Raport
Instytutu Automatyki i Informatyki Stosowanej
Politechniki Warszawskiej

Możliwości tworzenia aplikacji analizujących dane z wielu publicznie dostępnych źródeł polskich i zagranicznych

Mariusz Kamola

E-mail: M.Kamola@ia.pw.edu.pl

Praca naukowa nr: 2015-17

Warszawa, listopad 2015

Obfitość danych, wynikająca ze wzrostu liczby urządzeń podłączonych do globalnej sieci, stawia nowe wyzwania dla technologii ich gromadzenia, ale przede wszystkim ich przetwarzania. Stwarza też nowe możliwości łączenia danych, ujawniające ciekawe, nieoczywiste zależności. W sposób szczególny, ze względu na szeroki wachlarz zastosowań, interesujące są cyfrowe ślady aktywności społecznej ludzi – obserwowane wprost, albo poprzez informacje z otoczenia, w którym żyją.

Celem projektów studenckich realizowanych w ramach przedmiotu „Techniki Analizy Sieci Społecznych" jest tworzenie aplikacji lub analiz wydobywających dodatkową wiedzę z połączenia ogólnodostępnych danych. Zestaw proponowanych tematów projektowych został dobrany tak, aby studenci mieli możliwość zetknięcia się z całym wachlarzem problemów związanych z pozyskiwaniem i łączeniem danych heterogenicznych, takich jak:

- brak interfejsu programistycznego do danych (API),
- ograniczenia przepustowości kanału dostępu do danych,
- niespójność i zanieczyszczenie danych,
- niejednoznaczność złączenia danych z dwóch i więcej źródeł,
- niezgodność API z dokumentacją,
- ilość danych utrudniająca lub uniemożliwiająca ich analizę i wizualizację za pomocą standardowych procedur.

Studentom pozostawiono dowolność w doborze technologii, stawiając za priorytet przezwyciężenie problemów i dostarczenie wyników choćby w minimalnym zakresie.

W dalszym ciągu zostaną zaprezentowane projekty i ich rozwiązania, w ustandaryzowanej formie. Każdy opis charakteryzuje się:

- tematem zadania,
- listą wykorzystanych źródeł danych i sposobem ich pozyskania,
- sposobem połączenia danych z różnych źródeł, albo wewnętrznego złączenia danych z jednego źródła,
- opisem implementacji, zawierającym nazwy użytych kluczowych technologii, rozwiązań autorskich, odstępstw od zakresu zadania oraz problemów związanych ze źródłami danych bądź z samymi danymi,
- możliwościami dalszego wykorzystania wyników projektu,
- najciekawszymi wynikami projektu, najczęściej – wizualnymi.

Przedstawione projekty dzielą się na dwie zasadnicze kategorie, ze względu na ich produkt.

1. Projekty dostarczające aplikacji interaktywnych, zasadniczo pobierających i analizujących niezbędne dane na bieżąco, a zatem, siłą rzeczy, w ograniczonym zakresie.
2. Projekty analityczne, łączące i przetwarzające jednorazowo wszystkie pobrane dane, w celu znalezienia odpowiedzi na pewne zagadnienia badawcze.

W dalszej części raportu, formularze dotyczące obu typów projektów zostały wyróżnione kolorami: granatowym dla projektów aplikacyjnych, a brązowym – dla analitycznych.

1 Weryfikacja popularności doniesień na podstawie alternatywnego serwisu społecznościowego

Źródła danych

wykop.pl × twitter

Doniesienia użytkowników o wydarzeniach	Serwis społecznościowy; krótkie informacje

Pozyskanie danych

Analiza HTML	API, max. 100 wiadomości naraz

Złączenie danych

URL źródła wiadomości

Uwagi implementacyjne

Przyjęto, że współczynnik popularności wiadomości jest ważoną sumą popularności w Wykopie (liczba wykopań) oraz liczby retransmisji (retweet) na Twitterze. Wykonano aplikację sieciową w Javie.

Możliwe zastosowania

Stworzenie usługi agregatora doniesień o wydarzeniach, z wewnętrzną weryfikacją i rankingiem popularności wydarzeń.

Ilustracje działania

Title	Url	Author	Author gender	Total retweets	Total favourites	on Wykop	Popularity coeff.
Nie żyje Christopher Lee odtwórca roli Saurumana we "Władcy Pierścieni"	http://www.telegraph.co.uk/news/celebritynews/11666316/christopher-lee-dies-live.html	Opisarmoli_dlhydtoohbrdum	male	501395	3827	2187	389722.1704
Amerykanie mają ból dupy o to, że w Wiedźmine 3 nie ma murzynów [ENG]	http://www.polygon.com/2015/5/3/8719389/colorblind-on-witcher-3-rust-and-gaming's-race-problem	wiecejniizjednozwierze	female	13613	34	3405	11329.4704
Sprowadźmy chrześcijan zamiast 2 tys. muzułmanów!	http://www.youtube.com/watch?v=gyFce8HW3Ok	Unifikalizacja		0	0	6911	1545.9907
Real GTA	http://www.youtube.com/watch?v=0ZZquVyLEo	thuia	male	930	15	2183	1210.2961
Bardzo trafna opinia na temat przyjmowania imigrantów z Afryki do Polski	https://www.youtube.com/attribution_link?a=jAnDjMYLc8u=/watch?v=X5UCJHjsaA88feature=share	KenZoo	male	0	0	5039	1127.2243
Poważne włamanie do polskiego banku, skradzione dane, pieniądze i hasła klientów	http://zaufanatrzeciastrona.pl/postpowazne-wamanie-do-polskiego-banku-skradzione-dane-i-hasla-klientow/	mamypoops	male	633	29	2120	965.6419000000000
Słonega zaczął	http://www.wykop.pl/wpis/13042415/slonega-przecieki/	geekmaster	male	0	0	4183	935.7371

2 Weryfikacja rzetelności opisów znalezisk na podstawie porównania tekstu streszczenia z oryginalnym.

Źródła danych

wykop.pl × www.

Doniesienia użytkowników o wydarzeniach	Oryginalne strony www z wydarzeniami

Pozyskanie danych

Makieta ReST serwisu wykop.pl	Analiza HTML

Złączenie danych

Odległość Lewensztejna oraz podobieństwo cosinusowe dla słów

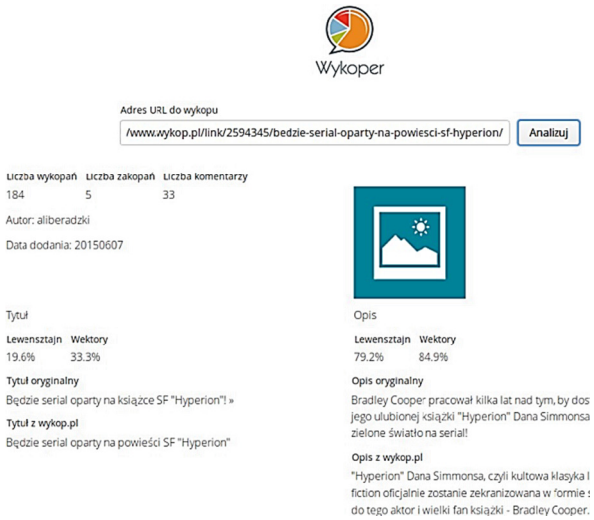
Uwagi implementacyjne

Wykonano aplikację sieciową w Javie. Wykorzystano serwis **mockable.io** do stworzenia namiastki API serwisu Wykop. Słowa treści wiadomości sprowadzane są do formy podstawowej za pomocą Morfologika.

Możliwe zastosowania

Stworzenie usługi agregatora doniesień o wydarzeniach, z wewnętrzną weryfikacją rzetelności zgłoszeń o wydarzeniach.

Ilustracje działania



3 Wyświetlenie wiadomości o wydarzeniach w pobliżu określonej lokalizacji

Źródła danych

wykop.pl × Google maps

Doniesienia użytkowników o wydarzeniach — Podkład mapowy

Pozyskanie danych

Analiza HTML — API

Złączenie danych

Zbiór nazw miejscowych w pobliżu określonej lokalizacji

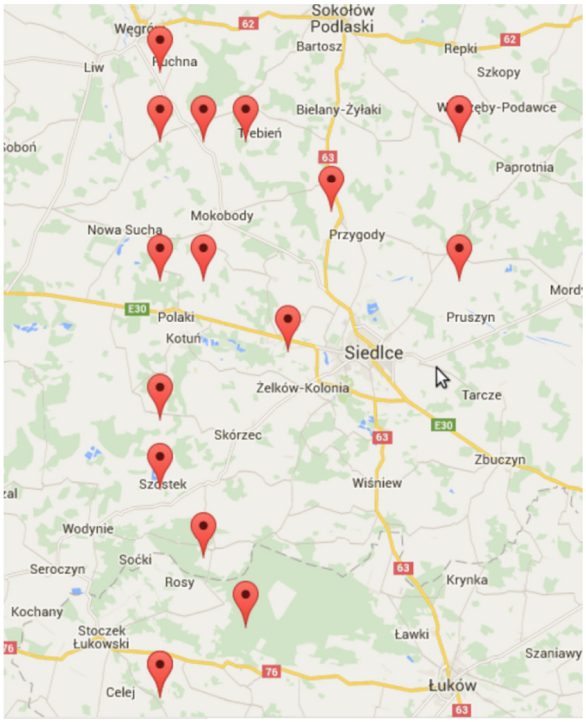
Uwagi implementacyjne

Wykonano aplikację graficzną w Pythonie i JavaScript. Dla zadanej nazwy lub współrzędnych, pobierany jest z Google Maps zbiór nazw okolicznych miejscowości – następnie wykorzystany jako kryterium wyszukiwania w Wykopie.

Możliwe zastosowania

Stworzenie usługi agregatora doniesień o wydarzeniach, uwzględniającego ich lokalizację.

Ilustracje działania



Mapa wydarzeń zaprezentowana użytkownikowi.

4 Pokazanie powiązań między artykułami The New York Times dla zadanych słów kluczowych, z wykorzystaniem kryteriów klasyfikacji określonych przez wydawcę

Źródła danych

The New York Times

Prasowy serwis informacyjny

Pozyskanie danych

API o ograniczonej wydajności

Złączenie danych

Poprzez współwystępowanie słów kluczowych przypisanych do artykułów, z uwzględnieniem ich istotności wg. dostawcy treści

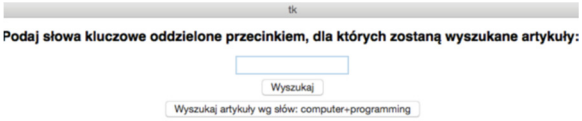
Uwagi implementacyjne

Wykonano aplikację graficzną w Pythonie. Wiele artykułów NYT nie ma przypisanych słów kluczowych i dlatego do budowy grafu powiązań byłoby potrzebne wyszukiwanie pełnotekstowe.

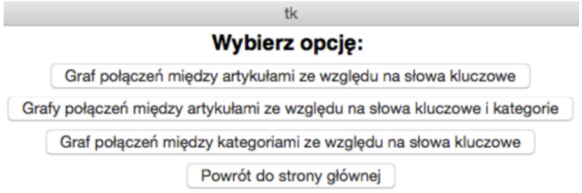
Możliwe zastosowania

Mechanizm rekomendacji nowych artykułów.

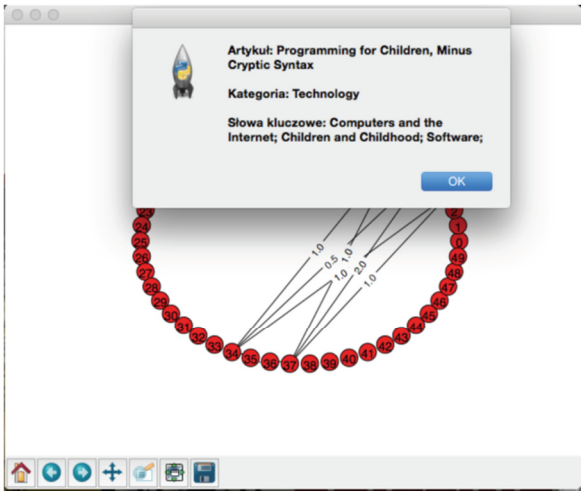
Ilustracje działania



Okno wyszukiwania artykułów



Okno wyboru możliwych opcji generowania grafów



Graf połączeń między artykułami ze względu na słowa kluczowe

5 Pokazanie grafu użytkowników gry Warcraft dla wybranego „królestwa”, pogrupowanych wg podobieństwa

Źródła danych

Portal tematyczny wow-heroes.com

Pozyskanie danych

Analiza HTML

Złączenie danych

Wskazany przez użytkownika podzbiór możliwych atrybutów: rasa, klasa, specjalizacja, gildia, królestwo

Uwagi implementacyjne

Wykonano aplikację konsolową w Pythonie. Do grupowania węzłów w dużym grafie zastosowano algorytm *k-prototypes*, bazujący na *k-means*: http://www.cs.ust.hk/~qyang/Teaching/537/Papers/huang98extensions.pdf

Możliwe zastosowania

Doradztwo specjalistyczne dla graczy ws. doboru własnych bohaterów w grze.

Wyniki działania

Wśród najlepiej radzących sobie bohaterów gry, dominuje jedna i ta sama rasa. Pozostałe atrybuty mają mniejsze znaczenie.

6 Pokazanie grafu *ego* generowanego przez `traceroute` do predefiniowanego zestawu adresów

Źródła danych

`traceroute`	× Alexa
Aplikacja raportująca parametry trasy do adresu IP	Serwis badający popularność adresów internetowych

Pozyskanie danych

Wielokrotne uruchamianie na maszynie użytkownika aplikacji	Archiwum 10 tys. najpopularniejszych adresów wg kategorii

Złączenie danych

Docelowe adresy IP

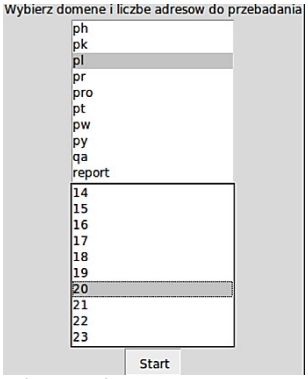
Uwagi implementacyjne

Wykonano aplikację graficzną w Pythonie, ale nie zinterpretowano w pełni wyników `traceroute`.

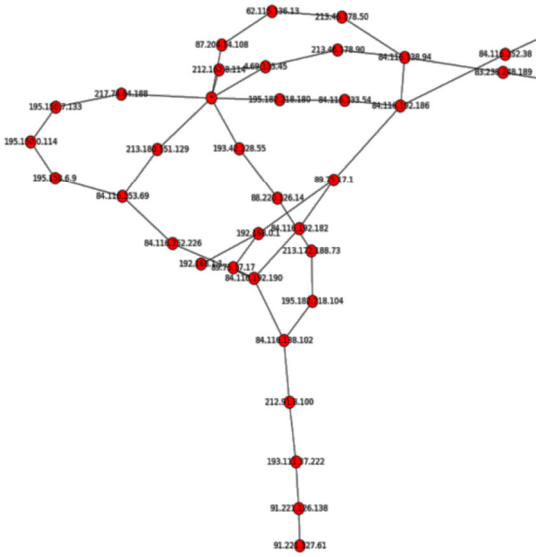
Możliwe zastosowania

Narzędzie do badania wydajności i niezawodności własnego połączenia z internetem.

Ilustracje działania



Okno wyboru parametrów analizy, tj. rodzaju i liczby adresów docelowych.



Uzyskana struktura połączeń.

7 Implementacja wyszukiwania pełnotekstowego w recenzjach książek

Źródła danych

lubimyczytać.pl
Twoja internetowa biblioteczka

Portal społecznościowy czytelników

Pozyskanie danych

Analiza HTML (JSoup)

Złączenie danych

Zaimplementowane wewnętrznie przez silnik Apache Lucene Solr (schemat TF-IDF)

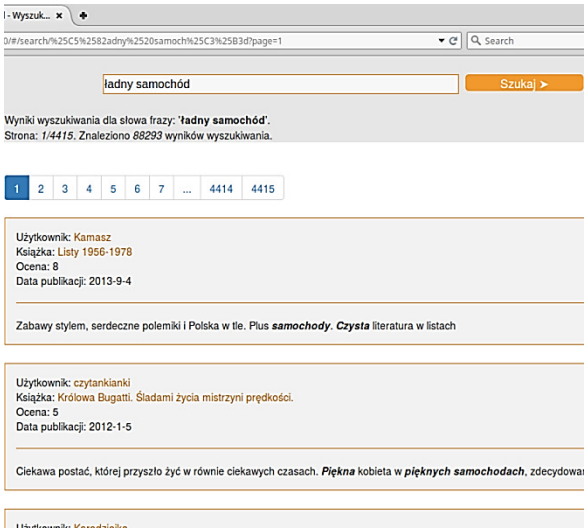
Uwagi implementacyjne

Wykonano aplikację sieciową z użyciem front-endu angularJS, komunikacji ReST, back-endu Java EE +Spring+PostgreSQL. Zbadano gruntownie wydajność aplikacji. Zaobserwowano znaczny spadek wydajności serwisu lubimyczytać przy pobieraniu kolejnych stron wyszukiwania o wysokich numerach.

Możliwe zastosowania

Mechanizmy rekomendacji i grupowania książek.

Ilustracje działania



Ekran wyników wyszukiwania.

8 Wyznaczenie trasy omijającej miejsca niebezpieczne

Źródła danych

dobraulica.pl
Sprawdź, zanim zamieszkasz
×
OpenStreetMap

Rejestr zdarzeń niebezpiecznych — Podkład mapowy

Pozyskanie danych

Analiza HTML — API

Złączenie danych

Lokalizacja GPS

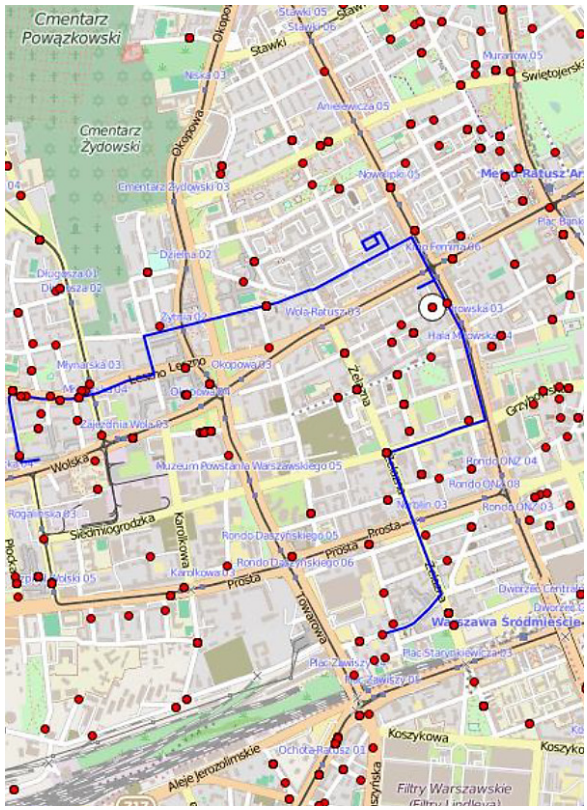
Uwagi implementacyjne

Wykonano aplikację w postaci JavaScriptu. Wobec problemów z implementacją własnego algorytmu trasowania, wykorzystano usługę OSRM (router.project-osrm.org). Do nałożenia warstwy na podkład użyto biblioteki OpenLayers. Z powodu odstępstw, aplikacja ma ograniczoną funkcjonalność.

Możliwe zastosowania

Rozwinięcie aplikacji, umożliwiające omijanie miejsc niebezpiecznych ze względu na kategorię zagrożenia, porę dnia i środek komunikacji.

Ilustracje działania



Mapa z wyznaczoną trasą.

9 Ukazanie połączeń pomiędzy dwiema osobami związanymi zawodowo

Źródła danych

KtoKogo.pl
KTO KOGO REPREZENTUJE WG KRS

Broker danych z Krajowego Rejestru Sądowego

Pozyskanie danych

Analiza HTML (BeautifulSoup)

Złączenie danych

Współwystępowanie 2 osób w tej samej organizacji

Uwagi implementacyjne

Wykonano aplikację konsolową w Pythonie. Niestety, nie rozwiązano problemu ujednoznacznienia osób o tych samych imionach i nazwiskach.

Możliwe zastosowania

Analiza powiązań pomiędzy osobami.

Ilustracje działania

10 Pokazanie lokalizacji określonych klubów piłkarskich; pokazanie klubów, z którymi rozgrywa spotkania wskazany klub

Źródła danych

Łączy nas piłka × OpenStreetMap

Portal piłkarski	Podkład mapowy

Pozyskanie danych

Analiza HTML	API

Złączenie danych

Siedziba klubu

Uwagi implementacyjne

Wykonano aplikację graficzną Javy. Aplikacja używa technologii: CouchDB, JSoup, oraz Nominatim – serwisu geokodowania. Problemem jest brak integralności struktury stron serwisu Łączynaspiłka.

Możliwe zastosowania

Rekreacyjne lub uatrakcyjnienie innych serwisów.

Ilustracje działania

11 Wyszukanie patentów osób o wskazanym nazwisku, z ujednoznacznieniem osób

Źródła danych

Amerykański urząd patentowy

Pozyskanie danych

Analiza HTML

Złączenie danych

Imię, nazwisko i adres (docelowo – lista współautorów; obecnie nie obsługiwana)

Uwagi implementacyjne

Wykonano aplikację sieciową w PHP, wykorzystując architekturę

Zaimplementowano cache zapytań z użyciem MySQL, aby przyspieszyć działanie, gdyż wydajność źródła danych okazała się zbyt mała.

Możliwe zastosowania

Aplikacja działa pod adresem www.tass.iplasota.vot.pl. Można ją wykorzystać do dalszej analizy kontaktów, działalności publikacyjnej i biznesowej wynalazców.

Ilustracje działania

Nazwisko (i opcjonalnie imię) wynalazcy: Davis Mark Szukaj

Znaleziono wyniki dla tej frazy.
Maksymalna liczba wczytanych patentów: 100.
Czas trwania analizy wszystkich stron: 00:01:38

Wynalazcy i wspólni współautorzy patentów

Imię wynalazcy	Miasto, Kraj/Stan	Liczba patentów	Współautorzy ogółem	Współautorzy wspólni z innymi
Burnapp; Mark	Bedfordshire, GB	1	3	2 z Davis; Mark James 2 z Davis; Paul James
Davis; Mark	Pasadena, CA	11	20	1 z Davis; Mark B. 2 z Davis; Mark E. 1 z Davis; Mark Edward
Davis; Mark Alan	Springville, UT	7	10	
Davis; Mark B.	Lake Jackson, TX	4	10	1 z Davis; Mark
Davis; Mark Bradley	Austin, TX	1	0	
Davis; Mark C.	Durham, NC	10	23	3 z Davis; Mark Charles

Wykaz patentów i współautorów

Imię wynalazcy	Współautorzy	Wykaz patentów
Burnapp; Mark	Davis; Mark James, Davis; Paul James, Hemmington; Sandra,	8,361,386,
Davis; Mark	Bernoulli; Carlo, Carr; Ryan, Cheng; Jianjun, Dredgelr.; John D., Friar; Gary, Gardner; Eric, Garnett; John, Howard; Richard D., Kahler; Andrew Wolf, Katz; Gary M., Khin; Kay T., Mendez; Jose Raul, Michie; William, Reeder; Paul A., Roe; Gene, Schindler; Peter, Wagner; Wesley J., Wang; Bin, Wiker; Nathan, Wilson; Debra,	8,404,662, 8,423,896, 8,475,781, 8,481,666, 8,531,286, 8,566,095, 8,650,065, 8,677,286, 8,682,644, 8,739,772, 8,755,113,
Davis; Mark Alan	Attavar; Sachin, Bangerter; Vern W., Blair; Steven M., Bradshaw; Simon Craig Caioneach, Chagovetz; Alexander, Decker; Keith W., Diwekar; Mohit, Dredge; John, Hullinger; Derek, Sabini; Eugene P.,	8,535,616, 8,580,658, 8,611,007, 8,698,091, 8,873,144, 8,913,320, 8,979,476,
Davis; Mark B.	Frick; Michael D., Kao; Sun-Chueh, Kapur; Mridula, Kwalk; Tae Hoon, Michielr.; William J., Mizeur; David Richard, Pade; Stan, Schindler; Peter, Wiker; Nathan J., Wilson; Debra R.,	8,486,323, 8,540,568, 8,742,039, 9,028,317,
Davis; Mark Bradley		8,599,863,
Davis; Mark C.	Bin; Li, Challener; David C., Cromer; Daryl C., Dixon; Martin G., Farcy; Alexandre J., Feghali; Wajdi K., Gopal; Vinodh, Guilford; James D., Hortenstus; Peter, Kelso; Scott E., Locker; Howard, Locker; Howard J., Loktyukhin; Maxim, Mirkes; Sean P., Ozturk; Erdinc, Perrin; Steven R., Resnick; Russell A., Roper; Matthew, Sheng; Wang, Springfield; Randall S., Toll; Bret L., Waltermann; Rod D., Wolrich; Gilbert M.,	8,473,752, 8,504,807, 8,549,264, 8,566,480, 8,607,085, 8,667,577, 8,682,962, 8,694,797, 8,984,200, 9,002,925,
Davis; Mark Charles	Cannady; Stacy John, Challener; David Carroll, Clark; Jeffrey, Cromer; Daryl, Dubs; Justin Tyler, Kelso; Scott Edwards, Li; Bin, Locker; Howard J., Matthews; Michael Thano, Perrin; Steven Richard, Rivera; David, Roper; Matthew, Rutledge; James Stephen, Springfield; Randall Scott, Tang; Liang, Ulrich; Sean Michael, Waltermann; Rod D., Waltermann; Rod David, Wang; Sheng, Zawacki; Jennifer Greenwood, Zhang; Dekui, Zhou; Yi,	8,452,877, 8,490,177, 8,533,506, 8,539,114, 8,706,642, 9,026,824,
Davis; Mark E.	Alabi; Akinleye, Archer; Raymond, Bellocq; Nathalie C., Cheng; Jianjun, Czaja; Stanislaw, Hwang; Jungyeon, Ke; Tianyi, Khin; Kay T., Lim; Ching-Jou, Mack; Brendan C., Moliner-Marin;	8,357,377, 8,389,499, 8,399,431, 8,497,365,

Przykładowe wyniki wyszukiwania.

12 Wizualizacja lokalizacji darczyńców i kwot darowizn na rzecz kandydatów wyborach, z uwzględnieniem pośrednictwa komitetów wyborczych

Źródła danych

×

Komisja wyborcza USA

Podkład mapowy

Pozyskanie danych

Pliki tekstowe do pobrania

API

Złączenie danych

Lokalizacja darczyńcy

Uwagi implementacyjne

Wykonano aplikację graficzną w JavaScript i architekturze AngularJS, współpracującą z danymi z FEC umieszczonymi uprzednio w bazie CouchDB. Interaktywne pobieranie danych z FEC okazało się zbyt wolne; natomiast takie same dane udostępniane przez The NY Times przez API są ograniczone do pierwszych 100 pozycji. Geolokalizacja poprzez statyczną, lokalną bazę kodów pocztowych GreatData.

Możliwe zastosowania

Prezentacja preferencji politycznych i hojności darczyńców w aspekcie geograficznym. Analiza roli komitetów wyborczych.

Ilustracje działania

13 Zarekomendowanie nowego filmu, na podstawie preferencji użytkowników portalu wobec zadanej listy filmów

Źródła danych

FILMWEB

Portal społecznościowy opinii o filmach

Pozyskanie danych

Analiza HTML (JSoup)

Złączenie danych

Imię, nazwisko i adres (docelowo – lista współautorów; obecnie nie obsługiwana)

Uwagi implementacyjne

Wykonano aplikację graficzną w Javie, z wykorzystaniem technik: gradle, JPA+PostgreSQL i Spring. Dla każdego z zadanych filmów, aplikacja wyszukuje z forum filmu użytkowników zadowolonych i sporządza ranking średnich ocen filmów wysoko przez nich ocenianych.

Możliwe zastosowania

Autorski, alternatywny system rekomendacji.

Ilustracje działania

14 Wyszukanie instytucji naukowych realizujących w określonym dniu projekty badawcze i zweryfikowanie wpływu projektów na działalność publikacyjną kierowników tych projektów

Źródła danych

OPI × Google Scholar

Ośrodek Przetwarzania Informacji	Wyszukiwarka publikacji naukowych

Pozyskanie danych

Analiza HTML	API (nieoficjalne)

Złączenie danych

Personalia kierowników projektów

Uwagi implementacyjne

Wykonano aplikację konsolową w Pythonie, z wykorzystaniem BeutifulSoup, Mechanize i github.com/ckreibich/scholar.py. Nie zaimplementowano wyszukiwania aktywnych projektów po dacie, bo OPI tego nie umożliwia.

Możliwe zastosowania

Analiza pożyteczności programów badawczych.

Ilustracje działania

```
Podaj tytuł projektu:
Podaj imie: Roman
Podaj nazwisko:
Podaj słowa kluczowe:
Roman Zawodny był kierownikiem projektu w roku 2000
Aktywność optyczna dielektryków w przybliżeniu kwadrupulowo-elektryc
Średnia liczba cytowań rocznie przed projektem: 39.7058823529
Średnia liczba cytowań rocznie po projekcie: 195.8
Współczynnik wzrostu cytowań: 4.93125925926
---------------------------
Krzysztof Kazimierz Myszkowski był kierownikiem projektu w roku 2010
Akcja misyjna Kościoła prawosławnego wśród grekokatolików w Polsce w
Średnia liczba cytowań rocznie przed projektem: 53.0625
Średnia liczba cytowań rocznie po projekcie: 246.4
Współczynnik wzrostu cytowań: 4.64358068316
---------------------------
Roman Słowiński był kierownikiem projektu w roku 2003
Algorytmy inteligentnej eksploracji danych i wspomagania decyzji
Średnia liczba cytowań rocznie przed projektem: 59.1111111111
Średnia liczba cytowań rocznie po projekcie: 218.6
Współczynnik wzrostu cytowań: 3.69812030075
---------------------------
Robert Susmaga był kierownikiem projektu w roku 2001
Algorytmy dokładne redukcji danych oparte na teorii zbiorów przybliż
Średnia liczba cytowań rocznie przed projektem: 70.9230769231
Średnia liczba cytowań rocznie po projekcie: 250.2
Współczynnik wzrostu cytowań: 3.52776572668
---------------------------
Grzegorz Matusik był kierownikiem projektu w roku 2012
Algebraiczne i topologiczne własności rodziny funkcji Hamela
Średnia liczba cytowań rocznie przed projektem: 62.05
Średnia liczba cytowań rocznie po projekcie: 202.4
Współczynnik wzrostu cytowań: 3.26188557615
```

15 Wyszukanie informacji i zdjęć piłkarzy o określonym nazwisku

Źródła danych

Łączy nas piłka × Google Image Search

Portal piłkarski	Wyszukiwarka publikacji naukowych

Pozyskanie danych

Analiza HTML	API

Złączenie danych

Imię, nazwisko, wiek, nazwa i adres klubu, drużyna

Uwagi implementacyjne

Wykonano aplikację graficzną w Javie z użyciem JSoup. Ograniczeniem jest szybkość działania serwisu Łączynaspiłka. Konieczne okazało się dodatkowe filtrowanie wyszukanych zdjęć, uwzględniające dane w opisie zdjęcia.

Możliwe zastosowania

Rekreacyjne.

Ilustracje działania

16 Pokaż lokalizację osób lubiących wskazaną książkę

Źródła danych

lubimyczytać.pl Twoja internetowa biblioteczka × Global Administrative Areas Boundaries without limits

Portal społecznościowy czytelników	Podkład mapowy GADM

Pozyskanie danych

Analiza HTML	Pliki do pobrania

Złączenie danych

Nazwa geograficzna

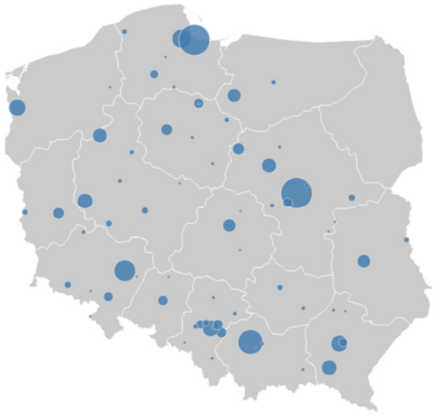
Uwagi implementacyjne

Aplikacja sieciowa napisana w Pythonie + BeautifulSoup, MongoDB, Google Geocoding. Podkład mapowy został pobrany do użytku lokalnego. Dane z Lubimyczytać zostały pobrane również do użytku lokalnego (ok. 20 tys. stron).

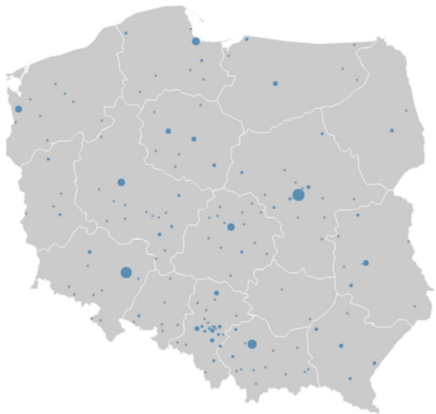
Możliwe zastosowania

Wsparcie przy organizowaniu imprez tematycznych, kampanii reklamowych, inicjatyw społecznych.

Ilustracje działania



Czytelnicy Terry'ego Prachetta



Czytelnicy serii „Zmierzch"

17 Wizualizacja obszaru i intensywności działania firm realizujących zamówienia publiczne

Źródła danych

OpenTED × Google maps

Baza zamówień publicznych	Podkład mapowy

Pozyskanie danych

Pliki tekstowe do pobrania	API

Złączenie danych

Lokalizacja kontrahenta

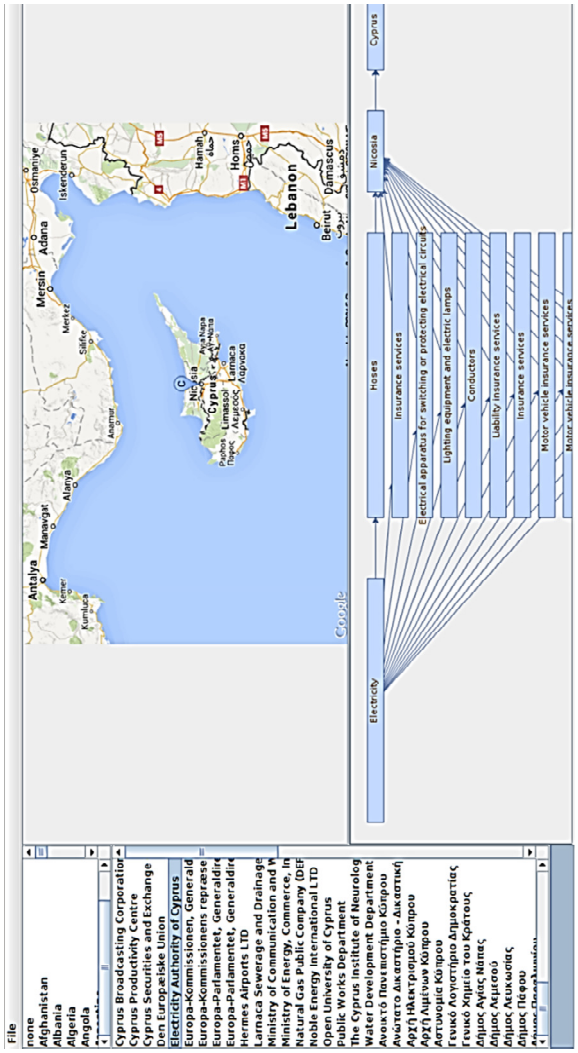
Uwagi implementacyjne

Wykonano aplikację graficzną w Javie, z wykorzystaniem mxGraph. Pozyskane dane o kontraktach często okazują się niepełne, co uniemożliwia poprawną geolokalizację i wizualizację.

Możliwe zastosowania

Wywiad gospodarczy, marketing i reklama.

Ilustracje działania



18 Pokaż zweryfikowane opinie o wskazanym lekarzu

Źródła danych

znanylekarz.pl × dobrylekarz.pl

Portal branżowy i rezerwacja terminów wizyt	Portal opinii pacjentów

Pozyskanie danych

Analiza HTML	Wyszukiwanie Google custom search + analiza HTML

Złączenie danych

Personalia, specjalizacja i miejscowość lekarza

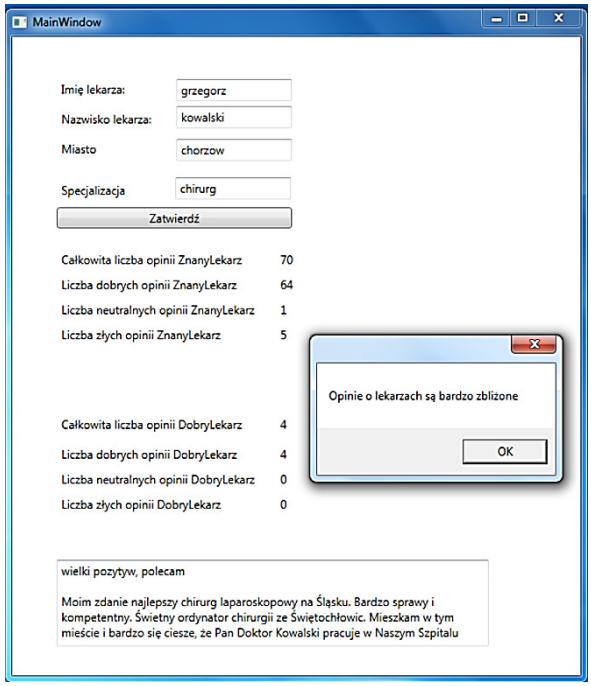
Uwagi implementacyjne

Aplikacja graficzna napisana w C# z wykorzystaniem HTMLAgilityPack. Klasyfikacja opinii (dobra/zła) na podstawie słowników wyrażeń. Problemem jest zamknięte API serwisu Znanylekarz, a także niewystarczająca liczba opinii weryfikujących w serwisie Dobrylekarz.

Możliwe zastosowania

Weryfikacja opinii o lekarzach, silnie moderowanych w jednym serwisie i zupełnie niekontrolowanych – w drugim.

Ilustracje działania



19 Weryfikacja wiarygodności opinii o połączeniach (wywoływacz,film) na podstawie danych z fotograficznego portalu społecznościowego

Źródła danych

FilmDev. × flickr

Portal technik fotograficznych	Fotograficzny portal społecznościowy

Pozyskanie danych

API	API

Złączenie danych

Nazwa filmu, wywoływacza, ew. identyfikator tej kombinacji

Uwagi implementacyjne

Analizę wykonano za pomocą zestawu skryptów w Pythonie. Dane zgromadzono w formacie GraphML.

Możliwe zastosowania

Zweryfikowana fachowa pomoc dla fotografów i fotografików.

Wyniki działania

filmdev_graph		flickr_graph	
film	l. zdj.	film	l. zdj.
Kodak Tri-X 400	3017	120	167231
Ilford HP5+ 400	1986	400TX	65712
Kodak T-Max 400	888	FP4+	52189
Shanghai GP3 100	781	Kodak Tri-X 400	47433
Fuji Neopan 400	753	Kodak Tri-x 400	46817
Fuji Neopan Acros 100	743	ilford hp5	39607
Ilford FP4+ 125	689	Neopan 400	21976
Agfa Rodinal	585	Kodak TMAX 400	20510
Foma Fomapan 100	488	Kodak T-Max 400	19487
Kodak T-Max 100	358	KODAK 400TX	18736

Tabela 1: Lista najpopularniejszych 10 filmów.

filmdev_graph		flickr_graph	
wywoływacz	l. zdj.	wywoływacz	l. zdj.
Agfa Rodinal	2824	Rodinal	75523
Kodak D-76	2429	rodinal	74676
Kodak HC-110	1951	Rodinal	74607
Agfa R09 One Shot	1120	C-41	56546
Kodak XTOL	1093	d-76	52492
Ilford ID11	1031	D76	49897
Foma Fomadon R09	599	d 76	46094
Adox Adonal	408	d76	43963
Kodak T-MAX	376	D-76	41994
Tetenal Ultrafin Plus	373	hc-110	36165

Tabela 2: Lista najpopularniejszych 10 wywoływaczy.

20 Grupowanie tematyczne tagów na podstawie współwystępowania ich w wypowiedziach na forum programistycznym

Źródła danych

stackoverflow

Forum tematyczne

Pozyskanie danych

API (otwarte, bez ograniczeń)

Złączenie danych

Wspólne wystąpienie pary tagów w opisie poszczególnych pytań zadawanych na forum

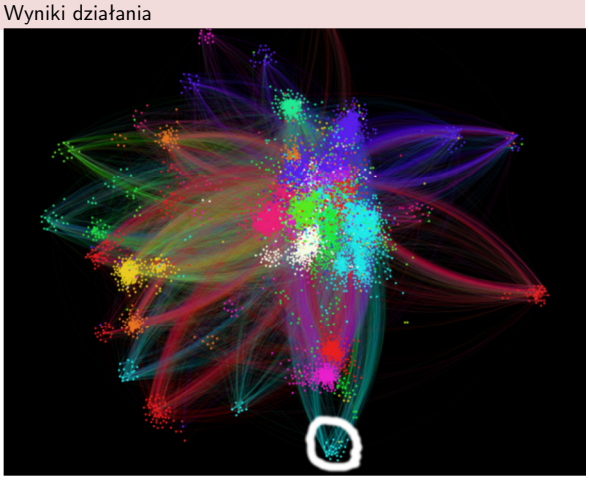
Uwagi implementacyjne

Analizę wykonano za pomocą zestawu skryptów w Pythonie. Dane zgrupowano niezależnie za pomocą algorytmów Louvain (użyty do kolorowania wierzchołków) oraz OpenOrd (do rozmieszczenia). Rozmieszczenie realizuje aplikacja Gephi.

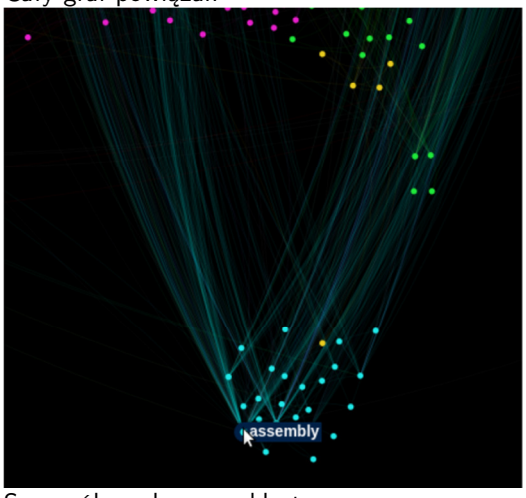
Możliwe zastosowania

Behawioralna kategoryzacja zagadnień informatycznych; oszacowanie popularności i powiązań między zagadnieniami.

Wyniki działania



Cały graf powiązań



Szczegóły wybranego klastra

21 Grupowanie książek na podstawie opinii użytkowników

Źródła danych

lubimyczytać.pl
Twoja internetowa biblioteczka

Forum tematyczne

Pozyskanie danych

Analiza HTML

Złączenie danych

Osoby wysoko oceniające obie książki

Uwagi implementacyjne

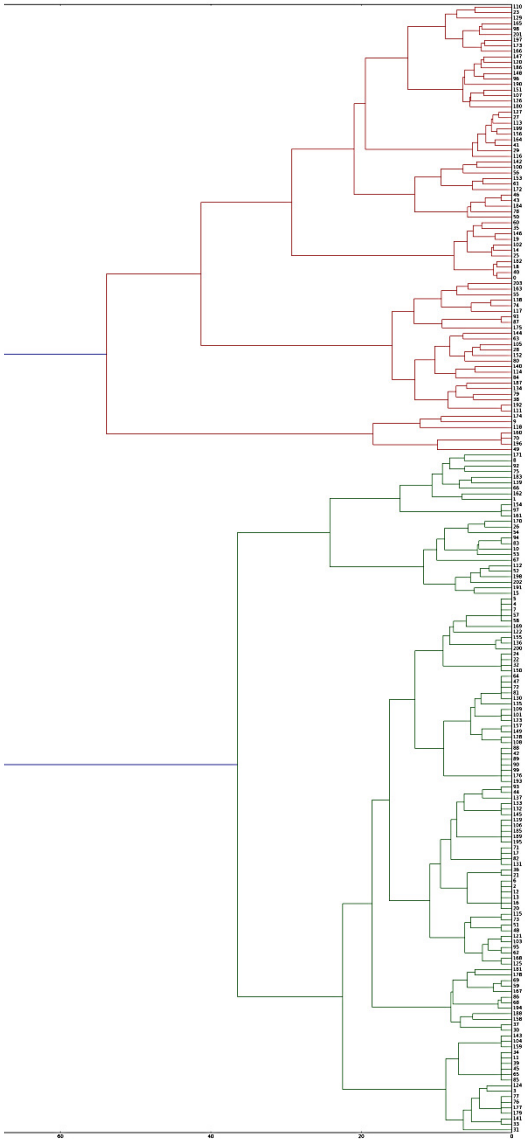
Wykonano zestaw aplikacji w Pythonie.

Możliwe zastosowania

Mechanizm rekomendacji.

Wyniki działania

Przeanalizowano 258 książek.
Książka o największej liczbie wspólnych czytelników: Złodziejka książek (wyd. Nasza Księgarnia).



Grupowanie aglomeracyjne książek metodą Warda.

22 Utworzenie sieci społecznej interakcji grup łosi i jej analiza

Źródła danych

DATA BASIN × Arc GIS ESRI

Ślady mobilności zwierząt	Podkład mapowy

Pozyskanie danych

Pobrane pliki	Pobrane pliki

Złączenie danych

Bliskość (w czasie i przestrzeni) grup łosi.

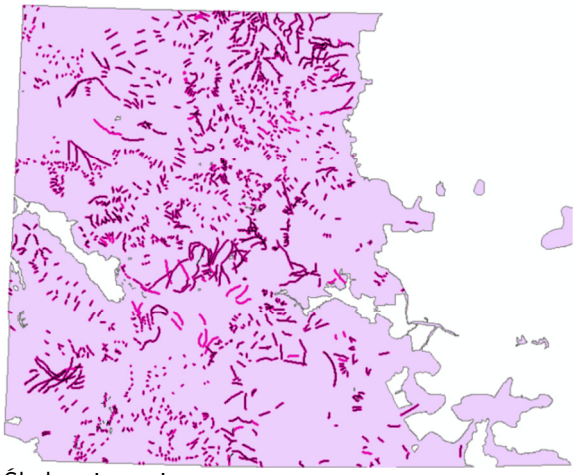
Uwagi implementacyjne

Analizę wykonano za pomocą zestawu skryptów w Pythonie. Dane zgromadzono w formacie GraphML.

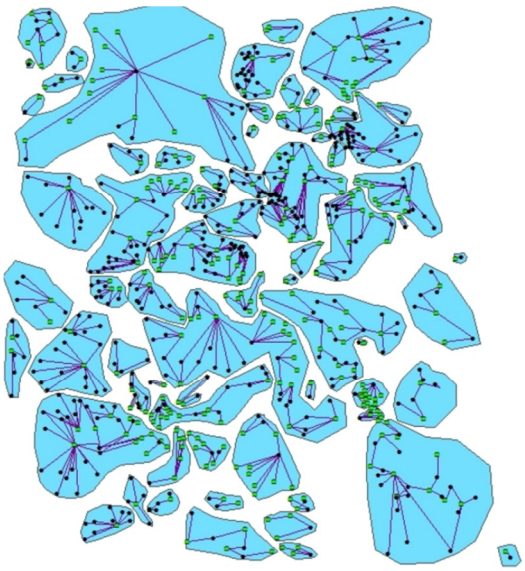
Możliwe zastosowania

Wsparcie prac w dziedzinie badań przyrodniczych.

Wyniki działania



Ślady migracyjne.



Zrekonstruowane sieci społeczne łosi.

23 Analiza powiązań artykułów w poszczególnych dziedzinach i pomiędzy dziedzinami

Źródła danych

arXiv.org

Repozytorium publikacji o otwartym dostępie

Pozyskanie danych

Analiza HTML

Złączenie danych

Zakwalifikowanie artykułu naukowego do 2 lub więcej dziedzin

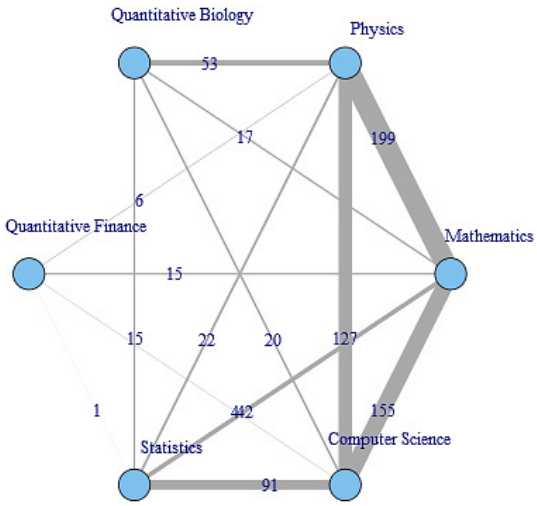
Uwagi implementacyjne

Wykonano zestaw skryptów w R.

Możliwe zastosowania

Edukacyjne i rekreacyjne.

Wyniki działania



Powiązania między dziedzinami.

Dziedzina	Liczba wspolautorow
Computer Science	3123
Mathematics	3968
Physics	17916
Quantitative Biology	529
Quantitative Finance	124
Statistics	622

Liczba współautorów dla poszczególnych dziedzin.

24 Analiza podobieństwa pseudonimów w grze World of Warcraft do imion bohaterów książek

Źródła danych

Portal gry zawierający informacje o aukcjach wirtualnych rekwizytów	Autorski słownik imion

Pozyskanie danych

API (przez Apigee.com)	Stworzony z korpusu językowego książek

Złączenie danych

Imię lub pseudonim; rozmyte dopasowanie Levenshteina

Uwagi implementacyjne

Analizę wykonano za pomocą zestawu programów napisanych w C++. Dalszą poprawę wyników można by uzyskać rozszerzając korpus językowy oraz dokładniej filtrując nazwy własne niebędące imionami. Wykryto mało podobieństw w nazewnictwie graczy, bo też liczba graczy była niewielka (7500).

Możliwe zastosowania

Marketing (nazwy gadżetów) i reklama.

Wyniki działania

Najpopularniejsze imiona, dokładnie dopasowane:
Loki, Aggo, Witch, Hatred, James, Wrong, Bold.
Popularne imiona, odległość edycyjna 1:
Scar, Sarah, Varo, Hord, Lily, Meera, Bank.

Najpopularniejszą inspiracją dla nazewnictwa postaci w grach jest seria *Gra o Tron*.

25 Analiza atrybutów gracza/awatara/przedmiotu, które mogą wpływać na wysokie przebicie ceny wywoławczej na aukcjach przedmiotów gry World of Warcraft

Źródła danych

Portal gry zawierający informacje o aukcjach wirtualnych rekwizytów	Opisy licytowanych rekwizytów
API (przez Apigee.com)	API (przez `curl`)

Złączenie danych

Identyfikator przedmiotu

Uwagi implementacyjne

Wykonano szereg regresji liniowych pomiędzy różnymi atrybutami licytowanych przedmiotów, a ich ceną. Problemem okazała się mała dynamika licytacji, tj. niewielka liczba przebić cen wywoławczych.

Możliwe zastosowania

Wsparcie strategiczne dla graczy.

Wyniki działania

Przeanalizowano wpływ atrybutów:

a) Ranking odczarowania przedmiotu
b) Cena zakupu w sklepie
c) Klasa przedmiotu
d) Czy jest zdolny do ekwipunku
e) Poziom
f) Jakość
g) Trwałość
h) Wymagany poziom
i) Obrona
j) Dodatkowe statystyki przedmiotu

Niektóre atrybuty wykazują silną korelację pomiędzy ceną zbycia a np. wymaganym poziomem postaci, niezbędnym do nabycia przedmiotu:

26 Analiza współwystępowania osób w spółkach z udziałem Skarbu Państwa

Źródła danych

Rejestry Ministerstwa Skarbu Państwa	Agregator danych, w tym KRS

Pozyskanie danych

Pobrane pliki	Analiza HTML

Złączenie danych

Nazwa spółki z udziałem Skarbu Państwa

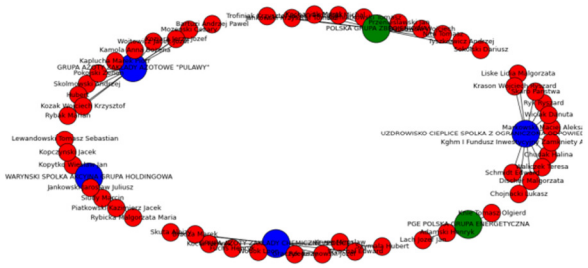
Uwagi implementacyjne

Analizę wykonano za pomocą zestawu programów napisanych w Pythonie. Problemem stało się dopasowanie spółek po nazwie, przy ich niespójnym nazewnictwie (MSP nie udostępnia numeru KRS spółek, w których ma udział).

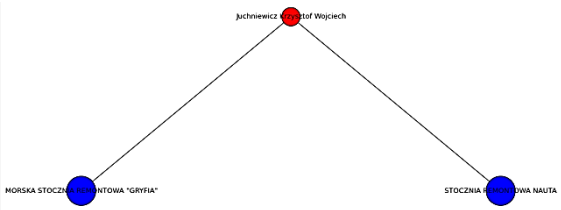
Możliwe zastosowania

Śledzenie nadużyć.

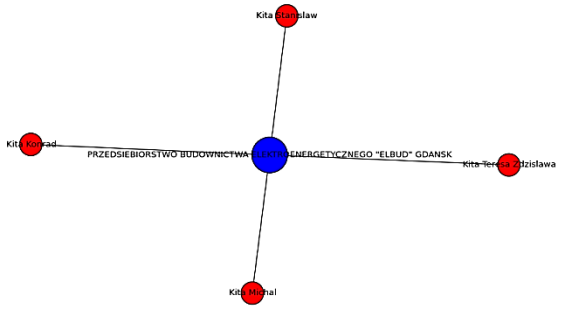
Wyniki działania



Spółki zawierające w nazwie słowo GRUPA.



Ta sama osoba działająca w konkurencyjnych spółkach.



Osoby o tym samym nazwisku występujące w tej samej spółce.

27 Analiza faktycznie konkurujących dostawców, tj. dostarczających takie same towary tym samym zamawiającym w drodze zamówień publicznych

Źródła danych

OpenTED

Baza zamówień publicznych

Pozyskanie danych

Pliki do pobrania

Złączenie danych

Autorski, dwupoziomowy słownik kategorii zamówień

Uwagi implementacyjne

Aplikację analityczną zaimplementowano w C++. Słownik kategorii dostarczony przez TED okazał się nieprzydatny, opracowano własną taksonomię słów kluczowych. Dane pierwotne są zanieczyszczone.

Możliwe zastosowania

Wywiad gospodarczy.

Wyniki działania (obok)

Intertrading System Technology Sp. z o.o.	4
Atende S.A.	3
Enigma SOI Sp. z o.o.	3
Qumak S.A.	3
Blue Energy Sp. z o.o.	2
CA Consulting S.A.	2
Decsoft S.A.	2
Deloitte Advisory Sp. z o.o.	2
ITC S.A.	2
Soflab Technology Sp. z o.o.	2
Wykonawcy wspólnie ubiegający się o udzielenie zamówienia: 1) Lider konsorcjum Enigma Systemy Ochrony Informacji sp. z o.o.	2
Wykonawcy wspólnie ubiegający się o udzielenie zamówienia: 1) Lider konsorcjum Trecom sp. z o.o. S.K.A.	2
Asseco Poland S.A.	1
Crann Spółka z ograniczoną odpowiedzialnością	1
CUBE.ITG S.A.	1
Data Techno Park Sp. z o.o.	1
DGT Sp. z o.o.	1
Engave Sp. z o.o.	1
Infonet Projekt S.A.	1
IT Expert Sp. z o.o.	1
MCX Telekom Sp. z o.o.	1
Wykonawcy wspólnie ubiegający się o udzielenie zamówienia: 1) Lider konsorcjum - Umbrella Consulting sp. z o.o.	1
Fagron Sp. z o. o.	1
GSK Services Sp. z o. o.	1
Intra Sp. z o. o.	1
KOMAK Sp. z o.o.	1
Konsorcjum firm: Amgen Sp. z o. o. i Nettle S.A.	1
konsorcjum firm: Anpharm Przedsiębiorstwo Farmaceutyczne S.A. i Servier Polska Services Sp. z o. o.	1
konsorcjum firm: Aspen Pharma Ireland Limited i Nettle Pharma Services Sp. z o. o.	1
konsorcjum firm: PGF Urtica Sp. z o. o. i PGF S.A	1
Med&Care Tomasz Witkowski	1

Wykonawcy zleceń publicznych w kategorii „IT & Electronics" dla Centrum Projektów Informatycznych (nazwa, liczba kontraktów)

Wszystkie przedstawione projekty dostarczyły użytecznych wyników lub oprogramowania. Studenci zastosowali bardzo zróżnicowane technologie implementacji, często zupełnie odmienne od proponowanych na wykładzie (np. R, C++ C#, PHP). To samo dotyczy algorytmów, zwłaszcza analizy stron HTML, grupowania wierzchołków i wizualizacji grafów. Poniżej przedstawiono wnioski szczegółowe dotyczące poszczególnych aspektów wykonania projektu.

Źródła i pozyskiwanie danych. Serwisy o zasięgu globalnym z reguły udostępniają API do danych, aczkolwiek zasady korzystania z niego mogą różnić się znacznie. Powszechne jest ograniczanie przepustowości kanału. Bardzo wiele serwisów, w tym większość polskich, nie udostępnia powszechnego API, ale umożliwia dostęp do treści przez odpowiednio sformułowane adresy URL; z reguły treść ta jest w formacie HTML, który musi być przeanalizowany. Jest to zadanie żmudne, ale głównym problemem w tym przypadku jest brak mechanizmu wymuszającego spójność (jednolitą strukturę) pobieranych danych oraz nierównomierna wydajność serwerów – w zależności np. od paginacji wyników. Dane przygotowane i pobierane w postaci plików z reguły nie sprawiają powyższych problemów, ale skutecznie uniemożliwiają stworzenie aplikacji interaktywnych, wymagających jedynie podzbioru danych.

Złączenie danych. Powszechnym problemem w łączeniu danych (nie tylko z wielu źródeł) jest brak unikatowego klucza złączenia, tj. wspólnego identyfikatora dla danych. Niekiedy jego udostępnienie nie wymagałoby żadnych wysiłków, np. Ministerstwo Skarbu Państwa mogłoby opatrzyć nazwy spółek, w których ma udziały, numerem KRS – ale tego nie robi. Podobnie, serwisy społecznościowe byłyby w stanie identyfikować użytkowników ich profilami z Facebooka czy Google+, skoro umożliwiają takie uwierzytelnienie. Nie przypuszczam jednak, że ulegnie to zmianie w przyszłości, choćby ze względu na troskę o prywatność użytkowników. Tymczasem jedynym sensownym podejściem w łączeniu takich danych jest zastosowanie dopasowania rozmytego (np. z użyciem odległości Levenshteina) oraz poszukiwanie dodatkowych atrybutów ujednoznaczniających złączenie (np. nazwy miejscowości i specjalności w przypadku lekarzy).

Implementacja. Studenci swobodnie implementują jednostanowiskowe aplikacje graficzne; większość jest również w stanie stworzyć bez wysiłku aplikację sieciową. To dobrze rokuje dla powstawania autorskich, różnorodnych aplikacji typu *mash-up*, ale nie umożliwia ich naturalnego rozwoju. Dlatego należałoby wyposażyć studentów w narzędzia do łatwego makietowania aplikacji działających od razu w chmurze. Obecnie nie widzę ani dość przystępnej architektury dla tego typu aplikacji, ani wystarczająco otwartego na nowości serwisu chmurowego.

Zastosowania. Większość zastosowań tego typu aplikacji i analiz lokuje się w obszarach marketingu, zarządzania i rozrywki. Wynika to ze specyfiki danych wejściowych, które z założenia miały opisywać interakcje społeczne. Jednakże w przypadku źródeł ściślej związanych z infrastrukturą, np. sieci internet, zastosowania również stają się bardziej „technologiczne" (badanie niezawodności, plany inwestycyjne).

Zrozumienie i prezentacja wyników. Możliwości studentów są pod tym względem najbardziej zróżnicowane. Niektórzy wykorzystali w maksymalnym (i właściwym) stopniu daną im swobodę, proponując własne miary podobieństwa węzłów (np. Twitter kontra Wykop); wielu jednak nie pojęło sensu wnioskowania o relacjach między węzłami sieci dwudzielnej, nie posuwając się nawet do normalizacji wyników np. względem stopni wierzchołków. W ogólnym interesie społeczeństwa jest, aby aplikacje i analizy tego drugiego typu nie rozpowszechniały się, gdyż wprowadzają w błąd, zamiast przekazywać konstruktywne wnioski. Łączenie i wnioskowanie z danych z wielu źródeł, zwłaszcza dotyczących ludzi, musi podlegać szczególnej weryfikacji poprawności.