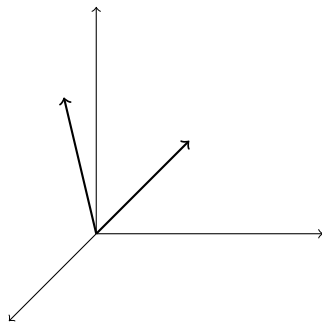


Wyszukiwanie dokumentów (ang. *document retrieval, text retrieval*) polega na poszukiwaniu dokumentów tekstowych z pewnego zbioru, które pasują do zapytania. **Wyszukiwanie informacji** (ang. *information retrieval*) jest bardziej ogólnym pojęciem – dokumenty mogą być multimedialne, wyszukiwanie dotyczy także metadanych itd.

Wyłuskiwanie informacji (ang. *information extraction*) polega na wypełnianiu formularzy z polami typu kto, gdzie, kiedy, jak, za ile itp. na podstawie informacji w zapisanych w języku naturalnym w dokumentach. Wyszukiwanie dokumentów często poprzedza wyszukiwanie informacji, a także odpowiadanie na pytania. Najpierw chcemy wiedzieć, gdzie jest informacja, potem ją analizujemy.

Model przestrzeni wektorowej

Jest to przykład bardziej ogólnego modelu **worka słów** (ang. *bag of words*) – modelu, w którym znaczenie dokumentu wynika ze złożenia znaczeń występujących w nich słów. Kolejność słów i kontekst ich występowania nie ma znaczenia. Zbiory słów reprezentowane są przez wektory w przestrzeni wielowymiarowej – jeden wymiar to jedno słowo.



Model przestrzeni wektorowej

Dokumenty i zapytanie opisywane są przez wektory wag słów. W najprostszym podejściu elementy wektora są równe 1, gdy dane słowo występuje w dokumencie, 0 w przeciwnym przypadku.

$$\vec{d}_j = (t_{1,j}, t_{2,j}, \dots, t_{N,j})$$
$$\vec{q}_k = (t_{1,k}, t_{2,k}, \dots, t_{N,k})$$

$$\text{sim}(\vec{q}_k, \vec{d}_j) = \sum_{i=1}^N t_{i,k} \cdot t_{i,j} = \vec{q}_k \cdot \vec{d}_j$$

Nauczyciele **akademiccy** mogą otrzymać za osiągnięcia naukowe, dydaktyczne lub organizacyjne **albo** za całokształt dorobku nagrody rektora oraz nagrody ministra właściwego do spraw szkolnictwa wyższego.

0 agentura
0 akademia
1 akademicki
0 akord
0 aktywista
0 alarm
1 albo
0 alert

...

Model przestrzeni wektorowej

Wagi można ulepszyć zastępując je np. **częstością użycia** danego słowa w danym dokumencie/zapytaniu. Otrzymujemy wówczas:

$$\begin{aligned} \vec{d}_j &= (w_{1,j}, w_{2,j}, \dots, w_{N,j}) \\ \vec{q}_k &= (w_{1,k}, w_{2,k}, \dots, w_{N,k}), \end{aligned} \quad \text{sim}(\vec{q}_k, \vec{d}_j) = \sum_{i=1}^N w_{i,k} \cdot w_{i,j} = \vec{q}_k \cdot \vec{d}_j$$

gdzie $w_{i,j}$ jest wagą (tu równą częstości) słowa i w dokumencie j .

Nauczyciele akademicki zatrudnieni **w** ...
uczelni wojskowej, **uczelni** służb państwowych, **uczelni** artystycznej, **uczelni** morskiej 2 nagrody
lub **uczelni** medycznej mogą otrzymywać 5 uczelnia
za osiągnięcia naukowe, dydaktyczne lub 4 w
organizacyjne albo **za** całokształt dorobku ...
nagrody rektora oraz **nagrody** właściwego 2 za
ministra wskazanego **w** art. 33 ust. 2, na ...
zasadach i **w** trybie określonych **w** ust. 3–7. ...

Wagi będące częstościami użycie słów w dokumencie kładą zbyt wielki nacisk na liczbę wystąpień. Im dłuższy dokument, tym ważniejszy. Dlatego wektor wag słów poddaje się **normalizacji** dzieląc wagi przez jego długość, czyli przez $\sqrt{\sum_{i=1}^N w_i^2}$. Wówczas podobieństwo:

$$\text{sim}(\vec{q}_k, \vec{d}_j) = \frac{\sum_{i=1}^N t_{i,k} \cdot t_{i,j}}{\sqrt{\sum_{i=1}^N q_i^2} \sqrt{\sum_{i=1}^N d_i^2}} = \frac{\vec{q}_k \cdot \vec{d}_j}{|\vec{q}_k| |\vec{d}_j|}$$

jest **cosinusem kąta** między wektorami jednostkowymi zapytania i dokumentu (lub dwóch dokumentów). Jeśli jest równe jeden, to dokumenty są identyczne, jeśli zero – niezależne (prostopadłe).

Częstość użycia poszczególnych słów w danym dokumencie nie oddaje ich wagi. Słowa, które występują w wielu dokumentach są mniej ważne niż te, które występują w nielicznych. Waga odwróconej częstotliwości występowania w dokumentach (ang. *inverse document frequency*) jest zdefiniowana jako:

$$\text{idf}_i = \log\left(\frac{N}{n_i}\right)$$

gdzie n_i to liczba dokumentów, w których występuje słowo i , a N to całkowita liczba dokumentów w zbiorze. Połączenie tej wagi z wagą częstości występowania słowa w dokumencie daje nam:

$$w_{i,j} = \text{tf}_{i,j} \cdot \text{idf}_i$$

Zapytania na ogół wcale nie przypominają dokumentów. Dlatego do obliczania wagi występującej w nich słów lepiej jest używać innych algorytmów. Salton i Buckley polecają następujący wzór, w którym $\max_j \text{tf}_{j,k}$ oznacza częstość występowania najczęstszego słowa w zapytaniu k :

$$w_{i,k} = \left(0.5 + \frac{0.5 \cdot \text{tf}_{i,k}}{\max_j \text{tf}_{j,k}} \right) \cdot \text{idf}_i$$

Najczęściej występujące w dokumentach (i zapytaniach) słowa (ang. *stop words*) mają niewielki wpływ (z powodu małej wagi) na znalezienie dokumentu, natomiast zajmują miejsce w plikach indeksowych, dlatego się je z tych plików usuwa.

Dla zapytań w języku angielskim **lematyzacja** (sprowadzenie formy odmienionej słowa do wyrazu hasłowego) często nie poprawia jakości wyszukiwania (bywa wręcz odwrotnie). Jednak dla języków z bogatą fleksją, a więc dla języków słowiańskich, lematyzacja pozwala na zwiększenie kompletności wyszukiwania i na wydatne zmniejszenie objętości plików indeksowych. Dla języków aglutynacyjnych jest wręcz konieczna.

Prosimy użytkownika o zaznaczenie tych z dokumentów przedstawionych mu jako odpowiedź na zapytanie, które uznaje za związane z zapytaniem. Następnie tworzymy nowe zapytanie używając innych wag. Chcemy „odepchnąć” wektor zapytania od niewłaściwych dokumentów w stronę tych właściwych. Uzyskujemy to przez dodanie uśrednionego wektora dla właściwych dokumentów i odjęcie uśrednionego wektora dla dokumentów niewłaściwych.

Niech \vec{q}_i reprezentuje wektor wag oryginalnego zapytania, R – liczbę dokumentów dotyczących pytania, S – liczbę dokumentów niezwiązanych z pytaniem. Niech γ i β przyjmują wartości należące do przedziału $\langle 0, 1 \rangle$ oraz $\gamma + \beta = 1$. Wówczas:

$$\vec{q}_{i+1} = \vec{q}_i + \frac{\beta}{R} \sum_{j=1}^R \vec{r}_j - \frac{\gamma}{S} \sum_{k=1}^S \vec{s}_k$$

Salton i Buckley osiągnęli dobre wyniki dla wartości $\beta = 0.75$ i $\gamma = 0.25$.

Oryginalne pytanie można rozszerzyć używając słownika **wyrazów bliskoznacznych** (ang. *thesaurus*). Zamiast stosować zwykły słownik wyrazów bliskoznacznych, można taki utworzyć samodzielnie poprzez **grupowanie słów** (ang. *term clustering*). Można to zrobić jednokrotnie dla całego zbioru dokumentów lub na podstawie dokumentów uznanych przez użytkownika za związane z pytaniem.

Dodatkowe materiały nt. wyszukiwania dokumentów/informacji

Dobrym źródłem są oba wymienione we wstępie podręczniki oraz materiały serii konferencji MUC (dostępne przez ACL Anthology). Poza tym:

- 1 Gerard Salton, James Allan, Chris Buckley, Amit Singhal, *Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts*, Science, vol. 264, 3 June 1994.
- 2 Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.
- 3 Stefan Büttcher, Charles L. A. Clarke, Gordon V. Cormack, *Information Retrieval. Implementing and Evaluating Search Engines*, MIT Press, 2010.

Kompletność (ang. *recall*) jest miarą tego, jak wiele istotnych danych zostało wyłuskanych z tekstu:

$$\text{Kompletność} = \frac{\text{Liczba prawidłowych odp. dostarczonych przez system}}{\text{Liczba możliwych prawidłowych odp. w tekście}}$$

Dokładność (ang. *precision, accuracy*) jest miarą poprawności odpowiedzi dostarczonych przez system:

$$\text{Dokładność} = \frac{\text{Liczba prawidłowych odp. dostarczonych przez system}}{\text{Liczba odp. dostarczonych przez system}}$$

Miara F (ang. *F-measure*, *F-score*) łączy kompletność z dokładnością (β jest parametrem):

$$\text{Miara F} = \frac{(\beta^2 + 1) \cdot \text{Dokładność} \cdot \text{Kompletność}}{\beta^2 \cdot \text{Dokładność} + \text{Kompletność}}$$

Pokrycie (ang. *coverage*) to udział elementów, dla których system udzielił odpowiedzi (poprawnej lub nie), w ogólnej ich liczbie:

$$\text{Pokrycie} = \frac{\text{Liczba elementów, dla których system może dostarczyć odp.}}{\text{Liczba wszystkich elementów}}$$

Przykłady miar

	prawda	system
1	V,N	V,N
2	V	V,N
3	V,A	V,N
4	V,N	N
5	N	V
6	V,N,A	V,N
7	A	-

kompletność = $7/(7+4) \approx 63.64\%$

dokładność = $7/(7+3) = 70.0\%$

pokrycie = $6/7 \approx 85.71\%$

β	1/2	1	2
F	$245/357$ $\approx 68.63\%$	$98/147$ $\approx 66.67\%$	$245/378$ $\approx 64.81\%$

Dla ostatniego elementu system nie mógł zostać zastosowany.