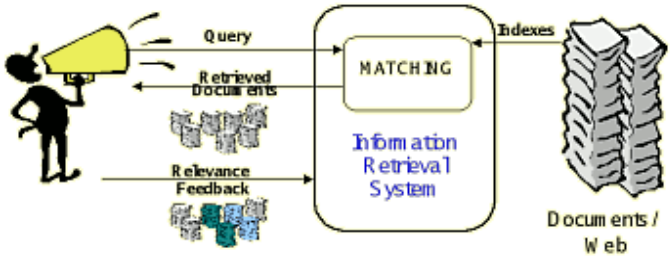


Wyszukiwanie i Przetwarzanie Informacji

Information Retrieval & Search

dr hab. inż. **Miłosz Kadziński**
dr inż. **Irmina Masłowska**

{milosz.kadzinski, irmina.maslowska}@cs.put.poznan.pl

http://www.cs.put.poznan.pl/imaslowska/wipi/

KARTA OPISU MODUŁU KSZTAŁCENIA

Nazwa modułu/przedmiotu **Wyszukiwanie i przetwarzanie zasobów informacyjnych**		Kod **1010511361010510091**
Kierunek studiów **Informatyka**	Profil kształcenia (ogólnoakademicki, praktyczny) **ogólnoakademicki**	Rok / Semestr **3 / 6**
Ścieżka obieralności/specjalność **-**	Przedmiot oferowany w języku: **polski**	Kurs (obligatoryjny/obieralny) **obieralny**
Stopień studiów: **I stopień**	**Forma studiów** (stacjonarna/niestacjonarna) **stacjonarna**	
Godziny Wykłady: **30** Ćwiczenia: **-** Laboratoria: **30** Projekty/seminaria: **-**		Liczba punktów **4**
Status przedmiotu w programie studiów (podstawowy, kierunkowy, inny) **kierunkowy**	(ogólnouczelniany, z innego kierunku) **z danego kierunku**	

Obszar(y) kształcenia i dziedzina(y) nauki i sztuki	**Podział ECTS (liczba i %)**
nauki techniczne	**4 100%**
nauki techniczne	**4 100%**

Odpowiedzialny za przedmiot / wykładowca:	**Odpowiedzialny za przedmiot / wykładowca:**
dr inż. Irmina Masłowska email: Irmina.Maslowska@cs.put.poznan.pl tel. 61 6652931 Instytut Informatyki ul. Piotrowo 2, 60-965 Poznań	dr hab. inż. Miłosz Kadziński email: Milosz.Kadzinski@cs.put.poznan.pl tel. 61 6653022 Instytut Informatyki ul. Piotrowo 2, 60-965 Poznań

Sposoby weryfikacji efektów kształcenia

Efekty kształcenia przedstawione wyżej weryfikowane są w następujący sposób:

Ocena formująca:

a) w zakresie wykładów:
- na podstawie odpowiedzi na pytania dotyczące materiału omówionego na poprzednich wykładach,

b) w zakresie laboratoriów / ćwiczeń:
- na podstawie oceny bieżącego postępu realizacji zadań.

Ocena podsumowująca:

a) w zakresie wykładów weryfikowanie założonych efektów kształcenia realizowane jest przez:
- ocenę wiedzy i umiejętności wykazanych na zaliczeniu pisemnym w formie testu składającego się z ok. 20 zadań otwartych: rozszerzonej odpowiedzi i/lub z krótką odpowiedzią, przy czym dla uzyskania oceny dostatecznej student musi zdobyć ponad 50% całkowitej liczby punktów,
- omówienie wyników zaliczenia,

b) w zakresie laboratoriów / ćwiczeń weryfikowanie założonych efektów kształcenia realizowane jest przez:
- ocenę umiejętności związanych z realizacją ćwiczeń laboratoryjnych,
- ocenę sprawozdania z realizacji zadań analitycznych i symulacyjnych przygotowywanego częściowo w trakcie zajęć, a częściowo po ich zakończeniu; ocena ta obejmuje także umiejętność pracy w zespole,
- ocenę kodu źródłowego z realizacji zadań programistycznych oraz „obronę” projektów przez studenta.

Uzyskiwanie punktów dodatkowych za aktywność podczas zajęć, a szczególnie za:
- omówienie dodatkowych aspektów zagadnienia,
- efektywność zastosowania zdobytej wiedzy podczas rozwiązywania zadanego problemu,
- wskazywanie trudności percepcyjnych studentów umożliwiające bieżące doskonalenia procesu dydaktycznego.

Metody dydaktyczne:

1. Wykład: prezentacja multimedialna ilustrowana przykładami podawanymi na tablicy.
2. Ćwiczenia laboratoryjne: rozwiązywanie zadań, ćwiczenia praktyczne, wykonywanie eksperymentów, dyskusja, praca w zespole, studium przypadków, demonstracja wybranych systemów przetwarzania informacji oraz pokaz multimedialny

Literatura podstawowa:

1. Eksploracja zasobów internetowych, Z.Markov, D.T.Larose, PWN, 2009

2. Introduction to Information Retrieval, Ch.D.Manning, P.Raghavan, H.Schütze, Cambridge University Press, 2008 (wersja poprawiona i uzupełniona w 2009 r. dostępna bezpłatnie on-line: http://nlp.stanford.edu/IR-book/)

3. Mining of Massive Datasets, Anand Rajaraman, Jeffrey David Ullman, Cambridge University Press, 2011 (wersja poprawiona i uzupełniona w 2012 r. dostępna bezpłatnie on-line: http://infolab.stanford.edu/~ullman/mmds/book.pdf)

4. Modern Information Retrieval, Ricardo Baeza-Yates, Berthier Ribeiro-Neto, Addison-Wesley, 1999

5. Data intensive text-processing with MapReduce, Jimmy Lin, Chris Dyer, University of Maryland, Morgan & Claypool Synthesis, 2010 (dostępna bezpłatnie on-line: https://lintool.github.io/MapReduceAlgorithms/MapReduce-book-final.pdf)

Literatura uzupełniająca:

1. Speech and Language Processing (3rd ed. draft), D. Jurafsky and J.H. Martin (wersja z 2018 dostępna bezpłatnie on-line: https://web.stanford.edu/~jurafsky/slp3)

2. Foundations of Statistical Natural Language Processing, Ch.D.Manning, H. Schütze, MIT Press, Cambridge Massachusetts, MIT Press Cambridge Mass, 1999

3. Web Data Mining: Exploring Hyperlinks, Contents and Usage Data. B. Liu, Springer, 2009

4. Mining the Web: Discovering Knowledge from Hypertext Data. S. Chakrabarti, Morgan Kaufmann, 2002

5. The Text Mining Handbook. R. Feldman, J. Sanger, Cambridge University Press, 2006

6. Felietony publikowane na bieżąco na http://searchenginewatch.com, http://searchengineland.com/

Information resources

Usługi (i protokoły) internetowe

- WWW (World Wide Web) (pages, services)
- Poczta elektroniczna (e-mails)
- Transfer plików (FTP – File Transfer Protocol, SFTP – Secure File Transfer Protocol)
- Serwisy społecznościowe
- Blogi, fora i listy dyskusyjne
- Komunikatory “instant messengers”, Telekonferencje
- VoIP czyli telefonia internetowa
- Radio i telewizja, video na żądanie – IPTV
- Telnet, SSH (Secure Shell)
- Sklepy i aukcje internetowe
- Bankowość elektroniczna
- Blockchain
- Gry online
- Sieci wymiany bezpośredniej P2P
- Czaty, jak IRC (Internet Relay Chat)
- Gopher ☺

Information resources

Usługi (i protokoły) internetowe

- WWW (World Wide Web) (pages, services)
- Poczta elektroniczna (e-mails)
- Transfer plików (FTP – File Transfer Protocol, SFTP – Secure File Transfer Protocol)
- Serwisy społecznościowe
- Blogi, fora i listy dyskusyjne
- Komunikatory "instant messengers", Telekonferencje
- VoIP czyli telefonia internetowa
- Radio i telewizja, video na żądanie – IPTV
- Telnet, SSH (Secure Shell)
- Sklepy i aukcje internetowe
- Bankowość elektroniczna
- Blockchain
- Gry online
- Sieci wymiany bezpośredniej P2P
- Czaty, jak IRC (Internet Relay Chat)
- Gopher ☺

Tradycyjne

- zasoby drukowane (książki, dzienniki, listy, czasopisma, poradniki, opracowania naukowe, zdjęcia, itp. dostępne poprzez różne biblioteki lub archiwa; wszelkie dokumenty danej firmy, organizacji, państwa ...)
- obiekty muzealne, dzieła sztuki, obiekty architektury, wykopaliska, skamieliny …
- wiedza ekspertów, doświadczenie jednostek …

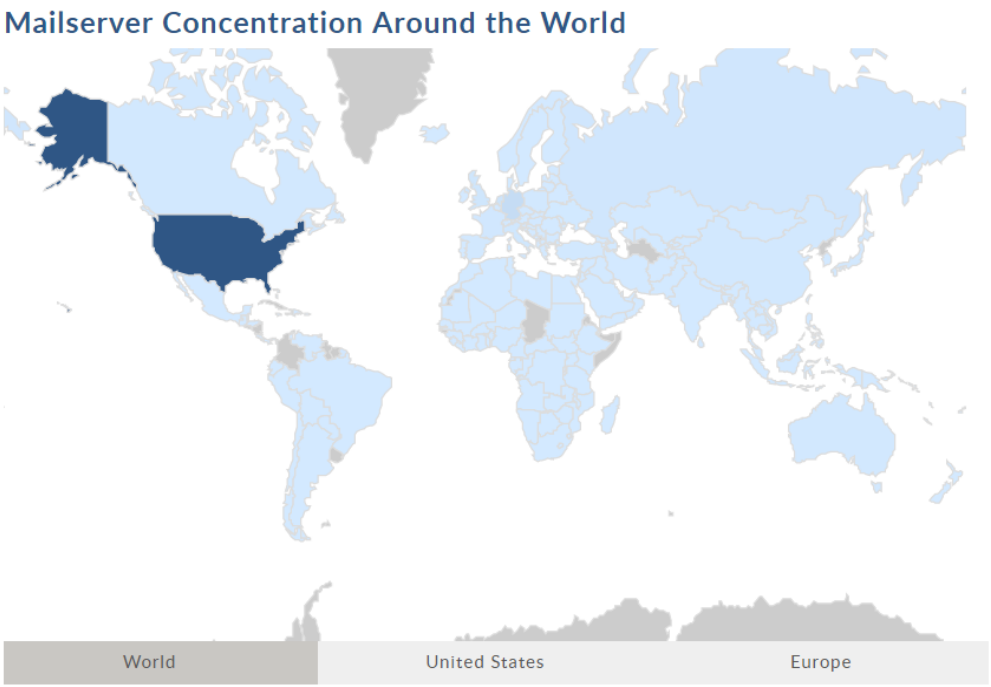
Access to Internet in numbers

Użytkownicy Internetu stanowią **58.7%** ludności świata (ponad 4 570 milionów)

Ameryka Północna:	94.6% (wzrost 2000-2020:	222%)
Europa:	87.2% (wzrost 2000-2020:	592%)
Australia i Oceania:	67.4% (wzrost 2000-2020:	277%)
Ameryka Południowa:	68.9% (wzrost 2000-2020:	2 411%)
Bliski Wschód:	69.2% (wzrost 2000-2020:	5 395%)
Azja:	53.6% (wzrost 2000-2020:	1 913%)
Afryka:	39.3% (wzrost 2000-2020:	11 559%)

http://www.internetworldstats.com/stats.htm

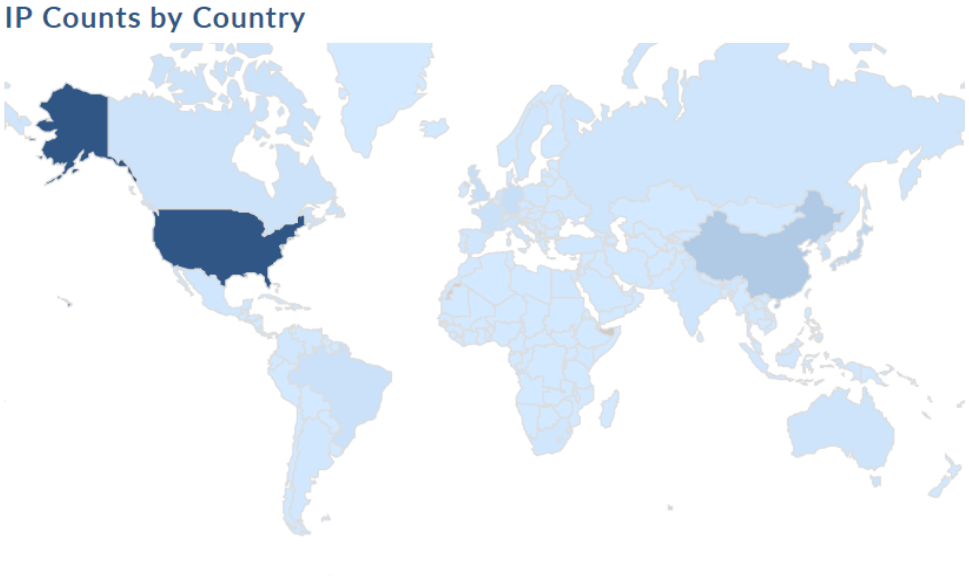
Access to Internet in numbers



Name	# Servers
United States	149313079
Germany	13200611
Netherlands	5398238
Great Britain (UK)	3554077
Russian Federation	2748165
Italy	2291137
France	1750514
Canada	1308124
Poland	1198903
Japan	1190011
China	1100841
Turkey	1010773
Czech Republic	955243

http://research.domaintools.com/statistics/ip-addresses/

Access to Internet in numbers



Country	# Addresses
United States	1,598,831,868
China	339,371,664
Japan	206,499,072
United Kingdom	124,376,381
Germany	123,579,519
Korea, Republic of	112,140,992
France	85,377,535
Brazil	84,825,164
Canada	70,567,354
Italy	54,493,830
Netherlands	50,930,861
Australia	50,594,448

http://research.domaintools.com/statistics/ip-addresses/

Access to Internet in numbers

In 2014, there were 2.4 billion internet users. That number grew to 3.4 billion by 2016, and in 2017 over 300 million internet users were added. As of June 2019 there were over 4.4 billion internet users. This is an 83% increase in the number of people using the internet in just 5 years!

Not only are there more people using the internet, but they are using it in many different ways

https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day

Information overload

Źródło: http://www.marketlogicsoftware.com/blog/detail/too-much-data-not-enough-insights

Information overload

Social Media – social media gains 840 new users each MINUTE

Since 2013, the number of Facebook Posts shared each minute has increased 22%, from 2.5 million to 3 million posts per minute in 2016. This number has increased more than 300%, from around 650,000 posts per minute in 2011

Every minute on Facebook: 510,000 comments are posted, 293,000 statuses are updated, and 136,000 photos are uploaded

Facebook users also click the like button on more than 4 million posts every minute

Since 2013, the number of Tweets each minute has increased 58% to more than 455,000 Tweets per minute in 2017

YouTube usage more than tripled from 2014-2016 with users uploading 400 hours of new video each minute. In 2017, users are watching 4,146,600 videos every minute

Instagram users post 46,740 pictures every minute

Worldwide, 15,220,700 texts are sent every minute

3,607,080 Google searches are conducted worldwide each minute of everyday

https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day

Information overload

Each DAY 1,209,600 new social media users

Over 2 billion monthly active Facebook users, compared to 1.44 billion at the start of 2015 and 1.65 at the start of 2016

Facebook has 1.32 billion daily active users on average as of June 2017

4.3 billion Facebook messages posted daily

5.75 billion Facebook likes every day

656 million tweets per day

More than 4 million hours of content uploaded to Youtube every day, with users watching 5.97 billion hours of Youtube videos each day

67,305,600 Instagram posts uploaded each day

22 billion texts sent every day

5.2 billion daily Google Searches in 2017

269 billion emails were sent daily in 2017, and this is expected to grow by 4.4% yearly to 319.6 billion in 2021

https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day

Data created by the Internet of Things (IOT)

Devices are a huge source of the 2.5 quintillion bytes of data we create every day – not just mobile devices, but Smart TV's, cars, airplanes, etc. – the internet of things is producing an increasing amount of data

IDC forecasts a 31% growth in wearable devices from 2016 – 2020. There were 28.3 million wearable devices sold in 2016, which would mean 82.5 million in 2020

Pratt & Whitney's Geared Turbo Fan (GTF) engine is fitted with 5,000 sensors and can generate up to 10GB of data each second

Business insider predicts that by 2020 75% of cars will come with built-in IoT connectivity

Uber is releasing 6 years of transportation data to cities to help them plan public transit

https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day

Web mining

>> wykorzystanie technik data mining
w automatycznym odkrywaniu i pozyskiwaniu
informacji z dokumentów i usług dostępnych
w sieci Web <<

O. Etzioni, *The World-Wide Web: Quagmire or gold mine?*
Communications of ACM, 39(11):65-68, 1996

jako dyscyplina naukowa leży na „przecięciu” badań

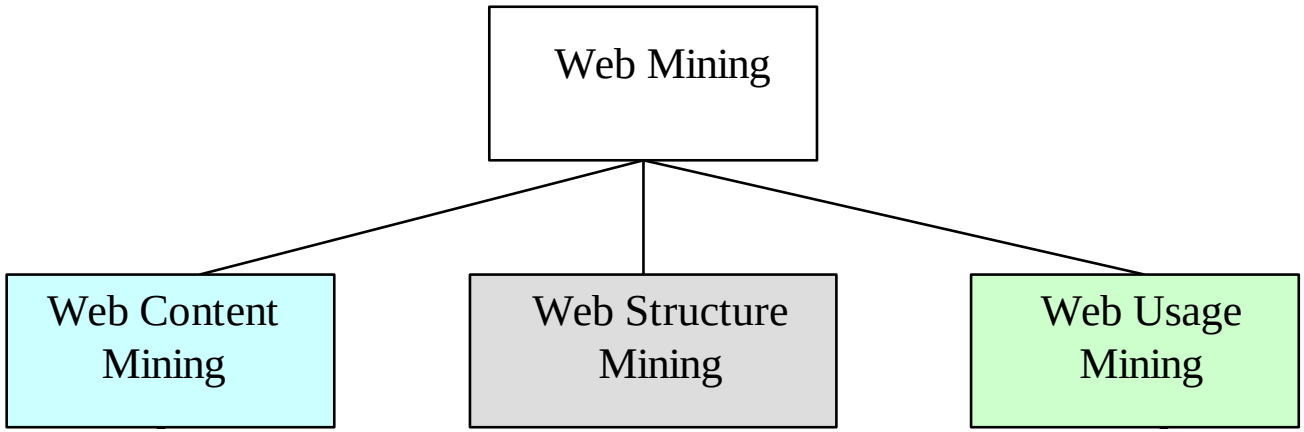
- baz danych i statystyki
- wyszukiwania informacji (Information Retrieval —IR)
- sztucznej inteligencji (w AI w szczególności czerpie z uczenia maszynowego — ML i przetwarzania języka naturalnego — NLP)

Główne problemy/zadania Web Mining

- dotarcie do adekwatnej informacji (wyszukiwanie) — głównie IR
- pozyskiwanie wiedzy z dostępnej informacji — głównie DM
- personalizacja informacji (indywidualizacja zarówno co do treści jak i formy)

3 nurty badań w ramach Web Mining

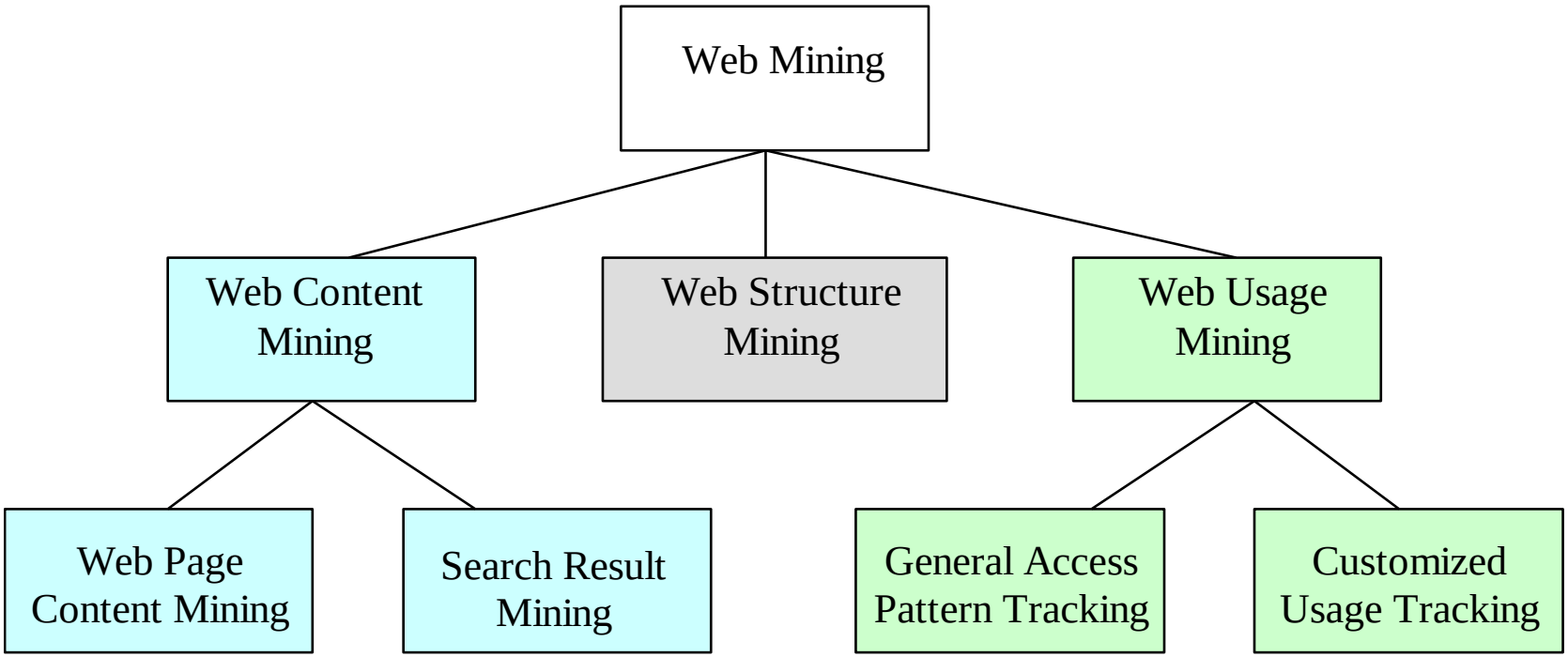
- eksploracja zawartości (treści) – Content mining
- eksploracja struktury – Structure mining
- eksploracja wykorzystania – Usage mining



Za: Jiawei Han, 1998

3 nurty badań w ramach Web Mining

- eksploracja zawartości (treści) – Content mining
- eksploracja struktury – Structure mining
- eksploracja wykorzystania – Usage mining



Za: Jiawei Han, 1998

Web mining

	Content Mining		Structure Mining	Usage Mining
	IR View	DB View		
View of Data	- Unstructured - Semi structured	- Semi structured - Web site as DB	- Links structure	- Interactivity
Main Data	- Text documents - Hypertext documents	- Hypertext documents	- Links structure	- Server logs - Browser logs
Representation	- Bag of words, n-grams - Terms, phrases - Concepts or ontology - Relational	- Edge-labeled graph (OEM) - Relational	- Graph	- Relational table - Graph
Method	- TFIDF and variants - Machine learning - Statistical (including NLP)	- Proprietary algorithms - ILP - (Modified) association rules	- Proprietary algorithms	- Machine learning - Statistical - (Modified) association rules
Application Categories	- Categorization - Clustering - Finding extraction rules - Finding patterns in text - User modeling	- Finding frequent substructures - Web site schema discovery	- Categorization - Clustering	- Site construction, adaptation and management - Marketing - User modeling

Za: Kosala-Blockeel, 2000

Information Rertrieval

>> IR is the automatic retrieval of ALL relevant documents while retrieving as FEW of the irrelevant as possible <<

van Rijsbergen C.J. (1979) *Information Retrieval*. Butterworths, London

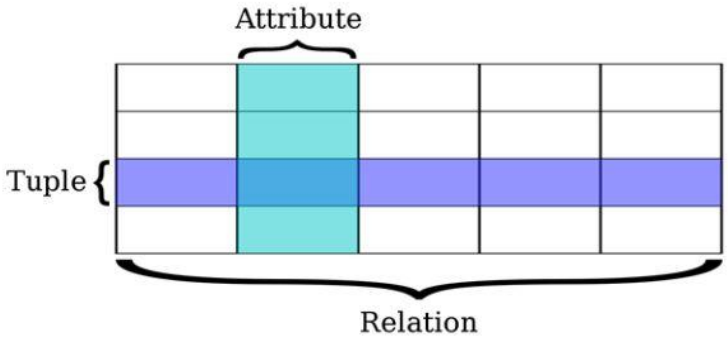
>> IR deals with the representation, storage, organization of, and access to information items <<

Baeza-Yates R., Ribeiro-Neto B. (1999) *Modern Information Retrieval*. Addison-Wesley, ACM Press, New York

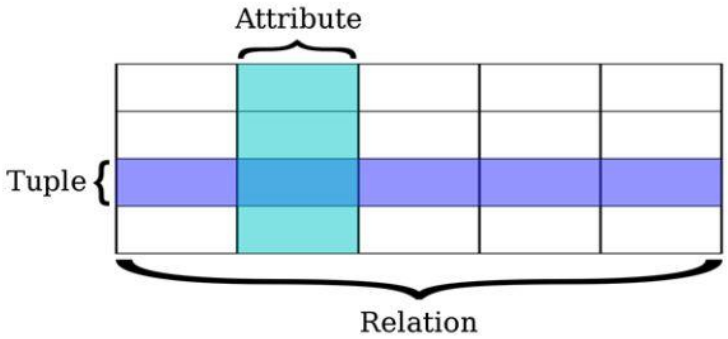
Badania IR dotyczą:

- szukania informacji w kolekcjach tekstów (search)
- indeksowania dokumentów (indexing)
- modelowania (modeling)
- klasyfikacji dokumentów (categorization)
- analizy skupień (clustering)
- architektury systemów (system architecture)
- interfejsów użytkownika (user interfaces)
- wizualizacji (visualization)
- filtrowania (filtering)
- ...

Information Retrieval vs. Data Retrieval



Information Retrieval vs. Data Retrieval



słaba strukturalizacja lub brak

$\Longleftrightarrow$

dobrze zdefiniowana struktura i semantyka

Information Retrieval vs. Data Retrieval

- Unorganized
- Unrelated
- In multiple places
- Reflects human behaviour
- Grows very rapidly

Coming from

- Payments
- Searches
- Choices
- Actions
- Images
- Recordings
- Opinions
- ...

www.linkedin.com/pulse/big-data-wealth-management-from-investment-consulting-zollinger-cfa

Unstructured data

90% of the data on the internet has been created since 2016, according to an IBM Marketing Cloud study

Research shows that 80% of worldwide data will be unstructured by 2025, meaning that it is typically **text heavy** and **does not follow a predefined data model**. That's where big data analytics and visibility come into play – they give context to massive amounts of unstructured data

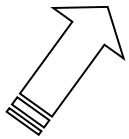
https://solutionsreview.com/data-management/80-percent-of-your-data-will-be-unstructured-in-five-years/

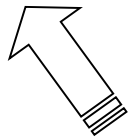
Access to relevant information

od czego zależy
efektywny dostęp do adekwatnej informacji

Access to relevant information

efektywny dostęp do adekwatnej informacji





działanie użytkownika

reprezentacja dokumentów

ROZDZIAŁ I MIRAŻE

Imię moje: Izmael. Przed kilku laty — mniejsza o ścisłość jak dawno temu — mając niewiele czy też nie mając wcale pieniędzy w sakiewce, a nie widząc nic szczególnego, co by mnie interesowało na lądzie, pomyślałem sobie, że pożegluję nieco po morzach i obejrzę wodną część świata. Taki mam właśnie sposób odpędzania splinu i regulowania krwiobiegu. Gdy tylko stwierdzę, że usta wykrzywiają mi się ponuro, gdy tylko do duszy mej zawita wilgotny, dżdżysty listopad, gdy złapię się na tym, że mimowolnie przystaję przed składami trumien albo podążam za każdym napotkanym pogrzebem, a w szczególności, gdy moja hipochondria tak mnie opanuje, iż potrzeba mi silnych zasad moralnych, by się powstrzymać od rozmyślnego wyjścia na ulicę i metodycznego strącania ludziom z głów kapeluszy — wtedy uznaję, że już wielki czas udać się na morze jak najrychlej. To jest moja namiastka pistoletu i kuli. Katon z filozoficzną oracją rzuca się na ostrze swego miecza; ja spokojnie siadam na okręt. Nie ma w tym nic zdumiewającego. Gdyby tylko zdawano sobie z tego sprawę, okazałoby się, że niemal wszyscy ludzie, każdy na swój sposób, w takiej czy innej chwili, żywią wobec oceanu niemal te same co ja uczucia.

Oto macie wyspiarski gród manhattańczyków, opasany przystaniami, jak indyjskie wyspy rafami koralowymi. Otaczają go zewsząd spienione nurty handlu. Z prawej i lewej strony ulice wiodą was ku wodzie. Najdalszym skrajem miasta jest plac Battery, kędy wspaniałe molo obmywają fale i chłodzą bryzy, które jeszcze kilka godzin temu nie widziały lądu. Spójrzcie na tłumy wpatrzone tam w wodę.

Powędrujcie wokół miasta w zadumane niedzielne popołudnie. Przejdźcie od Corlears Hook do Coenties Slip, a stamtąd przez Whitehall ku północy. Cóż ujrzycie? Rozstawieni wokoło całego miasta jak milczące szyldwachy, tkwią tysiącami i tysiącami śmiertelnicy zatopieni w oceanicznych marzeniach. Jedni wsparli się o pale, drudzy zasiedli na skrajach pomostów, ci wyglądają przez burty statków przybyłych z Chin, inni wdrapali się wy- soko na olinowanie, jak gdyby usiłując zyskać jeszcze lepszy widok na morze. Ale przecie to wszystko ludzie lądu, w powszedni dzień uwięzieni wśród desek i tynku — przywiązani do kontuarów, przygwożdżeni do ław, przytwierdzeni do biurek. Jakże więc to się dzieje? Czyżby zniknęły łany zielone? Co oni tu robią?

Text documents

Języki naturalne charakteryzują się:

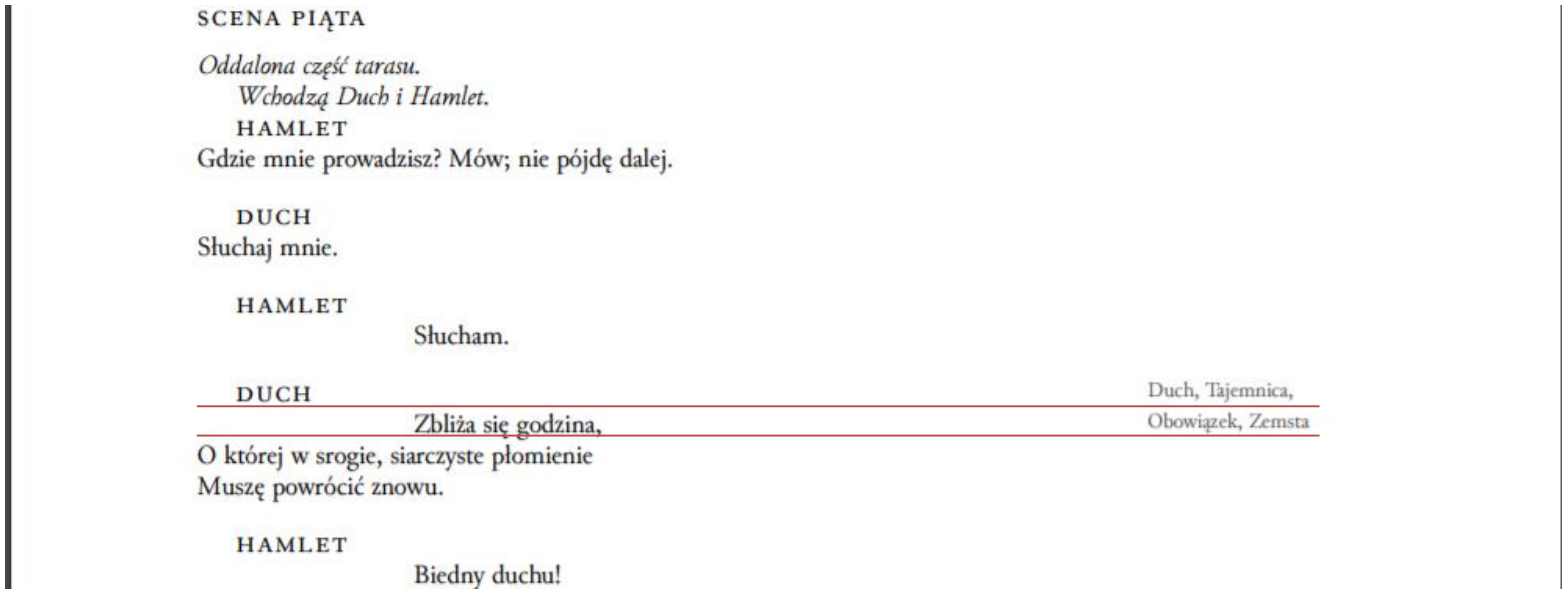
- *Nieprecyzyjnością* na każdym z poziomów opisu lingwistycznego:
  - fonetyka i morfologia (dźwięki i słowa)
  - składnia (struktura zdań)
  - semantyka (treść wypowiedzi, jej znaczenie)
  - pragmatyka (teoria sposobu używania języka)
- Koniecznością posiadania *wiedzy dziedzinowej*, aby *rzeczywiście rozumieć* tekst

Initial choices

- Określenie poprawnego kodowania plików
- Wyodrębnienie zawartości tekstowej i identyfikacja *strumienia* tekstu

Initial choices

- Określenie poprawnego kodowania plików
- Wyodrębnienie zawartości tekstowej i identyfikacja *strumienia* tekstu



SCENA PIĄTA

Oddalona część tarasu.
Wchodzą Duch i Hamlet.

HAMLET
Gdzie mnie prowadzisz? Mów; nie pójdę dalej.

DUCH
Słuchaj mnie.

HAMLET
Słucham.

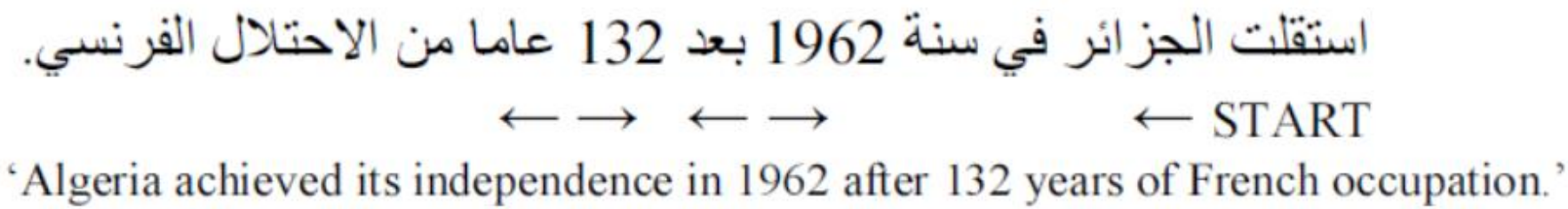
DUCH
Zbliża się godzina,
O której w srogie, siarczyste płomienie
Muszę powrócić znowu.

Duch, Tajemnica,
Obowiązek, Zemsta

HAMLET
Biedny duchu!

Initial choices

- Określenie poprawnego kodowania plików
- Wyodrębnienie zawartości tekstowej i identyfikacja *strumienia* tekstu



Initial choices

- Określenie poprawnego kodowania plików
- Wyodrębnienie zawartości tekstowej i identyfikacja *strumienia* tekstu

استقلت الجزائر في سنة 1962 بعد 132 عاما من الاحتلال الفرنسي.

← START ← → ← →

'Algeria achieved its independence in 1962 after 132 years of French occupation.'

- Ustalenie jaka jednostka tekstu będzie stanowić oddzielny *dokument*

  np. cała książka, czy rozdział? encyklopedia? blog? wolumen, numer czasopisma, czy pojedynczy artykuł? witryna sklepu, czy strony oferowanych towarów? mail z załącznikami to jeden dokument, czy kilka?

Preprocessing

Wstępne przetwarzanie tekstów ma na celu wyłonienie jednostek indeksujących (ang. *index terms*, *keywords*), które stworzą reprezentację dokumentów danej kolekcji

Główna motywacja:

- niektóre słowa tekstu mają większe znaczenie od innych; niosą więcej informacji o *zawartości/treści*, czy też wręcz *tematyce* dokumentu
- użycie wszystkich słów dokumentów kolekcji spowodowałoby zbytnie *zaszumienie* zadania wyszukiwania informacji

Preprocessing

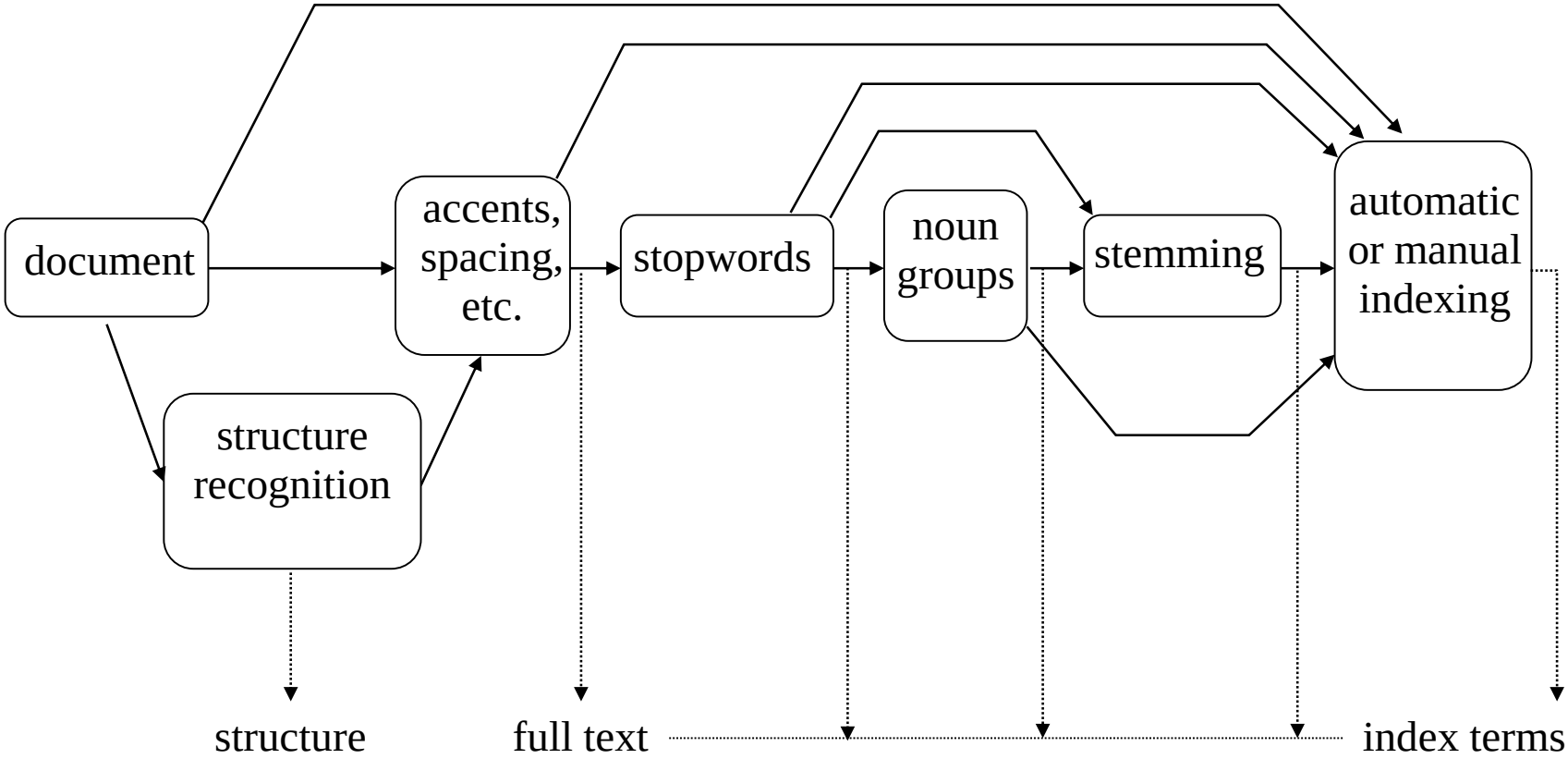
Jak podaje SIL International — światowa organizacja zajmująca się językami świata — język polski jest jednym z 6909 żywych języków świata (takich, którymi posługuje się jakaś grupa ludzi jako językami ojczystymi). Gdyby wziąć pod uwagę tylko języki, którymi posługuje się ponad milion osób można by się ograniczyć do 389 języków

Preprocessing

Wstępne przetwarzanie tekstów jest silnie uzależnione od języka tekstu. Jednak można wyróżnić pewne wspólne standardowe etapy/zadania

W systemie IR należy **spójnie** przetwarzać **dokumenty** oraz **zapytania** przychodzące od użytkowników

Preprocessing



Za: Baeza-Yates & Ribeiro-Neto, 1999

Tokenization

Tokenizacja wykorzystująca analizę leksykalną pozwala na wyodrębnienie z dokumentów jednostek leksykalnych stanowiących potencjalnie *słowa w języku naturalnym* lub inne *hasła*, które mogą być przydatne z punktu widzenia wyszukiwania

Tokenization

Zwykle stosuje się ogólne reguły jednocześnie z listą wyjątków, np. w postaci wyrażeń regularnych

Bazuje się na zdefiniowanych separatorach, którymi mogą być spacje i inne znaki niedrukowalne oraz określone znaki przestankowe, np. ? ! . , : ; ‘ — „ ” ( ) { } [ ] ...

Tokenization

ROZDZIAŁ I MIRAŻE

Imię moje: Izmael. Przed kilku laty — mniejsza o ścisłość jak dawno temu — mając niewiele czy też nie mając wcale pieniędzy w sakiewce, a nie widząc nic szczególnego, co by mnie interesowało na lądzie, pomyślałem sobie, że pożegluję nieco po morzach i obejrzę wodną część świata. Taki mam właśnie sposób odpędzania splinu i regulowania krwiobiegu. Gdy tylko stwierdzę, że usta wykrzywiają mi się ponuro, gdy tylko do duszy mej zawita wilgotny, dżdżysty listopad, gdy złapię się na tym, że mimowolnie przystaję przed składami trumien albo podążam za każdym napotkanym pogrzebem, a w szczególności, gdy moja hipochondria tak mnie opanuje, iż potrzeba mi silnych zasad moralnych, by się powstrzymać od rozmyślnego wyjścia na ulicę i metodycznego strącania ludziom z głów kapeluszy — wtedy uznaję, że już wielki czas udać się na morze jak najrychlej. To jest moja namiastka pistoletu i kuli. Katon z filozoficzną oracją rzuca się na ostrze swego miecza; ja spokojnie siadam na okręt. Nie ma w tym nic zdumiewającego. Gdyby tylko zdawano sobie z tego sprawę, okazałoby się, że niemal wszyscy ludzie, każdy na swój sposób, w takiej czy innej chwili, żywią wobec oceanu niemal te same co ja uczucia.

Oto macie wyspiarski gród manhattańczyków, opasany przystaniami, jak indyjskie wyspy rafami koralowymi. Otaczają go zewsząd spienione nurty handlu. Z prawej i lewej strony ulice wiodą was ku wodzie. Najdalszym skrajem miasta jest plac Battery, kędy wspaniałe molo obmywają fale i chłodzą bryzy, które jeszcze kilka godzin temu nie widziały lądu. Spójrzcie na tłumy wpatrzone tam w wodę.

Powędrujcie wokół miasta w zadumane niedzielne popołudnie. Przejdźcie od Corlears Hook do Coenties Slip, a stamtąd przez Whitehall ku północy. Cóż ujrzycie? Rozstawieni wokoło całego miasta jak milczące szyldwachy, tkwią tysiącami i tysiącami śmiertelnicy zatopieni w oceanicznych marzeniach. Jedni wsparli się o pale, drudzy zasiedli na skrajach pomostów, ci wyglądają przez burty statków przybyłych z Chin, inni wdrapali się wy- soko na olinowanie, jak gdyby usiłując zyskać jeszcze lepszy widok na morze. Ale przecie to wszystko ludzie lądu, w powszedni dzień uwięzieni wśród desek i tynku — przywiązani do kontuarów, przygwożdżeni do ław, przytwierdzeni do biurek. Jakże więc to się dzieje? Czyżby zniknęły łany zielone? Co oni tu robią?

Tokenization

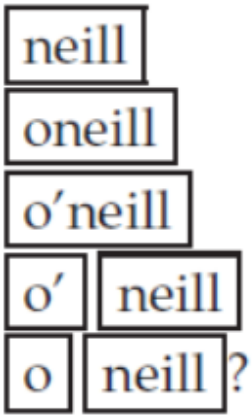
Już na tym etapie nawet dla stosunkowo „prostych” języków (jak angielski) mogą pojawić się problemy

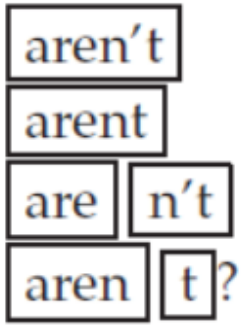
Tokenization

Już na tym etapie nawet dla stosunkowo „prostych” języków (jak angielski) mogą pojawić się problemy

Apostrofy służą zarówno do tworzenia form dzierżawczych jak i popularnych skrótów:

Mr. O’Neill thinks that the boys’ stories about Chile’s capital aren’t amusing.





Tokenization

Myślniki używane są w słowotwórstwie, do przenoszenia części wyrazu do następnego wiersza, ale także w nazwach własnych, wyrazach złożonych, itp.:

co-education, the hold-him-backand-drag-him-away maneuver, Hewlett-Packard

Tokenization

Myślniki używane są w słowotwórstwie, do przenoszenia części wyrazu do następnego wiersza, ale także w nazwach własnych, wyrazach złożonych, itp.:

co-education, the hold-him-backand-drag-him-away maneuver, Hewlett-Packard

Podobne problemy mogą wynikać z użycia spacji:

San Francisco, Los Angeles, York University, *ex aequo*, white space (vs. whitespace), lowercase (vs. lower-case and lower case), Mar 11, 1983

Tokenization

Myślniki używane są w słowotwórstwie, do przenoszenia części wyrazu do następnego wiersza, ale także w nazwach własnych, wyrazach złożonych, itp.:

co-education, the hold-him-backand-drag-him-away maneuver, Hewlett-Packard

Podobne problemy mogą wynikać z użycia spacji:

San Francisco, Los Angeles, York University, *ex aequo*, white space (vs. whitespace), lowercase (vs. lower-case and lower case), Mar 11, 1983

Na tym etapie tokeny bywają sprowadzane do zapisu małymi literami, ale: *Polish polish, Bush bush, Windows*...

Tokenization

When Google encounters a hyphen (–) in a query term, e.g., [e-mail], it searches for:

- the term with the hyphen: e-mail
- the term without the hyphen: email
- the term with the hyphen replaced by a space: e mail

[e-mail] matches “e-mail,” “e mail,” and “email”

[e mail] matches “e-mail” and “e mail”

If you aren’t sure whether a word is hyphenated, search for it with a hyphen

http://www.googleguide.com/favorite_results.html

Tokenization

ROZDZIAŁ I MIRAŻE

Imię moje: Izmael. Przed kilku laty — mniejsza o ścisłość jak dawno temu — mając niewiele czy też nie mając wcale pieniędzy w sakiewce, a nie widząc nic szczególnego, co by mnie interesowało na lądzie, pomyślałem sobie, że pożegluję nieco po morzach i obejrzę wodną część świata. Taki mam właśnie sposób odpędzania splinu i regulowania krwiobiegu. Gdy tylko stwierdzę, że usta wykrzywiają mi się ponuro, gdy tylko do duszy mej zawita wilgotny, dżdżysty listopad, gdy złapię się na tym, że mimowolnie przystaję przed składami trumien albo podążam za każdym napotkanym pogrzebem, a w szczególności, gdy moja hipochondria tak mnie opanuje, iż potrzeba mi silnych zasad moralnych, by się powstrzymać od rozmyślnego wyjścia na ulicę i metodycznego strącania ludziom z głów kapeluszy — wtedy uznaję, że już wielki czas udać się na morze jak najrychlej. To jest moja namiastka pistoletu i kuli. Katon z filozoficzną oracją rzuca się na ostrze swego miecza; ja spokojnie siadam na okręt. Nie ma w tym nic zdumiewającego. Gdyby tylko zdawano sobie z tego sprawę, okazałoby się, że niemal wszyscy ludzie, każdy na swój sposób, w takiej czy innej chwili, żywią wobec oceanu niemal te same co ja uczucia.

Oto macie wyspiarski gród manhattańczyków, opasany przystaniami, jak indyjskie wyspy rafami koralowymi. Otaczają go zewsząd spienione nurty handlu. Z prawej i lewej strony ulice wiodą was ku wodzie. Najdalszym skrajem miasta jest plac Battery, kędy wspaniałe molo obmywają fale i chłodzą bryzy, które jeszcze kilka godzin temu nie widziały lądu. Spójrzcie na tłumy wpatrzone tam w wodę.

Powędrujcie wokół miasta w zadumane niedzielne popołudnie. Przejdźcie od Corlears Hook do Coenties Slip, a stamtąd przez Whitehall ku północy. Cóż ujrzycie? Rozstawieni wokoło całego miasta jak milczące szyldwachy, tkwią tysiącami i tysiącami śmiertelnicy zatopieni w oceanicznych marzeniach. Jedni wsparli się o pale, drudzy zasiedli na skrajach pomostów, ci wyglądają przez burty statków przybyłych z Chin, inni wdrapali się wy- soko na olinowanie, jak gdyby usiłując zyskać jeszcze lepszy widok na morze. Ale przecie to wszystko ludzie lądu, w powszedni dzień uwięzieni wśród desek i tynku — przywiązani do kontuarów, przygwożdżeni do ław, przytwierdzeni do biurek. Jakże więc to się dzieje? Czyżby zniknęły łany zielone? Co oni tu robią?

Tokenization

Należy także rozważyć uwzględnienie pewnych specyficznych tokenów, jak numery telefonów, kody pocztowe, adresy e-mail, adresy IP, URL, hashtagi i inne wyjątki

np. C++, C#, F-16, M*A*S*H, B-tree

Tokenization

Należy także rozważyć uwzględnienie pewnych specyficznych tokenów, jak numery telefonów, kody pocztowe, adresy e-mail, adresy IP, URL, hashtagi i inne wyjątki

np. C++, C#, F-16, M*A*S*H, B-tree

Ostatecznie najczęściej stosuje się proste heurystyki i ogólne reguły z listą wyjątków, np. w postaci wyrażeń regularnych

Language specific problems

rzeczowniki złożone w j. niemieckim

Computerlinguistik – computational linguistics
Lebensversicherungsgesellschaftsangestellter – life insurance company employee

brak ‘spacji’ w językach dalekowschodnich

莎拉波娃现在居住在美国东南部的佛罗里达。今年 4 月
9 日，莎拉波娃在美国第一大城市纽约度过了 1 8 岁生
日。生日派对上，莎拉波娃露出了甜美的微笑。

Language specific problems

rzeczowniki złożone w j. niemieckim

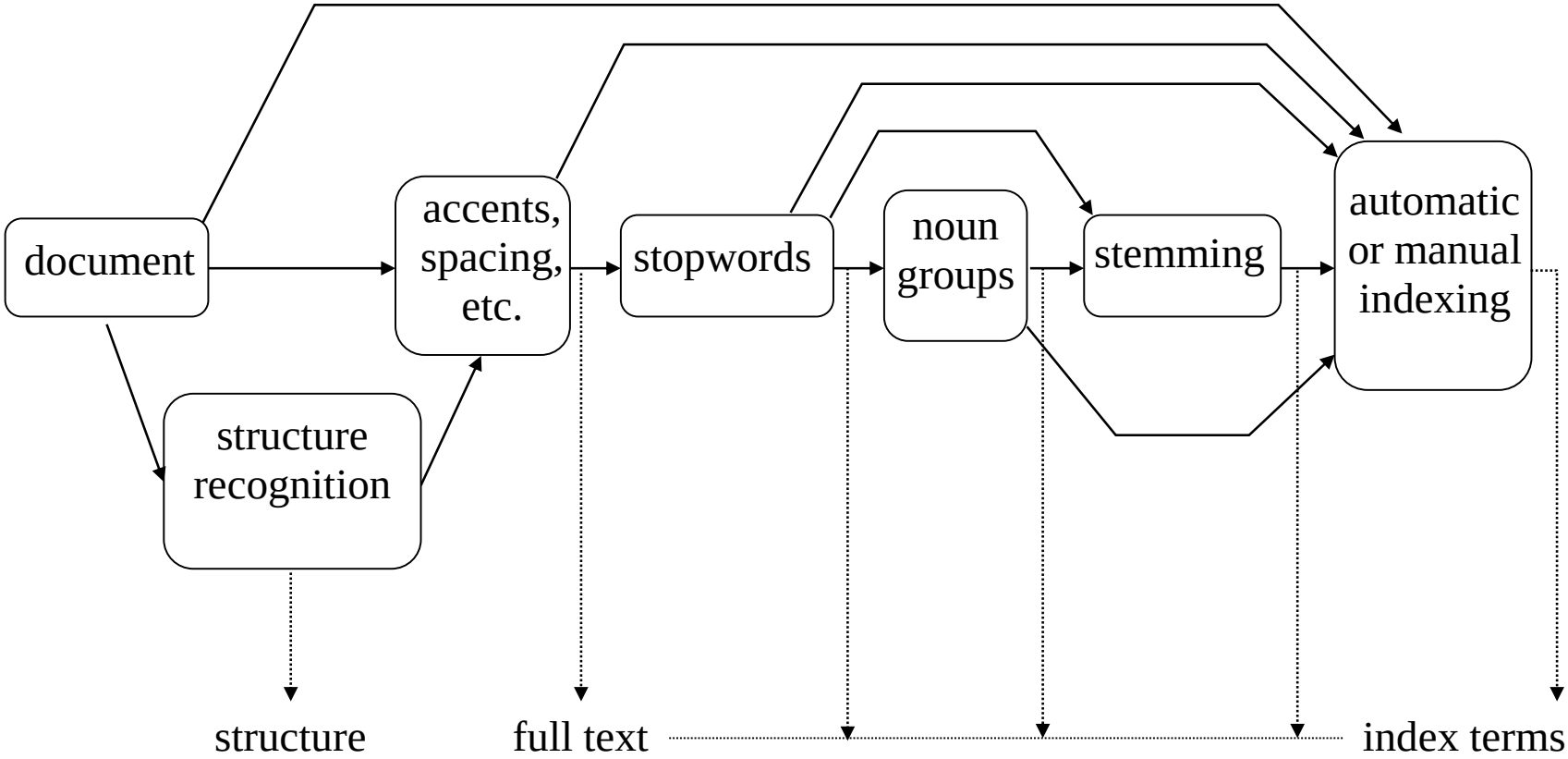
Computerlinguistik – computational linguistics
Lebensversicherungsgesellschaftsangestellter – life insurance company employee

brak ‘spacji’ w językach dalekowschodnich

莎拉波娃现在居住在美国东南部的佛罗里达。今年 4 月
9 日，莎拉波娃在美国第一大城市纽约度过了 1 8 岁生
日。生日派对上，莎拉波娃露出了甜美的微笑。

- dodatkowe mechanizmy segmentacji - wyodrębniające pojedyncze słowa
- wyodrębnianie ciągów *k* znaków zamiast prób identyfikacji słów

Preprocessing



Za: Baeza-Yates & Ribeiro-Neto, 1999

Filtering out stopwords

Słowa które występują w zbyt dużym odsetku dokumentów kolekcji (np. 80%) mogą być bezużyteczne z punktu widzenia wyszukiwania informacji. Słowa takie nie mają potencjału dyskryminacji tematycznej

Listy stopwords (*stop listy*) obejmują najczęściej przedimki, rodzajniki, spójniki, zaimki, liczebniki, oraz niektóre czasowniki, przysłówki i przymiotniki, itp.

Filtering out stopwords

a	are	click	every	here	least	nobody	please	such	thru	whenever
about	around	co	everyone	hereafter	less	none	put	system	thus	where
above	as	com	everything	hereby	like	noone	rather	take	to	whereafter
across	at	con	everywhere	herein	ltd	nor	re	ten	together	whereas
after	away	could	except	hereupon	made	not	s	th	too	whereby
afterwards	back	couldn	few	hers	many	nothing	same	than	top	wherein
again	be	cry	fifteen	herself	may	now	see	that	toward	whereupon
against	became	de	fill	him	me	nowhere	seem	the	towards	wherever
all	because	describe	find	himself	meanwhile	of	seemed	their	translate	whether
almost	become	detail	first	his	might	off	seeming	them	twelve	which
alone	becomes	did	five	how	mill	often	seems	themselves	twenty	while
along	becoming	do	for	however	mine	on	serious	then	two	whither
already	been	does	former	hundred	more	once	several	thence	un	who
also	before	done	formerly	i	moreover	one	she	there	under	whoever
although	beforehand	down	forty	ie	most	only	should	thereafter	until	whole
always	behind	due	found	if	mostly	onto	show	thereby	up	whom
am	being	during	four	in	move	or	side	therefore	upon	whose
among	below	each	from	inc	much	other	since	therein	us	why
amongst	beside	eg	front	indeed	must	others	sincere	thereupon	very	will
amoungst	besides	eight	full	interest	my	otherwise	six	these	via	with
amount	between	either	further	into	myself	our	sixty	they	want	within
an	beyond	eleven	get	is	name	ours	so	thick	was	without
and	both	else	give	it	namely	ourselves	some	thin	we	would
another	bottom	elsewhere	go	its	neither	out	somehow	third	web	yet
any	but	empty	had	itself	never	over	someone	this	well	you
anyhow	by	en	has	keep	nevertheless	own	something	those	were	your
anyone	call	enough	have	la	new	page	sometime	though	what	yours
anything	can	etc	he	last	next	part	sometimes	three	whatever	yourself
anyway	cannot	even	hence	latter	nine	per	somewhere	through	when	yourselves
anywhere	cant	ever	her	latterly	no	perhaps	still	throughout	whence	

Filtering out stopwords

Tradycyjne rozmiary list stopwords dla języka angielskiego to 200-500 pozycji

Dodatkowy zysk związany jest ze znacznym ograniczeniem rozmiaru przechowywanej reprezentacji dokumentów

Ale potencjalne problemy:

“to be or not to be”, „Dr. No”

wyszukiwanie jakichkolwiek fraz i cytatów

Filtering out stopwords

W systemie Carrot dla języka polskiego wykorzystano listę 169 stopwords

Stanowiły one 30% rozważanej kolekcji a zaledwie 0,02% wszystkich unikatowych termów

https://github.com/stopwords-iso/stopwords-pl/blob/master/stopwords-pl.txt
https://pl.wikipedia.org/wiki/Wikipedia:Stopwords

Filtering out stopwords

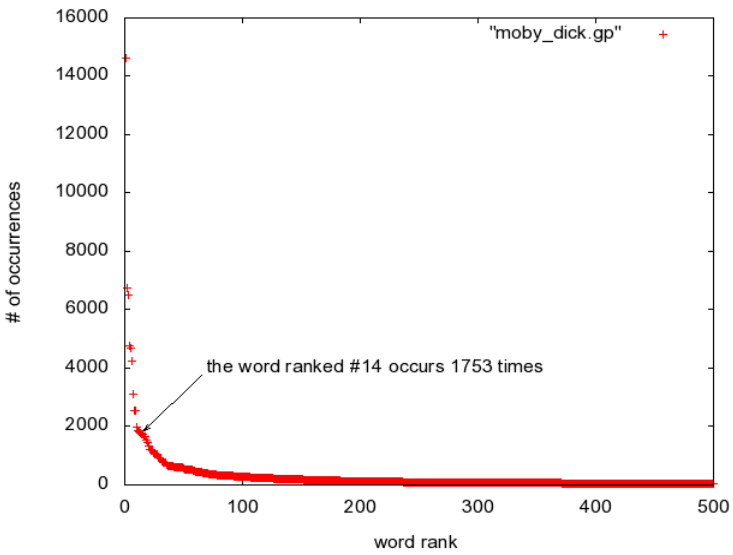
a	będzie	gdzie	jest	lat	nawet	ponad	są	tym
aby	będą	http	jeszcze	lecz	nich	ponieważ	t	tys
albo	chce	i	jeśli	lub	nie	poza	ta	tzw
ale	choć	ich	jeżeli	m	niej	proc	tak	tę
ani	co	im	już	ma	nim	przed	takich	u
aż	coraz	inne	ją	mają	niż	przede	takie	w
b	czy	iż	k	mamy	nowe	przez	także	we
bardzo	czyli	ja	kiedy	miał	np	przy	tam	wie
bez	często	jak	kilku	mimo	nr	r	te	więc
bo	d	jakie	kto	mln	o	raz	tego	wśród
bowiem	dla	jako	która	mogą	od	razie	tej	z
by	do	java	które	może	ok	roku	temu	za
byli	dwie	je	którego	można	on	rz	ten	zaś
bym	dwóch	jeden	której	mu	one	również	też	ze
był	e	jednak	który	musi	oraz	s	to	zł
była	g	jednym	których	na	p	się	trzy	że
było	gdy	jedynie	którym	nad	pl	sobie	tu	żeby
były	gdyby	jego	którzy	nam	po	strona	tych	
być	gdyż	jej	l	nas	pod	swoje	tylko	

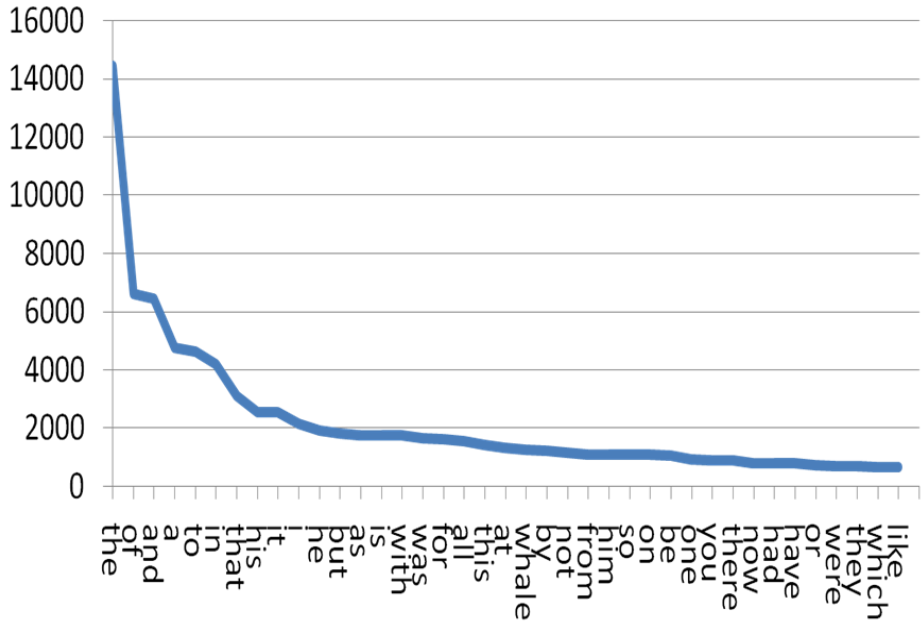
Zipf's law

W I połowie XX w. lingwista George Zipf przebadał typowe częstości słów dla kilku języków naturalnych i we wszystkich stosunkowo mały podzbiór słów był używany znacznie częściej od pozostałych

Na podstawie swych badań spopularyzował znaną wcześniej obserwację, że w reprezentatywnych kolekcjach tekstów najczęstsze słowo danego języka będzie występować 2 razy częściej niż 2. co do częstości, 3 razy częściej niż 3., itd.

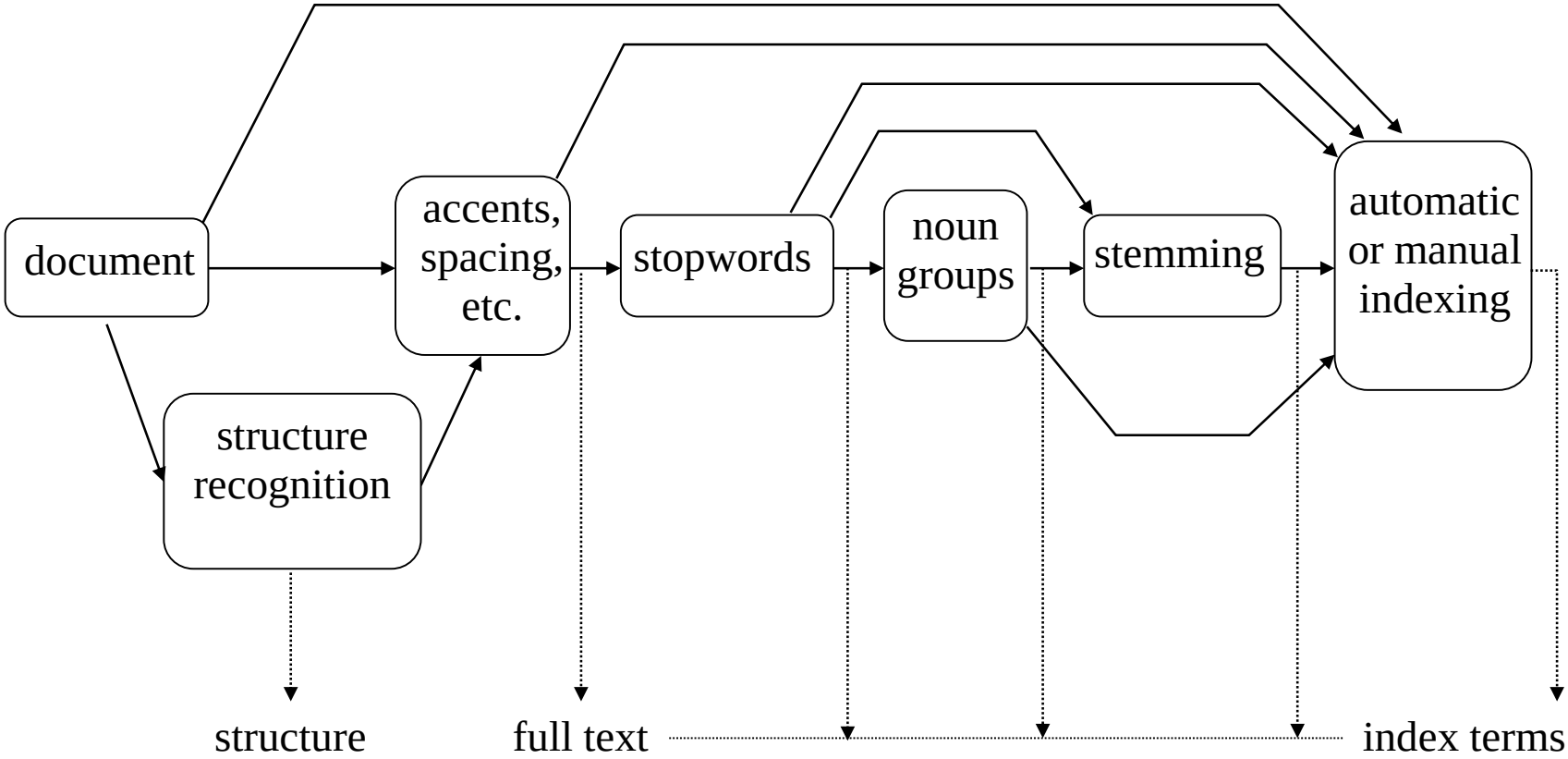
Zipf's law





Źródła: http://www.philippeadjiman.com/blog/2009/10/26/drawing-the-long-tail-of-a-zipf-law-using-gnuplot-java-and-moby-dick/ , http://www.ruwhim.com/?p=47532

Preprocessing



Za: Baeza-Yates & Ribeiro-Neto, 1999

Normalization

- Sprowadzenie różnych sposobów zapisu słowa/pojęcia do tego samego tokenu
- Określenie pewnych klas równoważności lub relacji między niektórymi termami (np. synonimy)
- Powinno się uwzględniać, jak użytkownicy formułują swoje zapytania

U.S.A. USA; naïve naive; colour color; 3/12/91 Mar. 12, 1991 czy 3 Dec 1991?
Chebyshev Tchebycheff; Beijing Pekin; Schütze Schuetze; piesc pięść czy pieść?

- Do reprezentacji dokumentu mogą zostać wprowadzone nowe tokeny (spoza zbioru słów dokumentu)

Stemming

Cele stosowania stemmingu:

- umożliwienie znalezienia tekstów, które zawierają odmiany fleksyjne słów kluczowych zawartych w zapytaniu
- oszczędność przestrzeni reprezentacji dokumentów (nawet o ok. 40%)

Stemming

Ściśle mówiąc:

Lematyzacja – odkrywa *lematy*, tj. formy podstawowe (słownikowe) wyrazów

Stemming (*hasłowanie*) – wyodrębnia *rdzenie* wyrazów

Co więcej, wiele stemmerów nie zapewnia tego, iż tworzone przez nie ciągi liter to rzeczywiste rdzenie wyrazów (lecz pewne *hasła*)

Algorytmy hasłowania w odróżnieniu od lematyzacji nie zwracają (zazwyczaj) informacji gramatycznej i przyjmują na wejściu tylko pojedyncze wyrazy. Mogą natomiast zwracać więcej niż jeden *stem* i zazwyczaj nie są oparte o słowniki

Stemming

Istnieje podział:

Stemmery specjalizowane do zastosowań lingwistycznych – generowane rdzenie powinny rzeczywiście odpowiadać rdzeniom w rozumieniu lingwistyki

Stemmery specjalizowane do zastosowań IR – szybkość działania jest zwykle ich cechą kluczową

Stemming

W języku angielskim wystarczy wziąć pod uwagę:

- czasowniki – czasy
- rzeczowniki – liczba mnoga, forma dzierżawcza

Stemming

W języku angielskim wystarczy wziąć pod uwagę:

- czasowniki – czasy
- rzeczowniki – liczba mnoga, forma dzierżawcza

Czasowniki nieregularne

eat, eats, eating, ate, eaten
catch, catches, catching, caught, caught
cut, cuts, cutting, cut, cut

Rzeczowniki

mouse/mice, goose/geese, ox/oxen
wyjątki i wyjątki od wyjątków: potatoes, tomatoes, photos

Stemming

W języku polskim, zależnie od części mowy, można wyróżnić odmianę przez

- przypadki (tj. mianownik, dopełniacz, celownik, biernik, narzędnik, miejscownik i wołacz)
- rodzaje (męski, żeński, nijaki, (nie)/męskoosobowy)
- liczby (pojedynczą i mnogą)
- osoby (pierwszą: ja, my; drugą: ty, wy; trzecią: on, ona, ono, oni, one)
- czasy (przeszły, teraźniejszy, przyszły)
- tryby (oznajmujący, przypuszczający, rozkazujący)
- strony (czynną, bierną, zwrotną)

Stemming

Odmiana wykazuje wiele nieregularności (*pies, psa; wrzeć, wrę*), a do tego dochodzą specyficzne dla naszego języka utrudnienia, jak *przegłos, itp.* (*latać, lecieć; pleść, plotę, plecie; kwiat, kwiecisty; miód, miodowy*)

Powyższe problemy komplikują zadania wyszukiwania informacji – często nie wystarczają proste heurystyki oparte na zastosowaniu zwykłego dopasowywania wzorców czy wyrażeń regularnych

Stemming

Homonimia występuje na różnych poziomach:

- morfologii (np. „dam” może być formą czasownika „dać” lub dopełniaczem liczby mnogiej rzeczownika „dama”)
- słownictwa (np. „rola”, która może być aktorską lub uprawną)
- składni (np. „zdrada przyjaciela” może oznaczać zarówno fakt, że przyjaciel zdradził, jak i to, że został zdradzony)

Stemming

Homonimia występuje na różnych poziomach:

- morfologii (np. „dam” może być formą czasownika „dać” lub dopełniaczem liczby mnogiej rzeczownika „dama”)
- słownictwa (np. „rola”, która może być aktorską lub uprawną)
- składni (np. „zdrada przyjaciela” może oznaczać zarówno fakt, że przyjaciel zdradził, jak i to, że został zdradzony)

Polisemia natomiast dotyczy wieloznaczności, tj. wielu znaczeń tego samego leksemu, a nie różnych leksemów („zamek” w drzwiach, a w spodniach)

Stemming

Homonimia, polisemia, itp.

Gdy takie zjawiska występują to, aby jednoznacznie określić leksem danego wyrazu, należy sięgnąć do *kontekstu*, w jakim ten wyraz się znajduje (np. do innych wyrazów w zdaniu lub wypowiedzi)

Zajmuje się tym *analiza składniowa* nie będąca przedmiotem typowych rozwiązań IR

Stemming

Problem stanowią też jednostki wielowyrazowe (np. *będę pracował*) będące formą jednego wyrazu, których wykrywanie jest często zbyt kosztowne obliczeniowo

W związku z tym hasłowanie wykonywane jest zasadniczo na pojedynczych wyrazach

Stemming

Możliwe strategie obejmują:

- usuwanie przyrostków
- wykorzystanie słowników
- zaawansowane techniki lingwistyczne
- wykorzystanie tzw. *k*-gramów

Jako *k*-gramy należy tu rozumieć podciągi *k*-literowe pochodzące od wyrazów, gdzie *k* jest liczbą naturalną

Przykładowo wszystkie *k*-gramy wyrazu „*samolot*” dla $k = 5$ to: „*samol*”, „*amolo*”, „*molot*”

Stemming

Dla języka angielskiego najpopularniejsze stemmery to

- Porter stemmer
- Lovins stemmer
- Paice stemmer

Wszystkie one wykorzystują ideę automatów skończonych

Bazują na technikach usuwania prefiksów/sufiksów

W j. ang. wyróżnia się zaledwie ok. 75 prefiksów i 250 sufiksów

Stemming

Pierwszy skuteczny algorytm dla angielskiego - **Lovin's stemmer** (1968) – stemmer jednoprzebiegowy, wykorzystujący tablicę 250 możliwych podstawień końcówek oraz dodatkowy etap postprocessingu – był projektowany jako uniwersalny

Najpopularniejszy stemmer – **Porter's stemmer** (1971), specjalizowany dla IR, wieloprzebiegowy, nie generuje poprawnych językowo rdzeni

Inne: **Krovets** (1993) – trójprzebiegowy stemmer wyłącznie fleksyjny, **Dawson** (1974) – poprawiona wersja stemmera Lovins, zawiera tablicę 1200 podstawień, **Lancaster** aka **Paice/Husk** (1990) – stemmer oparty na dopasowywaniu reguł, nie ma ograniczenia na liczbę kroków podstawień

Stemming

Sample text: Such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation

Lovins stemmer: such an analys can reve featur that ar not eas vis from th vari in th individu gen and can lead to a pictur of expres that is mor biolog transpar and acces to interpres

Porter stemmer: such an analysi can reveal featur that ar not easili visibl from the variat in the individu gene and can lead to a pictur of express that is more biolog transpar and access to interpret

Paice stemmer: such an analys can rev feat that are not easy vis from the vary in the individ gen and can lead to a pict of express that is mor biolog transp and access to interpret

http://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

Porter stemmer

Pomysł **Portera** polegał na tym, żeby formom odmienionym zamieniać końcówki (ang. *suffix*)

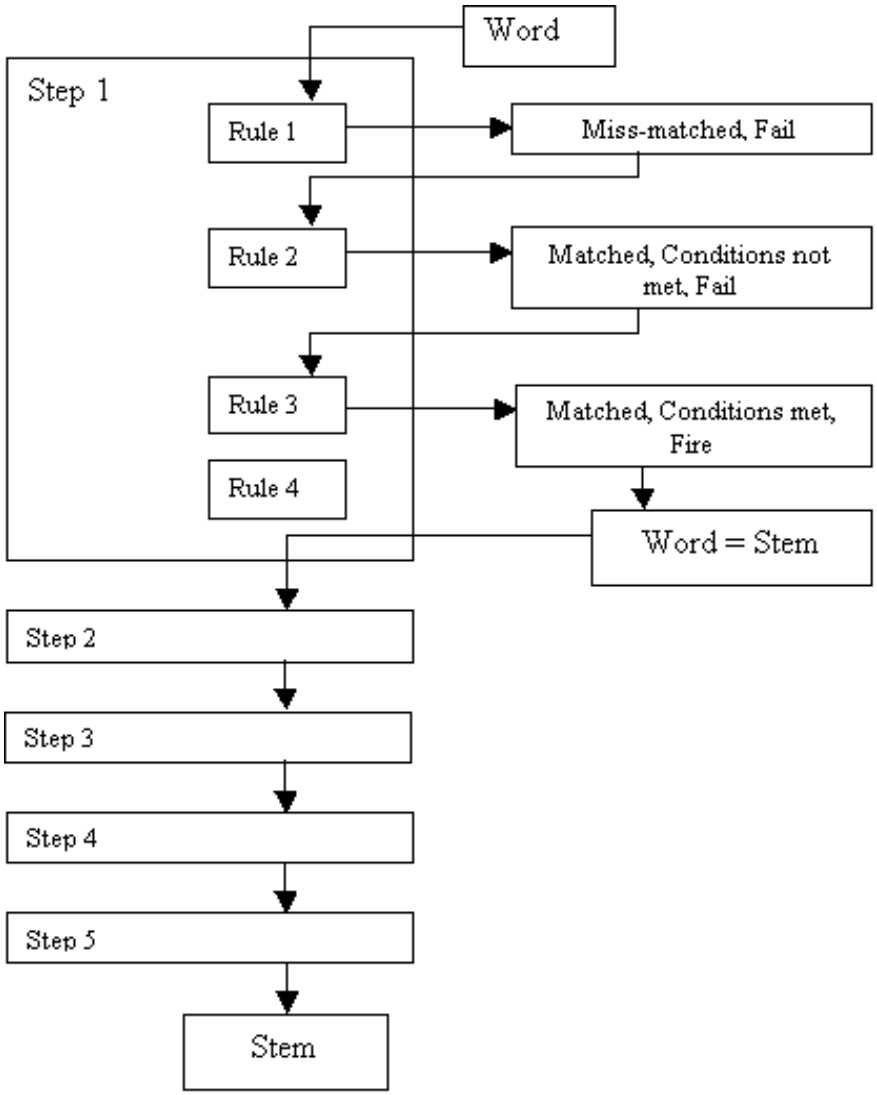
Dodatkowo dokonuje się to wieloprzebiegowo – możliwe jest wiele następujących po sobie zamian przyrostków dla jednego słowa

Operacje zostały szczegółowo zdefiniowane i opisane – zarówno rodzaj zamiany, warunki zajścia tej zamiany, jak i kolejność możliwych zamian

Zmianami są to takie operacje, jak np.:

- obcięcie końcówki [’], [’s] – np. w wyrazie *students’* lub *student ‘s*
- zamiana końcówki [ied] na końcówkę [i] (jeśli przed [ied] jest więcej niż jedna litera) lub [ie] (w przeciwnym razie) – np. w wyrazach „*cried*” $\rightarrow$ „*cri*”, „*tied*” $\rightarrow$ „*tie*”

Porter stemmer



Za: http://www.comp.lancs.ac.uk/

Porter stemmer

http://snowball.tartarus.org/algorithms/porter/stemmer.html

```
m will be called the \measure\ of any word or word part

m=0    TR,  EE,  TREE,  Y,  BY.
m=1    TROUBLE,  OATS,  TREES,  IVY.
m=2    TROUBLES,  PRIVATE,  OATEN,  ORRERY.


*S  - the stem ends with S (and similarly for the other letters).

*v* - the stem contains a vowel.

*d  - the stem ends with a double consonant (e.g. -TT, -SS).

*o  - the stem ends cvc, where the second c is not W, X or Y (e.g.
      -WIL, -HOP).
```

Porter stemmer

Step 1 deals with plurals and past participles.

Step 1a

SSES	-> SS	caresses	-> caress
IES	-> I	ponies	-> poni
		ties	-> ti
SS	-> SS	caress	-> caress
S	->	cats	-> cat

Step 1b

(m>0) EED	-> EE	feed	-> feed
		agreed	-> agree
(*v*) ED	->	plastered	-> plaster
		bled	-> bled
(*v*) ING	->	motoring	-> motor
		sing	-> sing

Porter stemmer

If the second or third of the rules in Step 1b is successful, the following is done:

AT -> ATE	conflat(ed)	->	conflate
BL -> BLE	troubl(ed)	->	trouble
IZ -> IZE	siz(ed)	->	size
(*d and not (*L or *S or *Z)) -> single letter			
	hopp(ing)	->	hop
	tann(ed)	->	tan
	fall(ing)	->	fall
	hiss(ing)	->	hiss
	fizz(ed)	->	fizz
(m=1 and *o) -> E	fail(ing)	->	fail
	fil(ing)	->	file

Step 1c

(*v*) Y -> I	happy	->	happi
	sky	->	sky

Porter stemmer

Step 2

(m>0)	ATIONAL	-> ATE	relational	->	relate
(m>0)	TIONAL	-> TION	conditional	->	condition
			rational	->	rational
(m>0)	ENCI	-> ENCE	valenci	->	valence
(m>0)	ANCI	-> ANCE	hesitanci	->	hesitance
(m>0)	IZER	-> IZE	digitizer	->	digitize
(m>0)	ABLI	-> ABLE	conformabli	->	conformable
(m>0)	ALLI	-> AL	radicalli	->	radical
(m>0)	ENTLI	-> ENT	differentli	->	different
(m>0)	ELI	-> E	vileli	- >	vile
(m>0)	OUSLI	-> OUS	analogousli	->	analogous
(m>0)	IZATION	-> IZE	vietnamization	->	vietnamize
(m>0)	ATION	-> ATE	predication	->	predicate
(m>0)	ATOR	-> ATE	operator	->	operate
(m>0)	ALISM	-> AL	feudalism	->	feudal
(m>0)	IVENESS	-> IVE	decisiveness	->	decisive
(m>0)	FULNESS	-> FUL	hopefulness	->	hopeful
(m>0)	OUSNESS	-> OUS	callousness	->	callous
(m>0)	ALITI	-> AL	formaliti	->	formal
(m>0)	IVITI	-> IVE	sensitiviti	->	sensitive
(m>0)	BILITI	-> BLE	sensibiliti	->	sensible

Porter stemmer

Step 3

(m>0)	ICATE	-> IC	triplicate	->	triplic
(m>0)	ATIVE	->	formative	->	form
(m>0)	ALIZE	-> AL	formalize	->	formal
(m>0)	ICITI	-> IC	electriciti	->	electric
(m>0)	ICAL	-> IC	electrical	->	electric
(m>0)	FUL	->	hopeful	->	hope
(m>0)	NESS	->	goodness	->	good

Porter stemmer

Step 4

(m>1) AL ->	revival	->	reviv
(m>1) ANCE ->	allowance	->	allow
(m>1) ENCE ->	inference	->	infer
(m>1) ER ->	airliner	->	airlin
(m>1) IC ->	gyroscopic	->	gyroscop
(m>1) ABLE ->	adjustable	->	adjust
(m>1) IBLE ->	defensible	->	defens
(m>1) ANT ->	irritant	->	irrit
(m>1) EMENT ->	replacement	->	replac
(m>1) MENT ->	adjustment	->	adjust
(m>1) ENT ->	dependent	->	depend
(m>1 and (*S or *T)) ION ->	adoption	->	adopt
(m>1) OU ->	homologou	->	homolog
(m>1) ISM ->	communism	->	commun
(m>1) ATE ->	activate	->	activ
(m>1) ITI ->	angulariti	->	angular
(m>1) OUS ->	homologous	->	homolog
(m>1) IVE ->	effective	->	effect
(m>1) IZE ->	bowdlerize	->	bowdler

Porter stemmer

Postprocessing

Step 5a

(m>1) E	->	probate	->	probat
		rate	->	rate
(m=1 and not *o) E	->	cease	->	ceas

Step 5b

(m > 1 and *d and *L) -> single letter

controll	->	control
roll	->	roll

Complex suffixes are removed bit by bit in the different steps. Thus GENERALIZATIONS is stripped to GENERALIZATION (Step 1), then to GENERALIZE (Step 2), then to GENERAL (Step 3), and then to GENER (Step 4). OSCILLATORS is stripped to OSCILLATOR (Step 1), then to OSCILLATE (Step 2), then to OSCILL (Step 4), and then to OSCIL (Step 5).

Porter stemmer

```
Suffix stripping of a vocabulary of 10,000 words
------------------------------------------------
Number of words reduced in step 1:    3597
                "                2:     766
                "                3:     327
                "                4:    2424
                "                5:    1373
Number of words not reduced:          3650

The resulting vocabulary of stems contained 6370 distinct entries.
Thus the suffix stripping process reduced the size of the vocabulary
by about one third.
```

Podejścia słownikowe

Wystarczy utrzymywać słownik zawierający dpowiednio dużą liczbę wyrazów danego języka i sprawdzać w nim od jakiego leksemu pochodzi dana forma odmieniona (ale możliwe *wielokrotne* dopasowania)

Podejścia heurystyczne

Opierają się na jakiejś transformacji algorytmicznej, zwracającej mniej lub bardziej unikalny token (ten sam dla wszystkich form danego leksemu)

Czasem tym tokenem może być lemat danego wyrazu, czasem rdzeń, czy stem lub hasło.

Algorytmy te nie weryfikują własnych wyników przez skonfrontowanie ich ze słownikiem poprawnych form

Podejścia heurystyczne

Przykładowymi rodzajami elementarnych operacji pozwalającymi na transformację odmienionego wyrazu mogą być:

- odcinanie końcówek i/lub przedrostków
- zamiana(y) końcówek i/lub przedrostków
- operacje na pojedynczych znakach (zamiana, wstawienie, usunięcie)
- operacje na większych jednostkach (sylaby, ciągi kilkuliterowe np. 2-gramy, 3-gramy itd.)

Podejścia mieszane (hybrydowe)

Algorytmy te charakteryzuje możliwość korzystania ze słownika z jednoczesnym stosowaniem reguł transformacji, które szacują lematy dla słów spoza tego słownika

Jeśli tylko algorytm stosujący podejście słownikowe dokonuje jakiegoś szacowania dla nieznanych mu form wyrazowych (spoza słownika) to można go nazwać hybrydowym

Algorytmy słownikowe mogą zapewniać poprawność wyników – bazując na słowniku poprawnych form

Jednak słownik nie musi być słownikiem poprawnych form (lematów). Może zawierać informacje, które tylko te lematy opisują, np. zbiór możliwych czteroliterowych ciągów, z których mogą składać się lematy

Algorytmy regułowe i hybrydowe dają wyniki przybliżone (pewne jedynie dla zwróconych wyrazów, których potwierdzenie znajdują w słowniku, z którego korzystają)

Narzędzia dla języka polskiego

Lametyzator Dawida Weissa — oprócz lematów zwraca on również informację gramatyczną (dla wyrazów ze słownika Ispell, które taką informację mają)

Stempel Andrzeja Białeckiego — algorytm regułowy

Stempelator Dawida Weissa — połączenie Lametyzatora z algorytmem Stempel (algorytm hybrydowy)

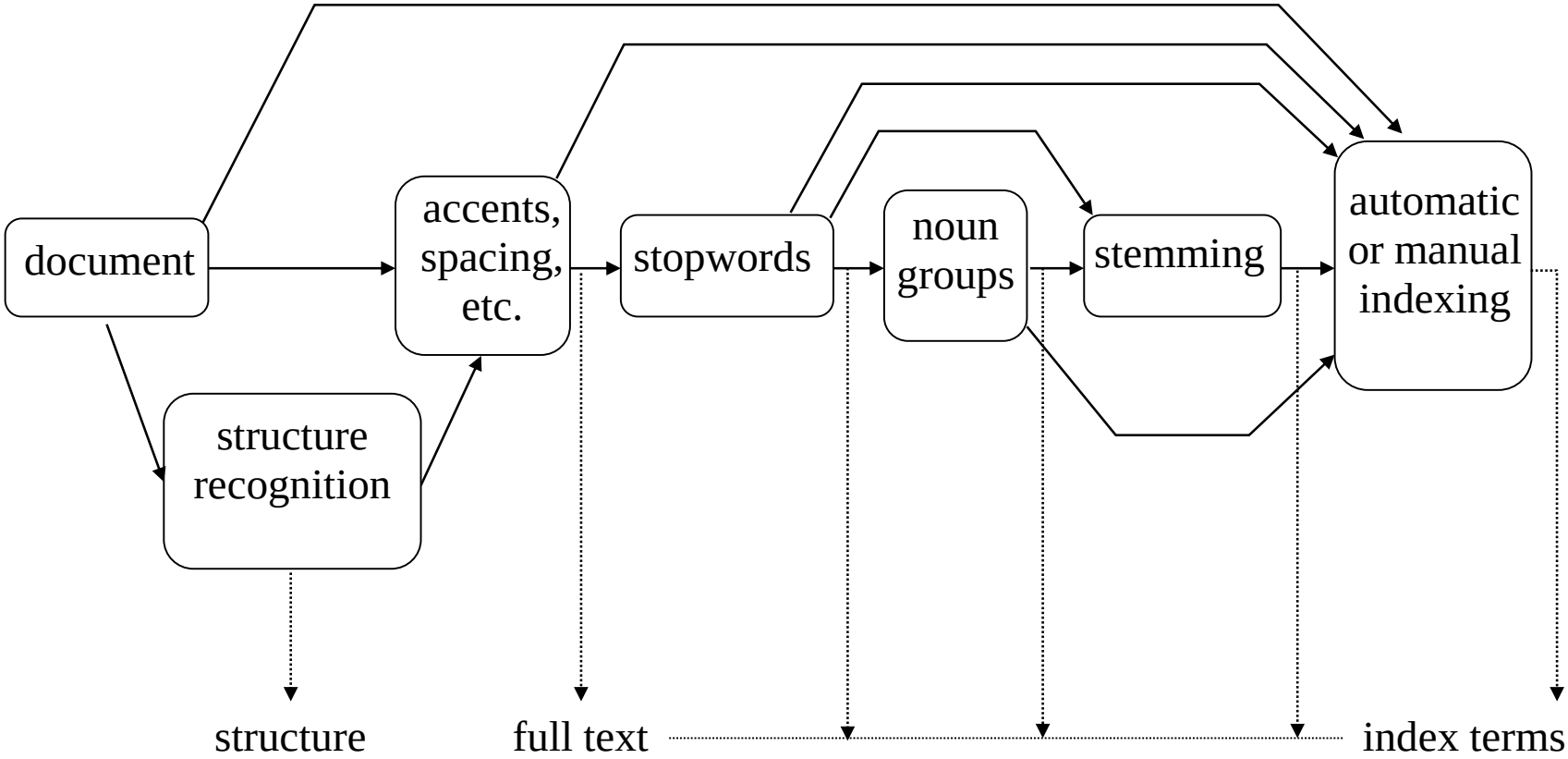
Morfeusz (analizator morfologiczny) Marcina Wolińskiego Oferuje słownikowe metody hasłowania i oznaczania części mowy (ang. POS-Tagging)

SAM-95 (analizator morfologiczny) Krzysztofa Szafrana

TaKIPI Macieja Piaseckiego, Grzegorza Godlewskiego, Adama Radziszewskiego, Bartosza Brody i Adama Wardyńskiego Przyporządkowuje opis morfo-syntaktyczny do wyrazów w zdaniu i określa znaczenia poszczególnych wyrazów

LEMOT Tomasza Dragosza — algorytm hybrydowy stworzony na potrzeby lematyzacji słów pozasłownikowych – praca mgr. w IIn PP

Preprocessing



Za: Baeza-Yates & Ribeiro-Neto, 1999

Final vocabulary

Ostateczny wybór jednostek indeksujących może być automatyczny lub półautomatyczny (z wykorzystaniem dodatkowych słowników/taksonomii)

Przykładowo automatyczna selekcja np. rzeczowników lub fraz rzeczownikowych w celu wyłonienia jednostek indeksujących o jak „największej semantyce”

W przypadku fraz jednostki indeksujące przestają być pojedynczymi słowami kluczowymi. Frazy rzeczownikowe mogą być identyfikowane automatycznie jako zbiory rzeczowników oddalonych od siebie w tekście o nie więcej niż zadaną liczbę pozycji (np. 3) lub z wykorzystaniem informacji lingwistycznej

Potencjalny problem: konieczność kosztownej analizy

POS—tagging (ang. Part Of Speech tagging), np.:
N - proper nouns; X - function words including articles and prepositions
extended biword - any string of terms of the form NX*N
‘renegotiation of the constitution’ → ‘renegotiation constitution’
N X X N NN

Final vocabulary

Metody opracowania słowników:

- Słowniki ‘ręcznie’ utrzymywane przez edytorów
  MeSH - Medical Subject Headings
  https://www.nlm.nih.gov/pubs/factsheets/mesh.html
- WordNet – *http://wordnet.princeton.edu*, itp.
- Słowniki tworzone automatycznie (na zasadzie obserwacji współwystępowania słów)
- Analiza zapytań użytkowników wyszukiwarek (*query log mining*)

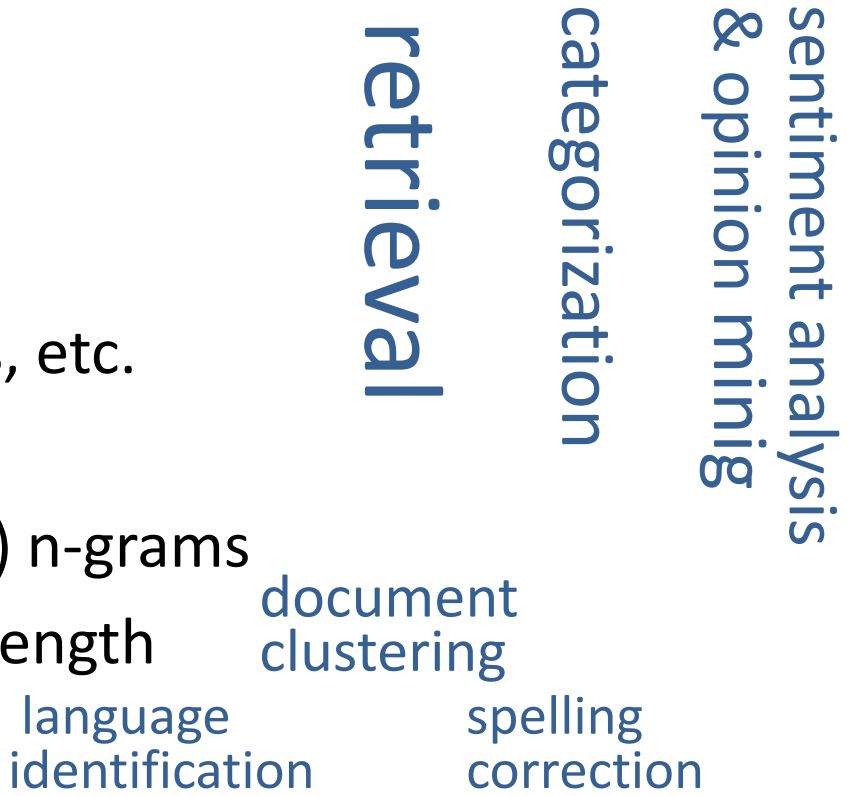
Representation

- Index terms:
  - Keywords
  - Bigrams
  - Word n-grams
  - Nouns, noun groups, etc.
  - Negation n-grams
  - POS (part of speech) n-grams
  - Phrases of variable length
  - Character k-grams

  ...
  - Word vectors (num. vector representations of words)

Representation

- Index terms:
  - Keywords
  - Bigrams
  - Word n-grams
  - Nouns, noun groups, etc.
  - Negation n-grams
  - POS (part of speech) n-grams
  - Phrases of variable length
  - Character k-grams
  - …
  - Word vectors (num. vector representations of words)



Podstawowe sposoby reprezentacji dokumentów tekstowych:

- Boolean representation (BIN - binary vectors)
  zachowuje informację jedynie o obecności lub nieobecności danego termu w danym dokumencie
- Bag-of-words representations (np. TF, TF-IDF)
  zachowują informację o liczbie wystąpień danego termu w danym dokumencie
- Reprezentacja pełna

  zachowuje informację o liczbie wystąpień danego termu w danym dokumencie wraz z miejscami wystąpień danego termu

Podstawowe sposoby reprezentacji dokumentów tekstowych:

- Boolean representation (BIN - binary vectors)
  $d_1 = (1, 1, 1, 1)$
  $d_2 = (1, 1, 0, 1)$
- Bag-of-words representations (np. TF, TF-IDF)
  $d_1 = (1, 1, 2, 2)$
  $d_2 = (1, 1, 0, 1)$

  d_1: Paul is quicker than John. Paul is quicker than George, too.
  d_2: John is quicker than George.

  index terms: george, john, paul, quicker

  a jak zapisać dokument d_3: George is quicker than John ?

Jaccard Coefficient

Do wyznaczania podobieństwa między dwoma dokumentami w binarnej reprezentacji BIN można stosować współczynnik (indeks) Jaccarda

Jest to statystyka używana do porównywania zbiorów. Współczynnik Jaccarda mierzy podobieństwo między dwoma zbiorami i jest zdefiniowany jako iloraz mocy części wspólnej zbiorów i mocy sumy tych zbiorów

W kontekście zbioru termów występujących w dokumencie $T(dj)$

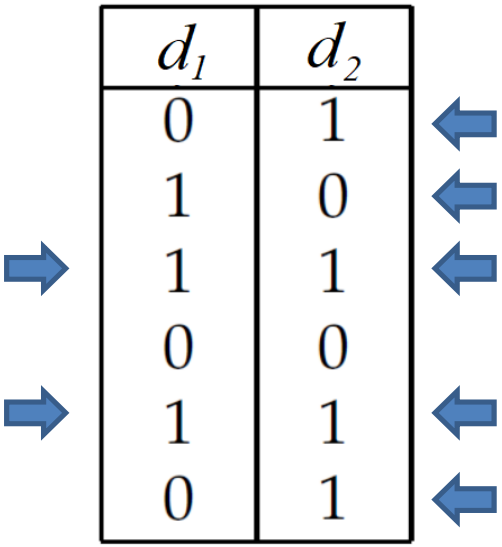
$$sim(d_1, d_2) = \frac{|T(d_1) \cap T(d_2)|}{|T(d_1) \cup T(d_2)|}$$

Jaccard Coefficient

d_1	d_2
0	1
1	0
1	1
0	0
1	1
0	1

$$sim(d_1, d_2) = \frac{|T(d_1) \cap T(d_2)|}{|T(d_1) \cup T(d_2)|}$$

Jaccard Coefficient



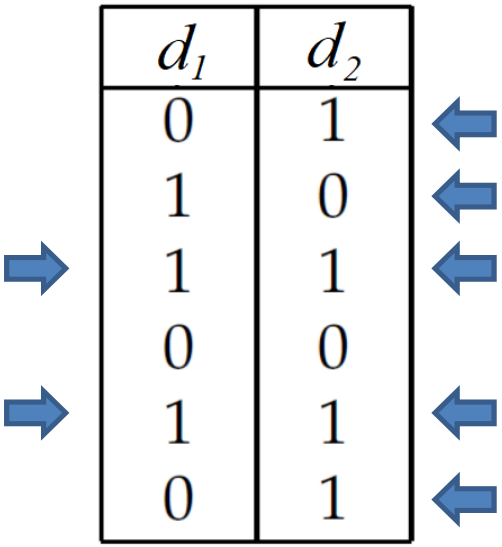
d_1	d_2
0	1
1	0
1	1
0	0
1	1
0	1

$$sim(d_1, d_2) = \frac{|T(d_1) \cap T(d_2)|}{|T(d_1) \cup T(d_2)|}$$

$$= 2/5$$

- Rozważa tylko niezerowe współrzędne wektorów
- Wystarczy porównywać pary jedynie tych dokumentów, które posiadają co najmniej jedną wspólną niezerową współrzędną

Jaccard Coefficient



d_1	d_2
0	1
1	0
1	1
0	0
1	1
0	1

$$sim(d_1, d_2) = \frac{|T(d_1) \cap T(d_2)|}{|T(d_1) \cup T(d_2)|}$$

$$= 2/5$$

- Rozważa tylko niezerowe współrzędne wektorów
- Wystarczy porównywać pary jedynie tych dokumentów, które posiadają co najmniej jedną wspólną niezerową współrzędną

Stosowany m.in. w wykrywaniu *prawie*-duplikatów (plagiatów) algorytmem *W-shingling* wykorzystującym reprezentację dokumentów opartą na n-gramach, czy w korekcji literówek przy wykorzystaniu reprezentacji słów opartej na k-gramach

Information Retrieval models

Klasyczne modele IR

- model Boole'owski
- model wektorowy (VSM – Vector Space Model)
- model probabilistyczny (BIR – Binary Independence Model)

Nieklasyczne modele IR

- model oparty na zbiorach rozmytych (fuzzy sets)
- rozszerzony model Boole'owski
- model LSI (Latent Semantic Indexing)
- model oparty na sieciach neuronowych (NN)
- uogólniony model wektorowy (Generalized VSM)
- nieklasyczne modele probabilistyczne (sieci Bayesowskie, belief networks, inference networks ...)

...

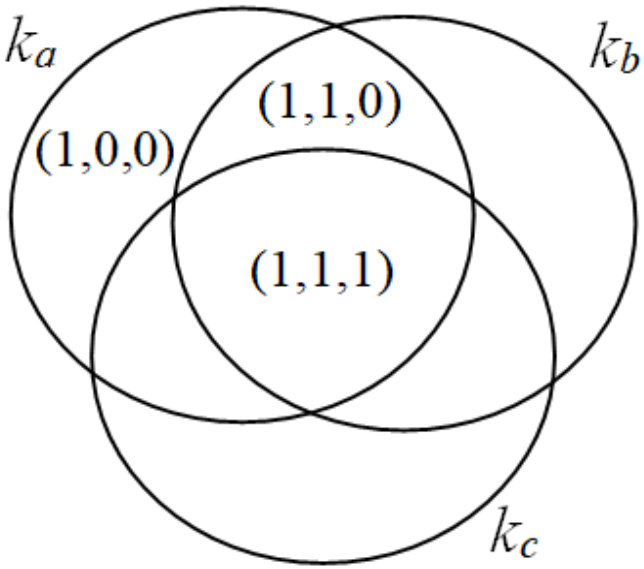
Information Retrieval models

- Niech T będzie liczbą termów indeksujących (rozmiarem słownika), k_i – i-tym słowem kluczowym, D – kolekcją, czyli zbiorem wszystkich dostępnych dokumentów, Q – zbiorem zapytań
- $K=\{ k_1, k_2, ..., k_T \}$ jest zbiorem wszystkich termów indeksujących
- Z każdym słowem kluczowym k_i dokumentu d_j związana jest waga $a_{ij} > 0$ (ew. $a_{ij} \geq 0$), dla słów kluczowych niewystępujących w tekście dokumentu $a_{ij} = 0$
- Stąd każdemu dokumentowi przyporządkowany jest wektor $d_j = (a_{1j}, a_{2j}, ..., a_{Tj})$
- Niech g_i będzie funkcją, która zwraca wagę związaną ze słowem kluczowym k_i dowolnego T-wymiarowego wektora, np.: $g_i(d_j) = a_{ij}$
- $sim(q, d_j)$ jest funkcją rangującą, która przyporządkowuje wartości rzeczywiste parom (q, d_j): $q \in Q$, $d_j \in \mathrm{D}$
- Funkcja sim definiuje uporządkowanie (ranking) dokumentów względem zapytania

Boolean model

- oparty na teorii zbiorów i algebrze Boole'a
- dokumenty i zapytania w postaci binarnych wektorów
- zapytania – wyrażenia boole'owskie o precyzyjnej semantyce (reprezentowalne w postaci DNF)
- binarna decyzja dotycząca adekwatności (brak funkcji rangującej)
- raczej data retrieval niż information retrieval (*exact match model)*
- obserowowano trudności użytkowników w wyrażeniu ich zapotrzebowań informacyjnych w postaci wyrażeń boole'owskich

Boolean model



$$q = k_a \wedge (k_b \vee \neg\, k_c)$$

$$q_{dnf} = (1,1,1) \vee (1,1,0) \vee (1,0,0)$$

ogólnie:

$$q_{dnf} = cc_1 \vee cc_2 \vee \ldots \vee cc_p$$

$$sim(d_j, q) = \begin{cases} 1 & if\ \exists cc_l \mid (cc_l \in q_{dnf}) \wedge (\forall k_i\ g_i(d_j) = g_i(cc_l)) \\ 0 & w\ przeciwnym\ wypadku \end{cases}$$

Boolean model

- Zalety modelu boole'owskiego
  - ☺ prostota i szybkość
  - ☺ dobre umotywowanie formalne
- Wady modelu boole'owskiego
  - ☹ ignorowanie informacji nt. częstości termów
  - ☹ dokładne dopasowanie dokumentów do zapytania często prowadzi do zbyt małych (np. pustych) lub zbyt dużych zbiorów wyników
  - ☹ brak funkcji rangującej powoduje, że odpowiedź systemu będzie często bezużyteczna (not manageable)

Exercises

I. Najczęstsze słowo „*i*” w pewnej dużej kolekcji polskojęzycznych tekstów wystąpiło około 120 tys. razy, trzecie co do częstości słowo w tej kolekcji to „*się*” – ilu wystąpień tego słowa można się spodziewać – zgodnie z prawem Zipfa?

II. Wyobraź sobie prosty stemmer, który dane słowo przekształca na pojedynczy *k-gram* obejmujący początkowe znaki słowa. Podaj efekt zastosowania takiej heurystyki na poniższym tekście przyjmując k=4.

„dawno temu w odległej galaktyce”

III. Dla podanej kolekcji 3 krótkich dokumentów a) uzupełnij ich reprezentację binarną b) podaj odpowiedź na zapytanie q w opraciu o model boole'owski

d_1: ala ma kota (1, 0, 1, 1)
d_2: kota ma alan (0, 1, 1, 1)
d_3: ala ma kota ma (?, ?, ?, ?)
q: ma $\wedge$ kota $\vee \neg$ alan