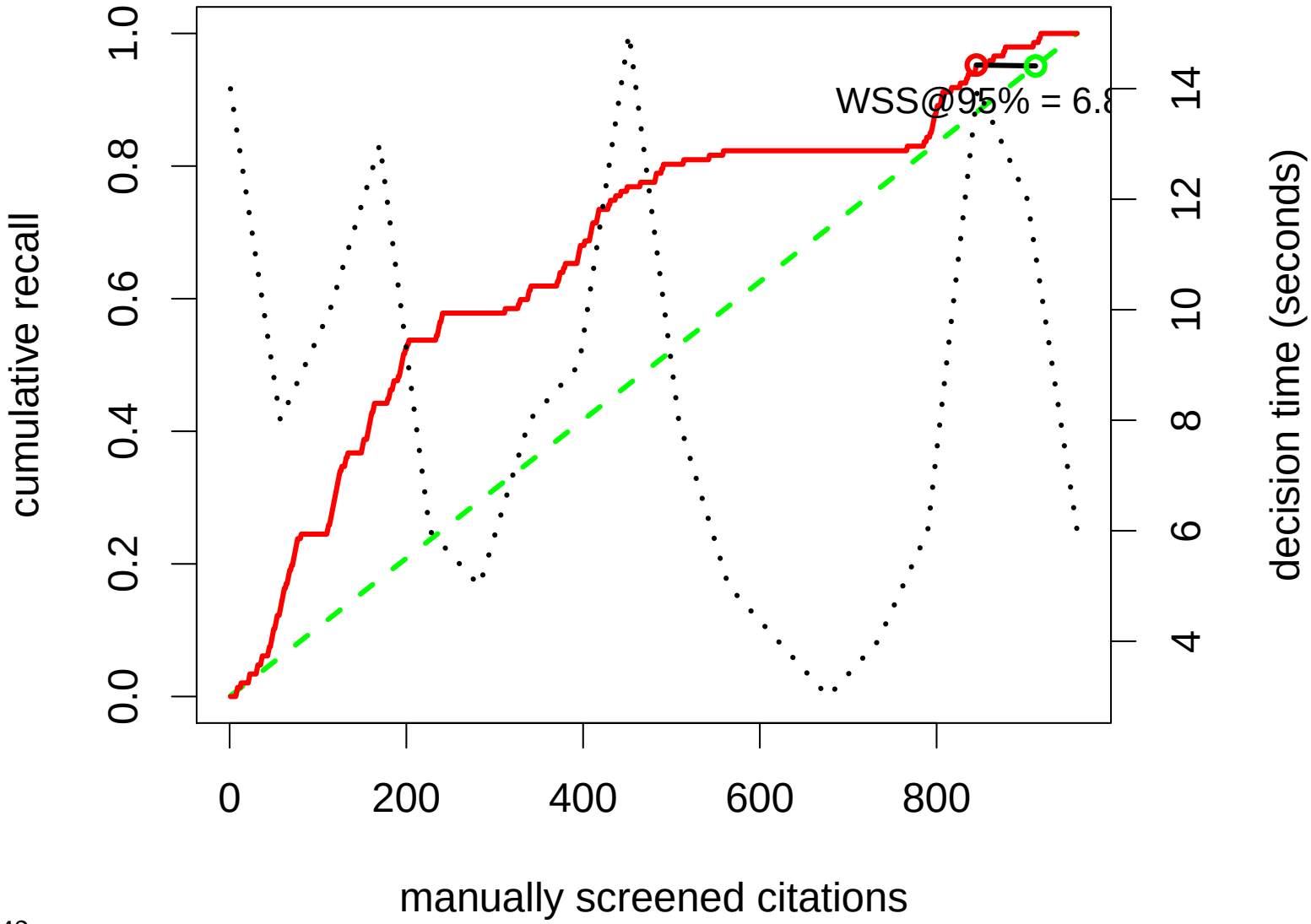
Pełna ewaluacja [Przybyła et al., w recenzji] obejmowała 22 przeglądy wykonane w całości.

Kolekcja	Temat	Pochodzenie	Rozmiar	Włączenia
TUB	Tuberculosis	NICE	4678	2.42%
BC	Behaviour change: individual approaches	NICE	1502	13.72%
BC-S	Behaviour change: individual approaches (surv.)	NICE	937	21.66%
BC-C	Choice architecture in behaviour change (surv.)	NICE	959	15.33%
WC-D	Walking and cycling (surv., database search)	NICE	304	27.30%
WC-C	Walking and cycling (surv., citation search)	NICE	468	12.18%
WC-F	Walking and cycling (surv., focused search)	NICE	86	9.30%
PAP	Physical activity and pregnancy	NICE	320	11.88%
WGP	Weight gain and pregnancy	NICE	110	11.82%
PW-S	Preventing excess weight gain (surv., self-weighing)	NICE	157	8.28%
PW-E	Preventing excess weight gain (surv., eating patterns)	NICE	719	5.15%
WM	Weight management (surv.)	NICE	665	29.62%
SH	Sexual health	NICE	3760	1.36%
QSH	Quality and safety in hospitals	IUMSP	4964	18.63%
LD	Learning difficulties	NICE	2148	0.28%
OCM	Osteoarthritis: care and management (surv.)	NICE	2986	15.00%
HB	Hepatitis B: diagnosis and management (surv.)	NICE	1523	3.81%

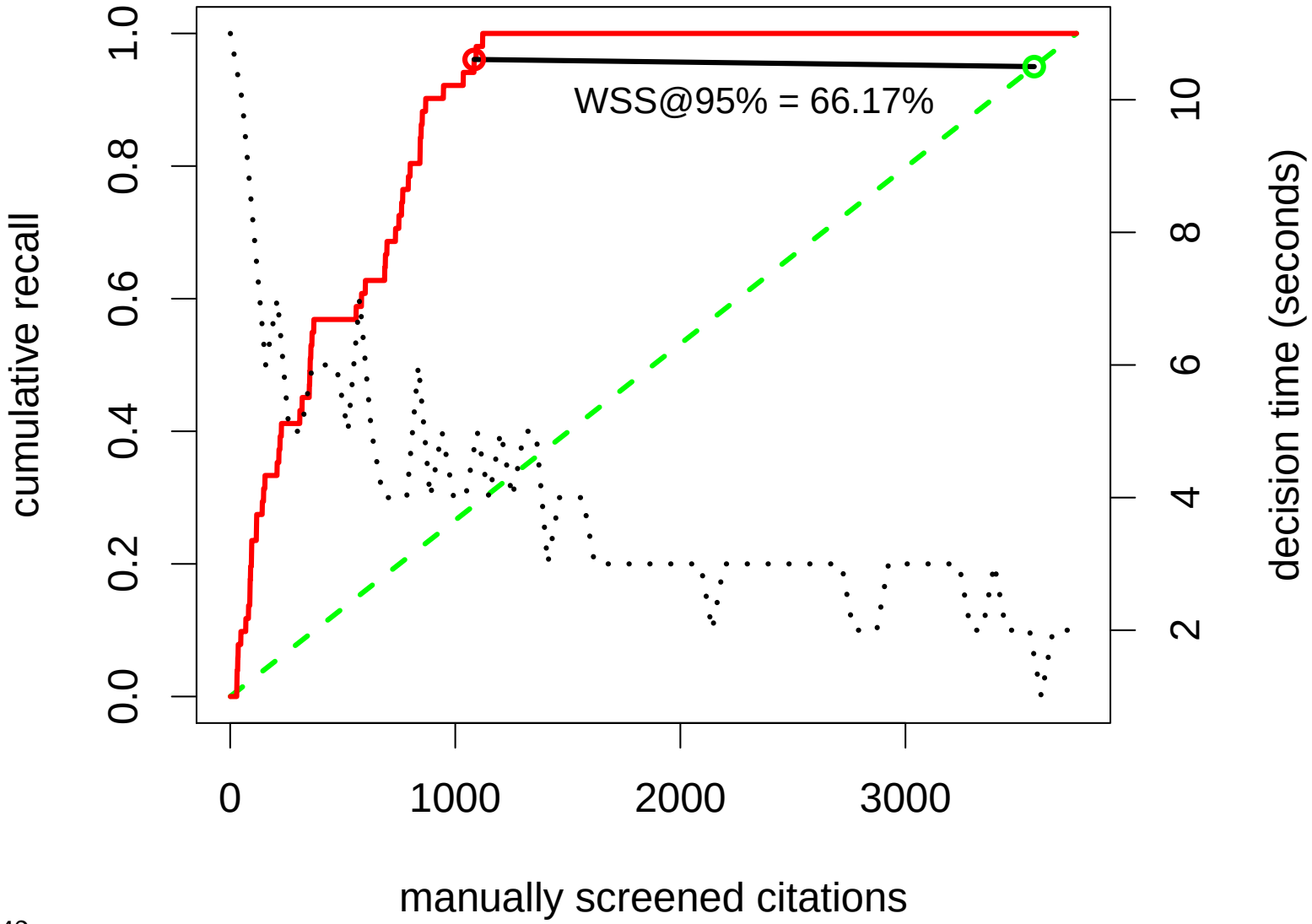
Wyniki

Kolekcja	Analityk	WSS@95%	Uczenie aktywne
BC-C	senior	* 6.89%	Yes
WC-D	senior	* 29.54%	Yes
WC-C	senior	* 22.35%	Yes
PAP	senior	* 40.63%	Yes
WGP	senior	* 36.82%	Yes
PW-S	senior	* 63.15%	Yes
PW-E	senior	* 38.81%	Yes
WM	senior	* 23.72%	Yes
SH	senior	* 66.17%	Yes
QSH	junior	* 39.84%	Yes
QSH	senior	* 31.32%	Yes
LD	senior	* 50.45%	Yes
OCM	senior	* 63.99%	Yes
BC-S	senior	* 9.41%	No
WC-F	senior	8.95%	No
HB	junior	-3.62%	No

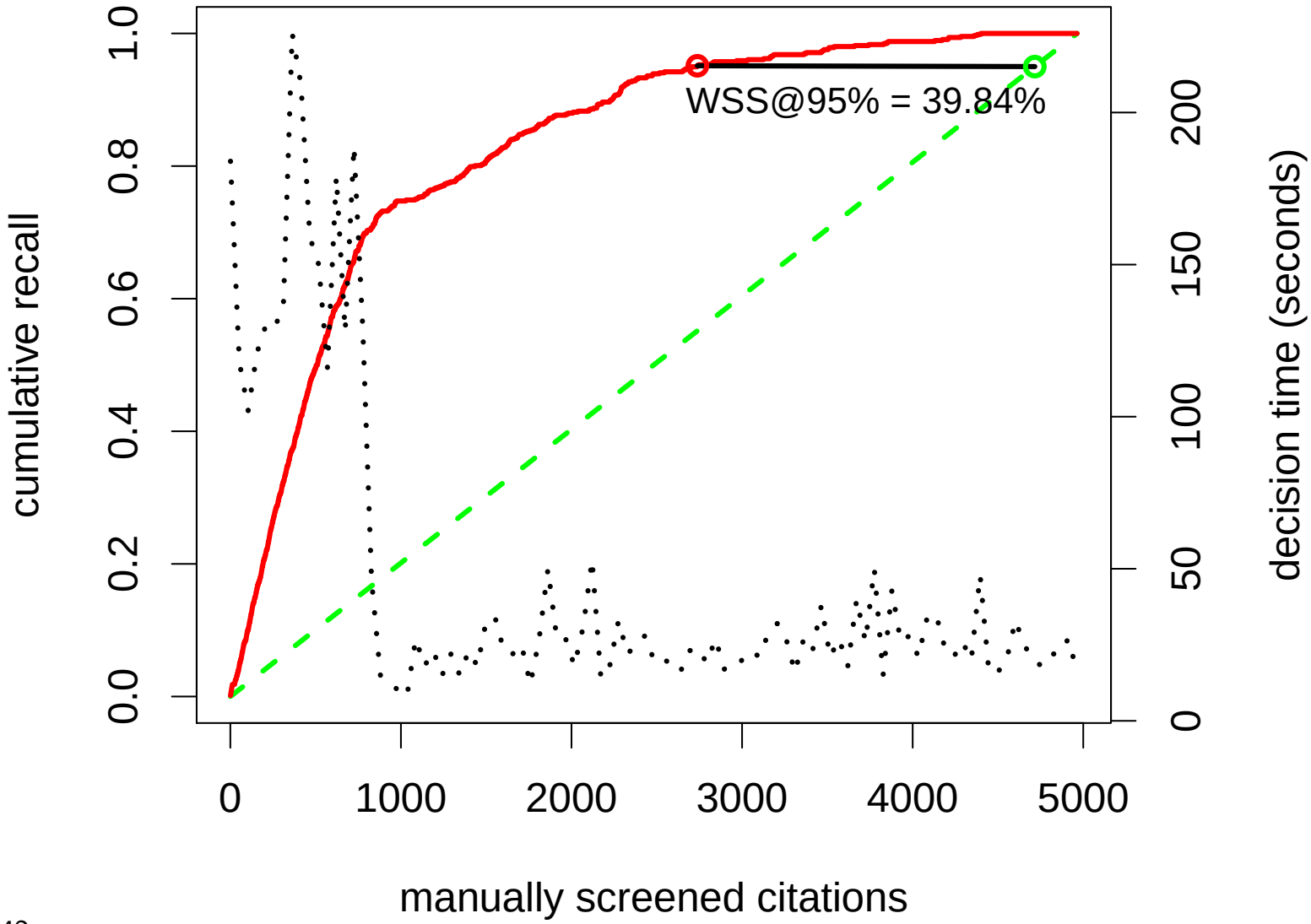
Przykład: trudny (choice architecture in BC)



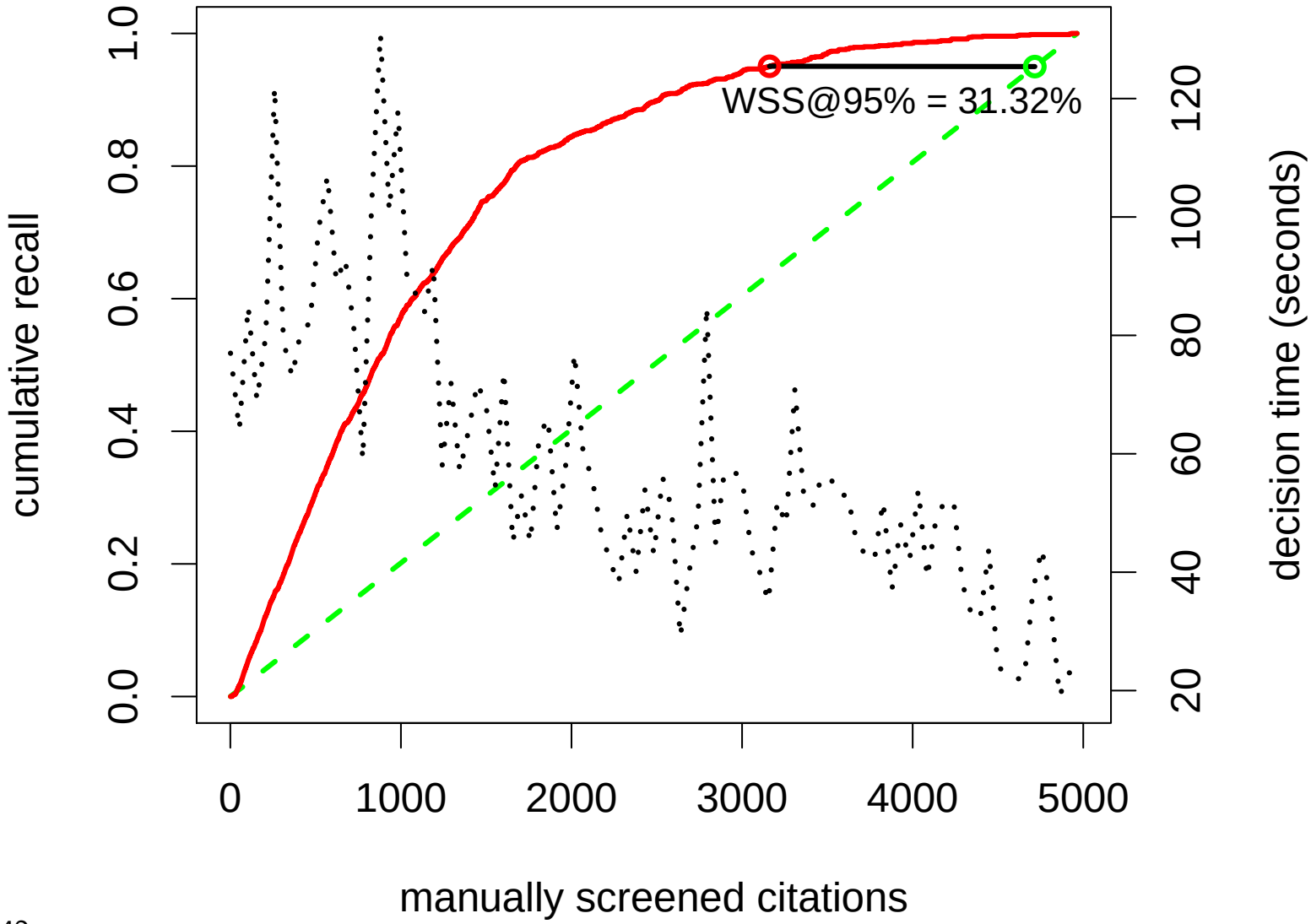
Przykład: łatwy (sexual health)



Przykład: młodszy analityk



Przykład: starszy analityk



Rozszerzenia

Grupowanie

Cel: podział zbioru dokumentów na spójne tematycznie grupy:

- grupowanie twarde (podział),
- słownictwo spójne w ramach grupy,
- cechy opisowe (słowa, wyrażenia) pozwalające zrozumieć grupę.

Samo-strojące grupowanie opisowe [Brockmeier et al., 2018]

Grupowanie

Grupowanie opisowe na k grup:

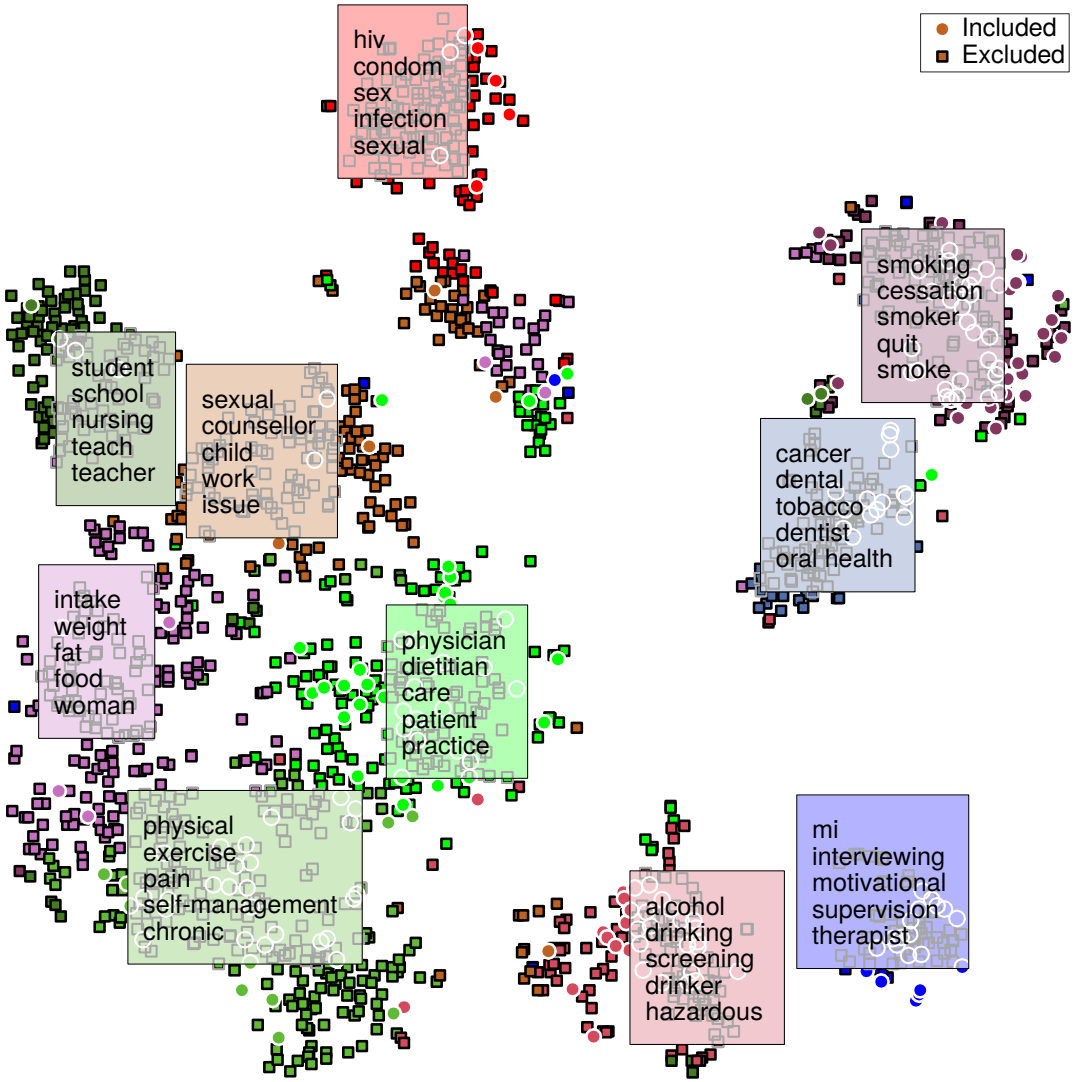
1. Budowa macierzy $A_{N\times N}$ podobieństwa cosinusowego wektorów TF/IDF (implicite),
2. Grupowanie spektralne w przestrzeni k wektorów własnych laplasjanu $L = D^{-1/2}AD^{-1/2}$ z użyciem k-means lub innego algorytmu,
3. Wybór etykiet opisujących (słów lub wyrażeń wielowyrazowych) grupę jako cech ją przewidujących z użyciem kryterium CMIM (conditional mutual information maximisation) i wybór liczby cech przez BIC (Bayesian information criterion).

BC: grupy i opisy

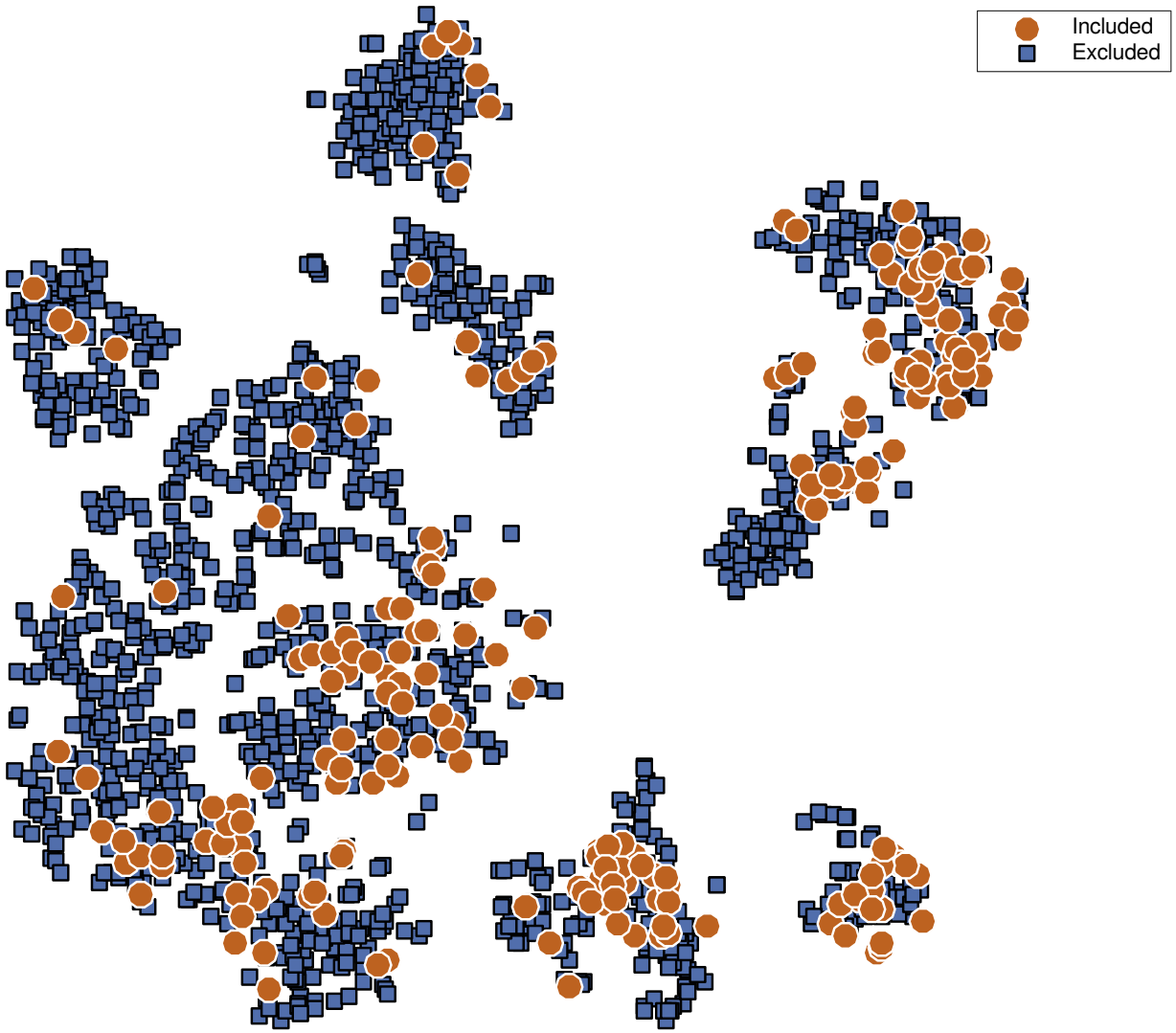
Behaviour change (10 grup z 1667 dokumentów):

- physical, pain, activity, trial, diabete (250)
- physician, dietitian, contraception, care (224)
- woman, fat, food, weight, intake, dietary (222)
- sexual, work, **young people**, child (194)
- **smoking cessation**, abstinence, nicotine (174)
- hiv, condom, testing, sex, prevention, hiv/aids (147)
- alcohol, drinking, screening, **brief intervention** (144)
- student, school, nursing, teach, performance (134)
- cancer, dental, tobacco, dentist, **tobacco use** (91)
- **motivational interviewing**, motivational (87)

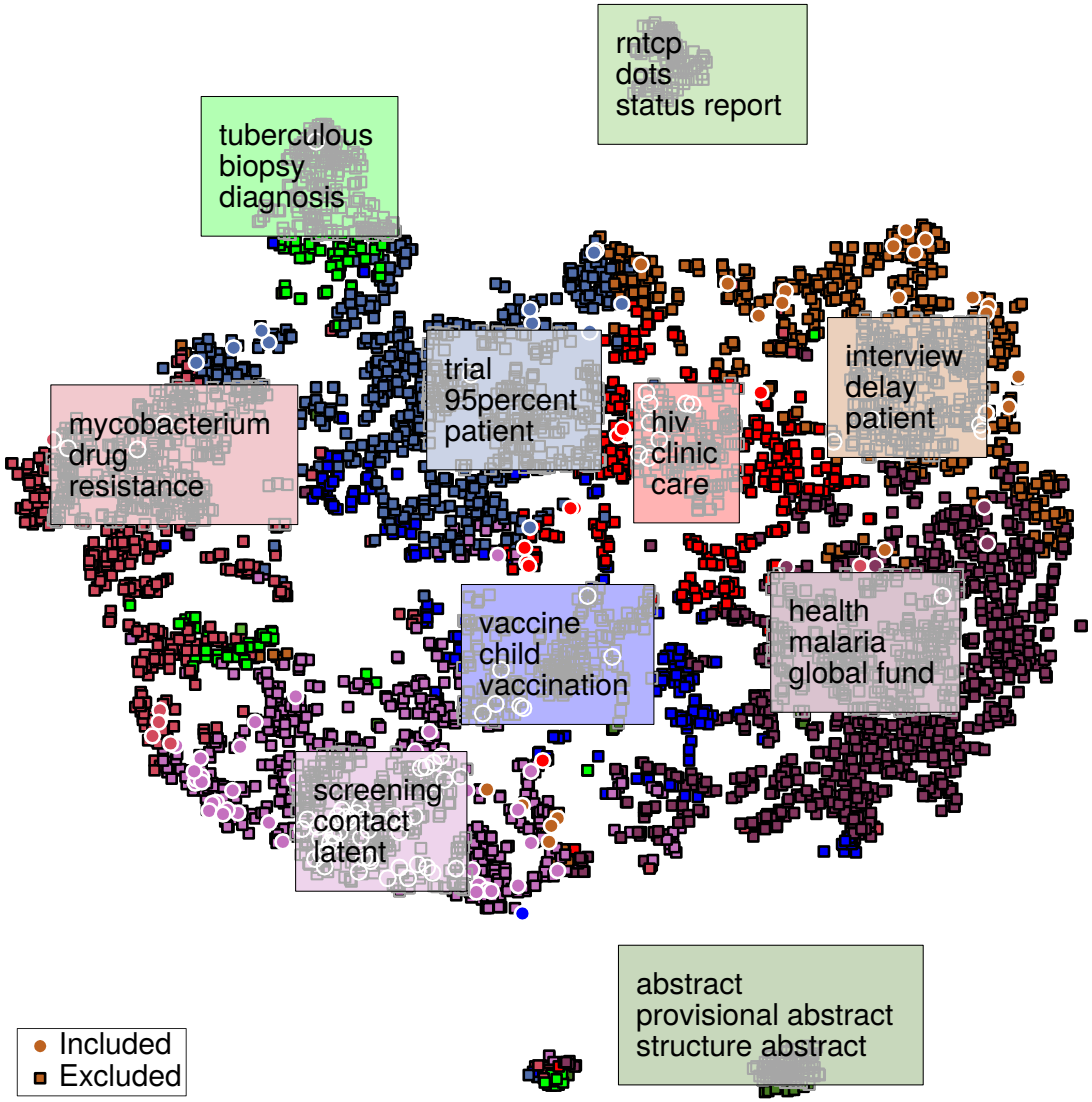
BC: wizualizacja t-SNE



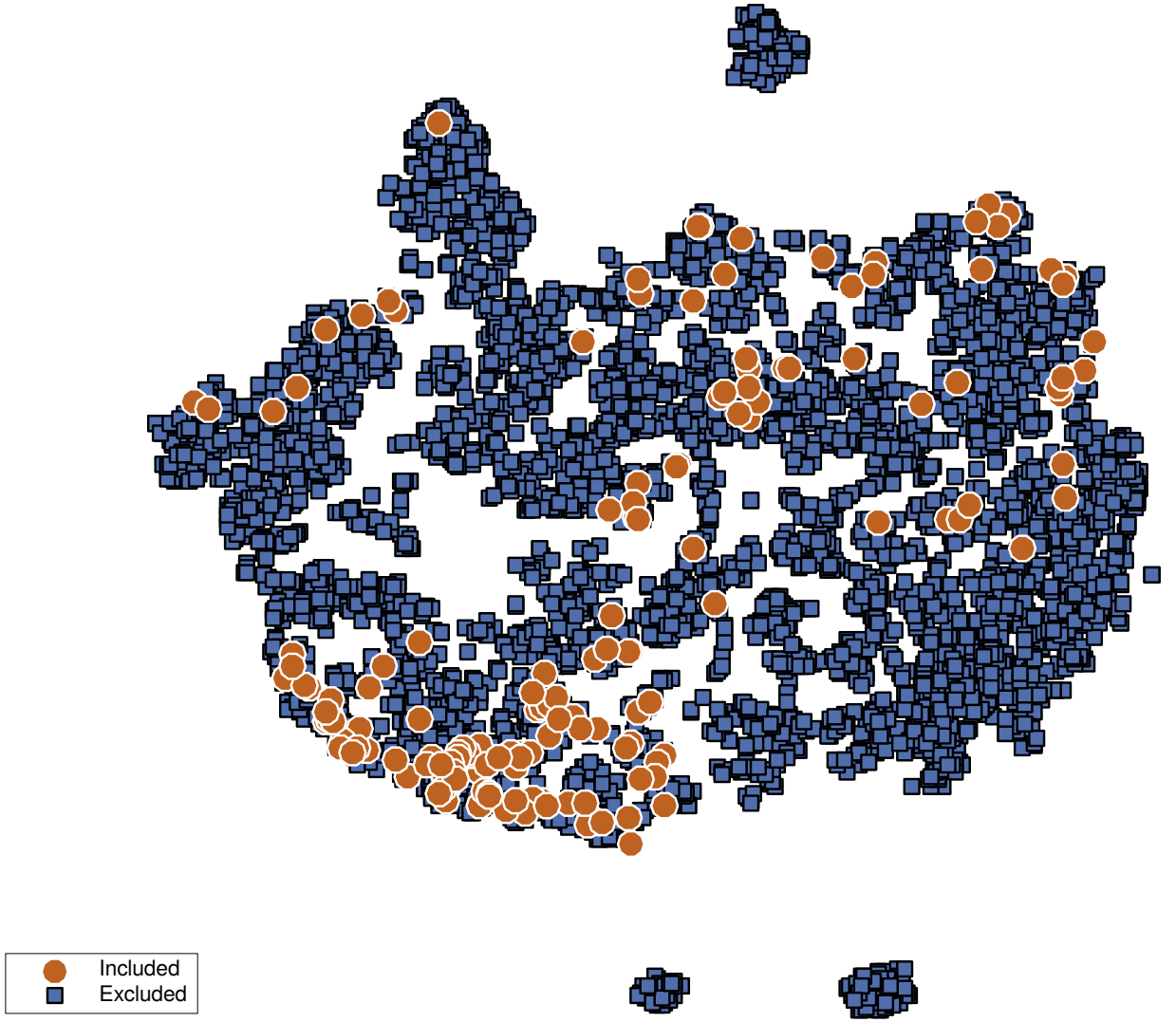
BC: włączenia



Tuberculosis: wizualizacja t-SNE



Tuberculosis: włączenia



Przyspieszanie uczenia

Początek przesiewania:

- Mało danych: AL dostarcza kilkadziesiąt przypadków, w tym kilka pozytywnych,
- Wydajność istotna: *rapid reviews*, orientacja w kolekcji, uniknięcie tendencyjności.

Hipoteza:

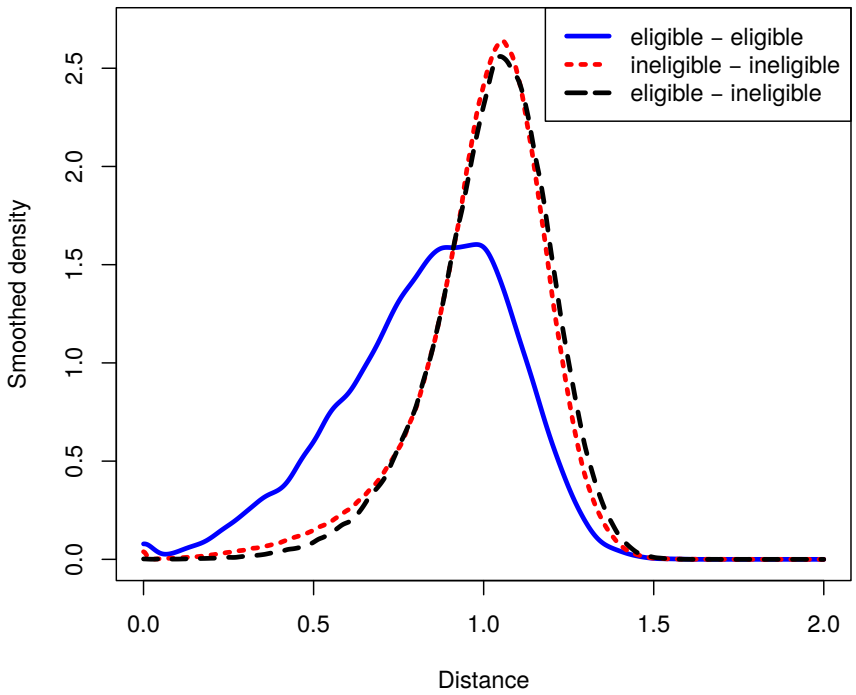
- W *dobrej* przestrzeni cechy etykiety są zachowane w bliskim sąsiedztwie.

Rozwiązanie – pół-nadzorowane uczenie z propagacją etykiet [Kontonatsios et al., 2017]

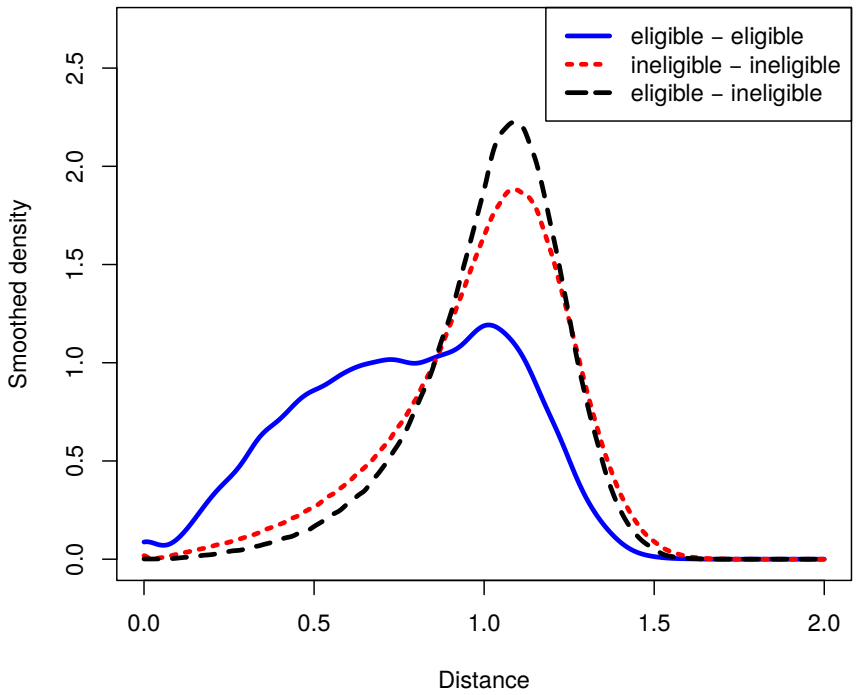
- Dla każdego przypadku z etykietą, przypisujemy ją do jego k najbliższych sąsiadów.

Hipoteza

COPD



Tobacco packaging



Przestrzeń sąsiedztwa

BoW lub *spectral embeddings* (osadzenia??).
Przekształcenie macierzy TF-IDF X:

1. normalizacja: $R_{d,w} = \sqrt{X_{d,w}} / \sqrt{\sum_{w'} X_{d,w'}}$,
2. macierz podobieństw: $C_{d,d'} = \langle \mathbf{r}_d, \mathbf{r}_{d'} \rangle \quad 0 \leqslant C_{d,d'} \leqslant 1$
3. symetryczna normalizacja: $\tilde{C} = D^{-1/2} C D^{-1/2}, D_{d,d} = (\sum_{d'} C_{d,d'})$
4. rozkład na wartości własne $\lambda_1 \geqslant \cdots \geqslant \lambda_p$ daje:

$$Z = [\mathbf{u}_1 \sqrt{\lambda_1}, \ldots, \mathbf{u}_p \sqrt{\lambda_p}] = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n]^{\mathbf{T}}$$

Otrzymujemy $\mathbf{z}_d$: p-wymiarowy wektor osadzenia dokumentu d.

Wyniki

Kolekcja	Metoda	Stopień ukończenia kolekcji			
		5%	10%	25%	100%
		$U_{cert.}$ / $U_{uncert.}$	$U_{cert.}$ / $U_{uncert.}$	$U_{cert.}$ / $U_{uncert.}$	$U_{cert.}$ / $U_{uncert.}$
COPD	AL	60.92 / 73.71	64.37 / 80.63	78.88 / 90.10	92.35 / 95.14
	SemiBow	65.33 / 77.26	75.41 / 80.57	86.19 / 89.36	94.19 / 94.91
	SemiSpectral	65.30 / **88.71**	74.35 / **86.87**	85.65 / **91.93**	94.06 / **95.55**
Proton Beam	AL	47.57 / 79.23	62.57 / 88.31	82.65 / 94.39	93.33 / 96.21
	SemiBow	50.68 / **79.61**	68.34 / 88.90	84.94 / 94.72	93.84 / 96.31
	SemiSpectral	53.65 / 79.57	70.49 / **89.08**	86.03 / **94.87**	94.12 / **96.37**
Cooking Skills	AL	46.66 / 47.59	59.40 / 60.56	75.17 / 73.69	89.68 / 88.59
	SemiBow	56.26 / 57.13	68.05 / 66.11	80.75 / 76.77	91.43 / 89.07
	SemiSpectral	**60.71** / 53.17	**70.65** / 64.98	**81.96** / 76.24	**91.78** / 88.98
Sanitation	AL	24.44 / **25.61**	32.10 / 32.23	52.09 / 48.49	82.49 / 80.63
	SemiBow	24.27 / 24.68	35.37 / 32.82	54.54 / 48.36	83.18 / 80.59
	SemiSpectral	24.37 / 17.30	**37.71** / 31.52	**57.38** / 54.03	**83.95** / 82.09
Tobacco Pack.	AL	45.70 / 43.48	53.96 / 55.79	75.35 / 72.70	90.85 / 90.06
	SemiBow	50.27 / 55.61	61.92 / 62.79	78.66 / 75.78	91.68 / 90.56
	SemiSpectral	54.70 / **60.78**	63.98 / **68.27**	**79.24** / 78.75	**91.84** / 91.42
Youth Dev.	AL	22.71 / 28.09	31.34 / 36.52	51.97 / 56.34	82.81 / 83.66
	SemiBow	32.61 / 41.43	42.62 / 46.48	61.77 / 60.02	85.69 / 84.43
	SemiSpectral	36.40 / **49.91**	44.13 / **53.73**	62.29 / **64.56**	**85.94** / 85.83

Analiza błędów

Przeglądy systematyczne badań neurologicznych *in vivo* (na zwierzętach): CAMARADES, uniwersytet w Edynburgu. Scenariusz:

- Istniejące poprzednie przeglądy ze znanymi decyzjami,
- Budowa modelu i stosowanie do nowych publikacji dla prioretyzacji.

Błędne decyzje człowieka:

- Niesłuszne włączenie: korekta w następnych etapach,
- Niesłuszne wyłączenie: brak korekty,
- Redukowane przez równoległe przesiewanie i uzgadnianie konfliktów (£££).

Czy można zastosować uczenie maszynowe do wykrywania błędów?

Metoda

Walidacja krzyżowa z sortowanie błędów [Bannach-Brown et al., 2018]:

- Budowa modelu regresji logistycznej z regularyzacją L1 (LASSO) na binarnej macierzy wystąpień słów przekształconej przez SVD (300 wymiarów) i LDA (300 tematów).
- Przypisanie prawdopodobieństw włączenia $\hat{y}_i$ do wszystkich przypadków w walidacji krzyżowej z $k = 5$,
- Sortowanie kolekcji według niezgodności z decyzją człowieka $e_i = y_i - \hat{y}_i$ i przegląd z obu końców,
- Koniec po napotkaniu 5 przypadków poprawnej decyzji pod rząd.

Wyniki

- Ponowny przegląd 85 rekordów,
- Zmiana decyzji dla 55 rekordów,
- Brakujące włączenia – 44 (4,77% całej kolekcji),
- Błędne włączenia – 11, np.

The kinetics of inhibition of hepatic drug metabolism by prostaglandins in **rabbits** The in vitro effects of prostaglandin A1 and prostaglandin E1 on the hepatic microsomal metabolism of type I (aminopyrine) and type II (p-chloro-N-methylaniline and aniline) drug substrates in **rabbits** were investigated. Both prostaglandins competitively inhibited aminopyrine N-demethylation to the same extent with a 500 microM prostaglandin concentration decreasing metabolism by approximately 50 percent. Prostaglandin A1 and prostaglandin F1 decreased p-chloro-N-methylaniline biotransformation, and prostaglandin E1 **depressed** aniline metabolims, via mixed-inhibition kinetics...

Zakończenie

Wdrożenie systemu

- RobotAnalyst pozwala na oszczędzenie 7-70% nakładu pracy w przesiewaniu literatury,
- System zaprezentowany na konferencjach dot. przyglądów systematycznych w formie warsztatów [Kontonatsios et al., 2016, Nolan et al., 2017b] i raportów z ewaluacji [Nolan et al., 2017a, Le Pogam et al., 2017b, Le Pogam et al., 2017a],
- Aktualnie wykorzystywany przez naukowców z $\sim$30 jednostek ze świata,
- Nagroda *Next Big Thing* (£10,000) dla badań z UoM o potencjale komercyjnym,
- Więcej: `http://nactem.ac.uk/robotanalyst/`

Otwarte pytania

- Skąd wiadomo, że osiągnięto zakładaną czułość i można zakończyć przegląd?
- Jak zastosować bardziej zaawansowaną analizę tekstu dla usprawnienia klasyfikacji?
- Jak zagwarantować wymaganą jakość procesu?
- Czy takie same korzyści można osiągnąć w innych zadaniach filtrowania dokumentów? (TAK)
- Czy można przyspieszyć proces przez ręczne wybieranie początkowych dokumentów? (TAK)
- Czy korzystanie z systemu priorytezacji może obniżać czujność? (TAK)
- Czy można zautomatyzować również ekstrakcję informacji z pełnego dokumentu? (MOŻE)

Literatura (1)

Bannach-Brown, A., Przybyła, P., Thomas, J., Rice, A. S., Ananiadou, S., Liao, J., and Macleod, M. R. (2018).
The use of text-mining and machine learning algorithms in systematic reviews: reducing workload in preclinical biomedical sciences and reducing human screening error.
bioRxiv, page 255760.

Brockmeier, A. J., Mu, T., Ananiadou, S., and Goulermas, J. Y. (2018).
Self-Tuned Descriptive Document Clustering using a Predictive Network.
IEEE Transactions on Knowledge and Data Engineering.

Kontonatsios, G., Batista-Navarro, R., Przybyła, P., and Ananiadou, S. (2016).
Text mining methods to support the development of sensitive search strategies in public health reviews.
In *Cochrane Colloquium*, Seoul, South Korea.

Kontonatsios, G., Brockmeier, A. J., Przybyła, P., McNaught, J., Mu, T., Goulermas, J. Y., and Ananiadou, S. (2017).
A semi-supervised approach using label propagation to support citation screening.
Journal of Biomedical Informatics, 72:67–76.

Le Pogam, M. A., Przybyła, P., Ojeda-Ruiz, E., Bacher, S., Ananiadou, S., and von Elm, E. (2017a).
Improving efficiency of reference screening in systematic literature reviews using the RobotAnalyst text mining application: performance assessment in a systematic review on patient safety.
In *Swiss Public Health Conference*, Basel, Switzerland.

Literatura (2)

Le Pogam, M. A., Przybyła, P., Ojeda-Ruiz, E., Bacher, S., Ananiadou, S., and von Elm, E. (2017b).

Using the RobotAnalyst text-mining application to boost efficiency of literature screening: experience from a systematic review in health services research.

In *Global Evidence Summit*, Cape Town, South Africa.

Miwa, M., Thomas, J., O'Mara-Eves, A., and Ananiadou, S. (2014).

Reducing systematic review workload through certainty-based screening.

Journal of Biomedical Informatics, 51:242–253.

Nolan, K., Ananiadou, S., Le Pogam, M.-A., von Elm, E., and Przybyła, P. (2017a).

Screening evidence for systematic reviews using a text-mining system: The RobotAnalyst.

In *Global Evidence Summit*, Cape Town, South Africa.

Nolan, K., Ananiadou, S., Przybyła, P., and Brockmeier, A. J. (2017b).

RobotAnalyst: An online system to support citation screening in evidence reviewing.

In *Global Evidence Summit*, Cape Town, South Africa.

Dziękuję za uwagę!

Zespół NaCTeM:

- Sophia Ananiadou,
- Austin J. Brockmeier,
- Georgios Kontonatsios,
- John McNaught,
- Antonis Antoniou,
- . . . (rekrutujemy!)

Analitycy:

- Kay Nolan,
- Marie-Annick Le Pogam,
- Erik von Elm,
- Alexandra Bannach-Brown,
- i inni.