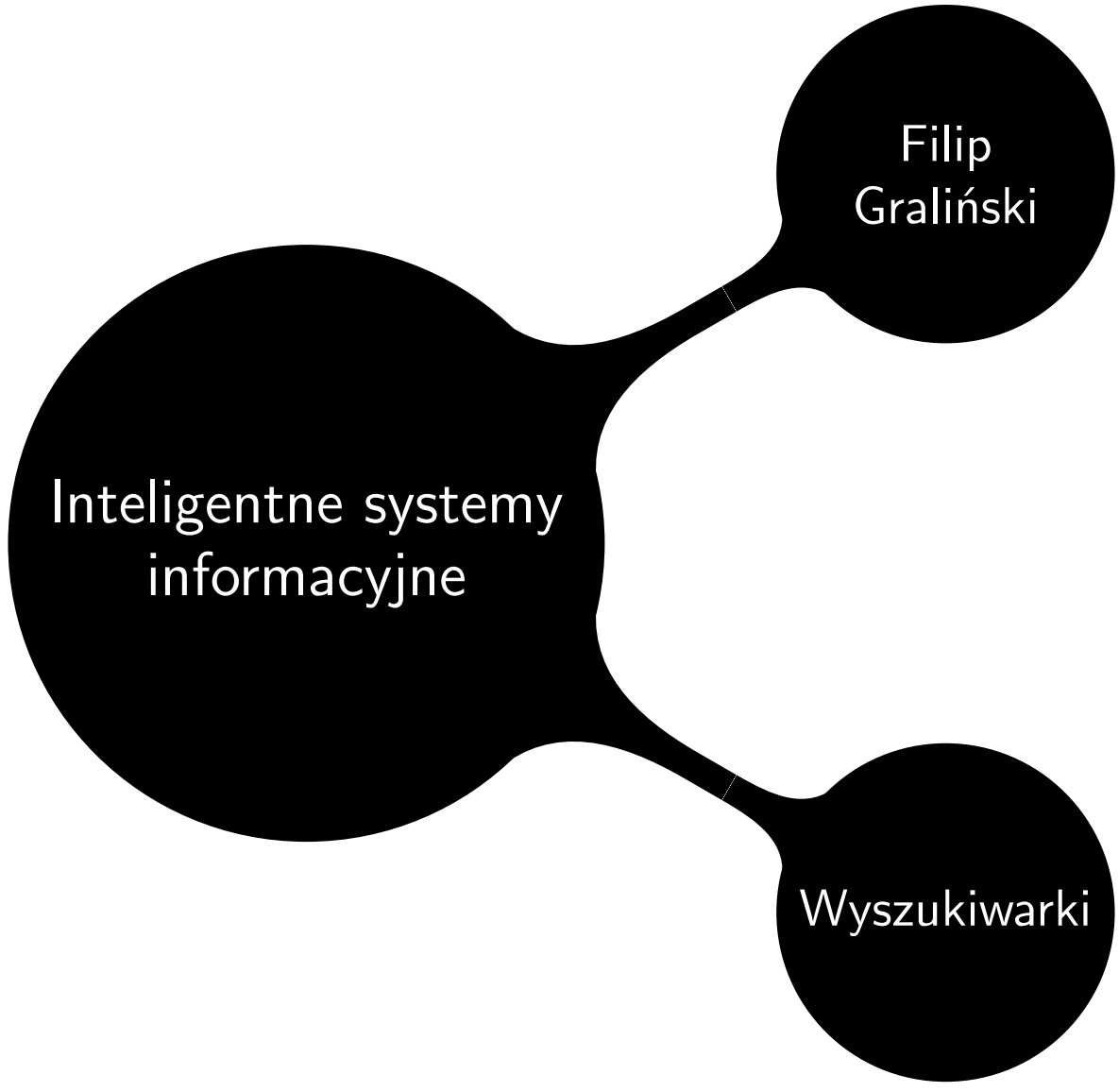
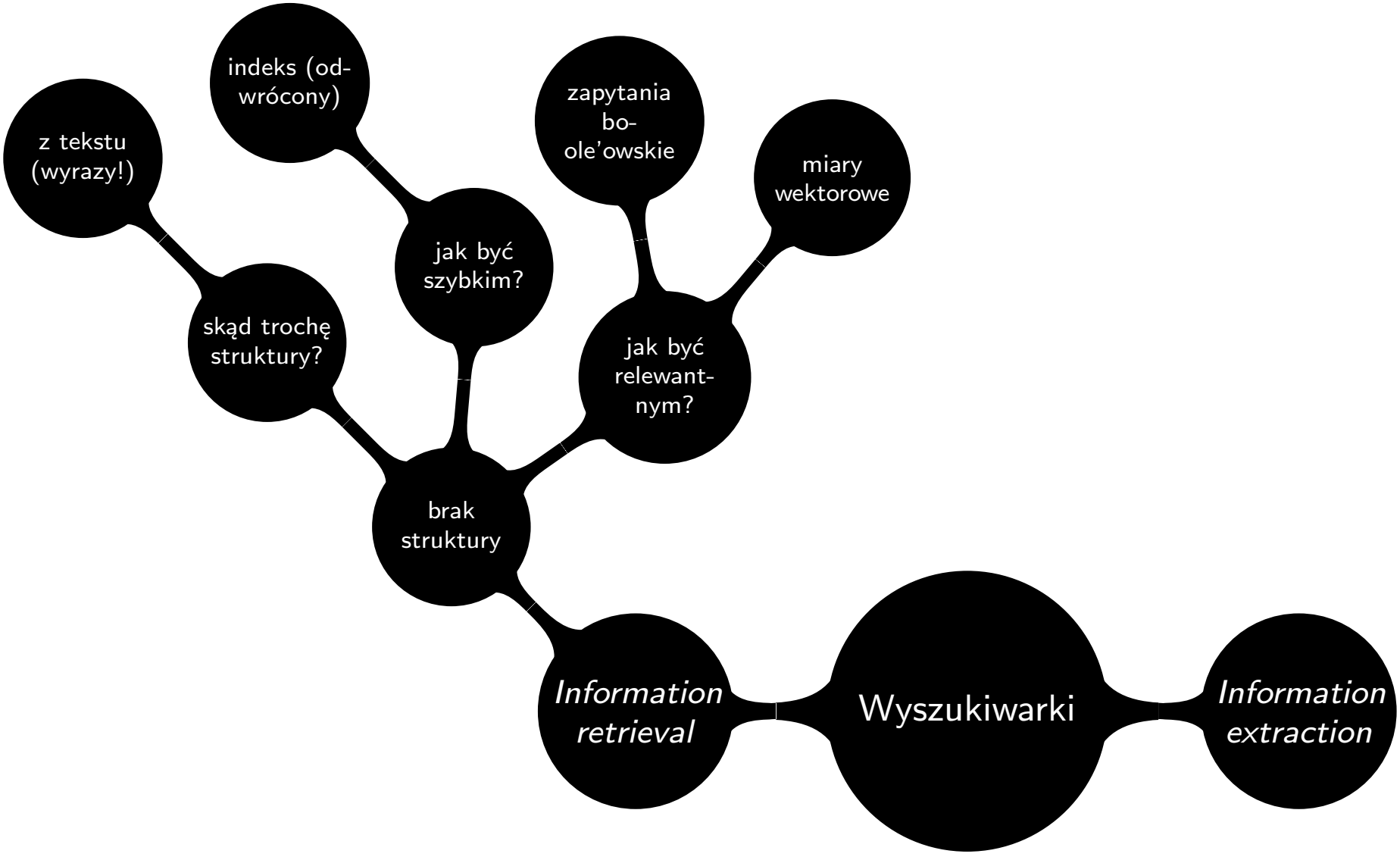

Filip
Graliński
Inteligentne systemy
informacyjne
Wyszukiwarki


Wyszukiwarki
Information retrieval
Information extraction
brak struktury
skąd trochę struktury?
jak być szybkim?
jak być relewant-nym?
z tekstu (wyrazy!)
indeks (od-wrócony)
zapytania bo-ole'owskie
miary wektorowe

dokument jako wektor

doc1	*Ala ma kota.*
doc2	*Podobno jest kot w butach.*
doc3	*Ty chyba masz kota!*
doc4	*But chyba zgubiłem.*

dokument jako wektor

doc1	*Ala ma kota.*
doc2	*Podobno jest kot w butach.*
doc3	*Ty chyba masz kota!*
doc4	*But chyba zgubiłem.*

	ala	*but*	*chyba*	*kot*	*mieć*	*podobno*	*ty*	*zgubić*
doc1	1	0	0	1	1	0	0	0
doc2	0	1	0	1	0	1	0	0
doc3	0	0	1	1	1	0	1	0
doc4	0	1	1	0	0	0	0	1

dokument jako wektor

doc1	*Ala ma kota.*
doc2	*Podobno jest kot w butach.*
doc3	*Ty chyba masz kota!*
doc4	*But chyba zgubiłem.*
doc5	*Kot ma kota.*

	ala	*but*	*chyba*	*kot*	*mieć*	*podobno*	*ty*	*zgubić*
doc1	1	0	0	1	1	0	0	0
doc2	0	1	0	1	0	1	0	0
doc3	0	0	1	1	1	0	1	0
doc4	0	1	1	0	0	0	0	1

dokument jako wektor

doc1	*Ala ma kota.*
doc2	*Podobno jest kot w butach.*
doc3	*Ty chyba masz kota!*
doc4	*But chyba zgubiłem.*
doc5	*Kot ma kota.*

	ala	*but*	*chyba*	*kot*	*mieć*	*podobno*	*ty*	*zgubić*
doc1	1	0	0	1	1	0	0	0
doc2	0	1	0	1	0	1	0	0
doc3	0	0	1	1	1	0	1	0
doc4	0	1	1	0	0	0	0	1
doc5	0	0	0	?	1	0	0	0

jak uwzględnić frekwencję wyrazu?

$$\mathrm{tf}_{t,d}$$

jak uwzględnić frekwencję wyrazu?

$$\mathrm{tf}_{t,d}$$

$$1 + \log(\mathrm{tf}_{t,d})$$

jak uwzględnić frekwencję wyrazu?

$$\mathsf{tf}_{t,d}$$

$$1 + \log(\mathsf{tf}_{t,d})$$

$$\begin{cases} 1, & \text{jeśli } \mathsf{tf}_{t,d} > 0 \\ 0, & \text{w przeciwnym razie} \end{cases}$$

jak uwzględnić frekwencję wyrazu?

$$\mathrm{tf}_{t,d}$$

$$1 + \log(\mathrm{tf}_{t,d})$$

$$\begin{cases} 1, & \text{jeśli } \mathrm{tf}_{t,d} > 0 \\ 0, & \text{w przeciwnym razie} \end{cases}$$

$$0.5 + \frac{0.5 \times tf_{t,d}}{max_t(\mathrm{tf}_{t,d})}$$

odwrotna częstość w dokumentach

Czy wszystkie wyrazy są tak samo ważne?

odwrotna częstość w dokumentach

Czy wszystkie wyrazy są tak samo ważne?

NIE. Wyrazy pojawiające się w wielu dokumentach są mniej ważne.

odwrotna częstość w dokumentach

Czy wszystkie wyrazy są tak samo ważne?

NIE. Wyrazy pojawiające się w wielu dokumentach są mniej ważne.

Aby to uwzględnić, przemnażamy frekwencję wyrazu przez odwrotną częstość w dokumentach (*inverse document frequency*):

$$\mathrm{idf}_t = \log \frac{N}{\mathrm{df}_t}$$

,

idf_t – odwrotna częstość wyrazu t w dokumentach

N – liczba dokumentów w kolekcji

df_f – w ilu dokumentach wystąpił wyraz t?

odwrotna częstość w dokumentach (cd.)

wyraz wystąpił...	$\text{idf}_t = \ldots$
w 1 dokumencie	

odwrotna częstość w dokumentach (cd.)

wyraz wystąpił...	$\mathrm{idf}_t = \ldots$
w 1 dokumencie	$= \log N/1 = \log N$

odwrotna częstość w dokumentach (cd.)

wyraz wystąpił...	$\text{idf}_t = \ldots$
w 1 dokumencie	$= \log N/1 = \log N$
2 razy w kolekcji	$= \log N/2$ lub $\log N$

odwrotna częstość w dokumentach (cd.)

wyraz wystąpił...	$\mathrm{idf}_t = \ldots$
w 1 dokumencie	$= \log N/1 = \log N$
2 razy w kolekcji	$= \log N/2$ lub $\log N$
w połowie dokumentów	$= \log N/(N/2) = \log 2$

odwrotna częstość w dokumentach (cd.)

wyraz wystąpił...	$\mathrm{idf}_t = \ldots$
w 1 dokumencie	$= \log N/1 = \log N$
2 razy w kolekcji	$= \log N/2$ lub $\log N$
w połowie dokumentów	$= \log N/(N/2) = \log 2$
we wszystkich dokumentach	$= \log N/N = \log 1 = 0$

odwrotna częstość w dokumentach (cd.)

wyraz wystąpił...	$\text{idf}_t = \ldots$
w 1 dokumencie	$= \log N/1 = \log N$
2 razy w kolekcji	$= \log N/2$ lub $\log N$
w połowie dokumentów	$= \log N/(N/2) = \log 2$
we wszystkich dokumentach	$= \log N/N = \log 1 = 0$

Zamiast $\text{tf}_{t,d}$ będziemy w wektorach rozpatrywać wartości:

$$\text{tf-idf}_{t,d} = \text{tf}_{t,d} \times \text{idf}_t$$

odwrotna częstość w dokumentach (cd.)

wyraz wystąpił...	$\text{idf}_t = \ldots$
w 1 dokumencie	$= \log N/1 = \log N$
2 razy w kolekcji	$= \log N/2$ lub $\log N$
w połowie dokumentów	$= \log N/(N/2) = \log 2$
we wszystkich dokumentach	$= \log N/N = \log 1 = 0$

Zamiast $\text{tf}_{t,d}$ będziemy w wektorach rozpatrywać wartości:

$$\text{tf-idf}_{t,d} = \text{tf}_{t,d} \times \text{idf}_t$$

Overlap score measure:

$$\text{Score}(q, d) = \sum_{t \in q} \text{tf-idf}_{t,d}$$

Opowieść A

Dziewczyna w akademiku kupila weza, z czasem podrosl. Zrzucone skory na scianach, rysunki, wycinki, waz slizgal sie luzem, sypal w lozku, itd. w pewnym momencie waz przestal jesc myszy, rozlozyl sie i zaczal "nadymac". Telefony do sklepu, weterynarza i kogostam jeszcze nie daly skutku i nikt nie potrafil doradzic dziewczynie jak to ogarnac. Waz lezal i sie "nadymal", a dziewczyna zamartwiala sie przekonana, ze choruje. Zadzwonila przypadkiem do zoo, opowiedziala cala sytuacje i uslyszala: Prosze wyjsc z pokoju i zamknac drzwi - waz jest zdrowy, a na lozku lezy i sie "nadyma" bo sprawdza czy jest w stanie pania zmiescic...

Opowieść B

Anegdotka o wężu Historia wydarzyła się koleżance. Dziewczyna hodowała pytona (pytony nie maja kłów jadowych, czy jak tam to się nazywa) . Wąż był już dość spory i nie trzymała go w żadnym specjalnym terrarium tylko luzem w mieszkaniu. Wąż snuł się w po domu, wylegiwał na meblach i parapetach. Ale pewnego dnia przestał jeść. Po kilku dniach jego głodówki zaniepokojona właścicielka zabrała go do weterynarza. Ten obejrzał i powiedział, że wygląda na zdrowego. Koleżanka zabrała go do chaty, ale sytuacja jeszcze przez kilka dni się utrzymywała - dalej nie jadł. Pewnej nocy obudziła się i zobaczyła go obok siebie na łóżku swojego węża wyciągniętego wzdłuż niej jak kij od miotły. W pierwszej chwili przestraszyła się, że zdechł, ale okazało się, że żyje. Następnego dnia zabrała go znowu do weterynarza i opowiedziała, ze dalej nie je, a poza tym wspomniała o dziwnej sytuacji w nocy.Dzięki wyjaśnieniom weterynarza okazało się, że pytony właśnie w taki sposób mierzą czy są wystarczająco długie, żeby połknąć swoją ofiarę. :) Łatwo się domyśleć, dlaczego przestał jeść...

Opowieść C

Smutne. Przypomina mi to historię z bałwankami. Był sobie wredny facet, który rozjeżdżał dzieciom samochodem bałwanki. Co dzieci ulepiły to on rozjeżdżał. Pewnego razu dzieci ulepiły ślicznego bałwanka na hydrancie. Duży im wyszedł ten bałwanek. I Pan Wredotek z impetem spróbował go rozjechać. Ja nie jestem tak subtelny, jak te dzieci. Ja nakarmiłbym człowieka tym, czym on nakarmił psa. I na pewno nie skracałbym mu cierpienia.

Opowieść D

Wczoraj w popołudniowej audycji w trójce prowadzący przeczytał maila od słuchaczki z Poznania: ...na pewnym osiedlu dzieci ulepiły bałwana - ktoś z premedytacją go rozjechał samochodem, następnego dnia ulepiły go jeszcze raz i znów go ktoś rozjechał, trzeciego dnia ulepiły go na hydrancie.... Nie wiem, czy to prawda, ale chciałbym zobaczyć minę kierowcy

podobieństwo cosinusowe

wektor dokumentu ($\vec{V}(d)$) wektor, którego składowe odpowiadają wyrazom

podobieństwo cosinusowe

wektor dokumentu ($\vec{V}(d)$) wektor, którego składowe odpowiadają wyrazom

podobieństwo cosinusowe ($\text{sim}(d_1, d_2)$) cosinus kąta między wektorami dokumentów:

$$\text{sim}(d_1, d_2) = \frac{\vec{V}(d_1) \cdot \vec{V}(d_2)}{|\vec{V}(d_1)||\vec{V}(d_2)|}$$

podobieństwo cosinusowe

wektor dokumentu ($\vec{V}(d)$) wektor, którego składowe odpowiadają wyrazom

podobieństwo cosinusowe ($\text{sim}(d_1, d_2)$) cosinus kąta między wektorami dokumentów:

$$\text{sim}(d_1, d_2) = \frac{\vec{V}(d_1) \cdot \vec{V}(d_2)}{|\vec{V}(d_1)||\vec{V}(d_2)|}$$

Traktując zapytanie q jako dokument otrzymamy...

Cosine score:

$$\text{Score}(q, d) = \frac{\vec{V}(q) \cdot \vec{V}(d)}{|\vec{V}(q)||\vec{V}(d)|}$$

modele języka

Model języka przypisuje prawdopodobieństwa napisom.

Jeśli M ma być modelem języka polskiego, oczekiwalibyśmy, że dla napisów:

- z_1 *W tym stanie rzeczy pan Ignacy coraz częściej myślał o Wokulskim.*
- z_2 *Po wypełniony zbiornik pełny i należne kwotę, usłyszała w attendant*
- z_3 *xxxxyźźźźźit backspace hooooooooooop x y z*

zachodzić będzie:

$$P(z_1|M) > P(z_2|M) > P(z_3|M)$$

modele języka w wyszukiwaniu informacji

Na podstawie każdego dokumentu d tworzymy model języka M_d.

ocena oparta na modelach unigramowych

$$\text{Score}(q, d) = P(q|M_d) = K_q \prod_{t \in q} P(t|M_d)^{\text{tf}_{t,d}}$$

. . . a tak naprawdę

1. Przyjmujemy sensowne założenia
2. . . .
3. (Okapi) BM25

$$\mathrm{BM25}(q,d) = \sum_{t \in q}^{n} \mathrm{idf}_t \cdot \frac{\mathrm{tf}_{t,d}}{\mathrm{tf}_{t,d} + k_1 \cdot \left((1-b) + b \cdot \frac{|d|}{\mathrm{avgdl}}\right)}$$

- $\mathrm{idf}_t = \log \frac{N - \mathrm{df}_t + 0.5}{\mathrm{df}_t + 0.5}$ (nieco inna definicja niż wyżej)
- $|d|$ – długość dokumentu d
- avgdl – średnia długość dokumentu
- k_1, b – parametry (ręcznie dostrajane)