CENTRUM CYFROWE

PIGUŁKA WIEDZY

EKSPLORACJA TEKSTU I DANYCH (TEXT AND DATA MINING)

Text and data mining (TDM) to metoda komputerowej analizy tekstu i danych, odgrywająca coraz większe znaczenie w sferze badań i rozwoju. Tylko z pomocą takich metod możemy w pełni korzystać z ogromnych zasobów danych i tekstu, generowanych i dostępnych w cyfrowym świecie.

Dzięki TDM możemy tworzyć nowe leki, dokonywać odkryć naukowych czy tworzyć produkty lepiej dopasowane do indywidualnych potrzeb klientów. TDM to szansa na jeszcze lepsze wykorzystanie potencjału nowych technologii dla rozwoju gospodarczego. Jednak, aby w pełni wykorzystać potencjał jaki drzemie w TDM, należy:

- uregulować kwestie prawne, tak aby zagwarantować szeroki zakres dopuszczalnych metod oraz wyeliminować wątpliwości co do dopuszczalnych działań;
- promować publikowanie danych i treści w otwartych formatach, umożliwiających automatyczny odczyt i analizę;
- podnieść świadomość społeczną o korzyściach, jakie przynosi TDM, otwieranie i dzielenie się danymi oraz o sposobach w jakim można to robić.

1. Dlaczego to zagadnienie jest ważne?

Dane rozumiane są dziś nie tylko jako liczby w tabelach czy suche fakty, ale też filmy, muzyka lub np. obrazy. Z kolei teksty to różnego rodzaju publikacje naukowe, ekspertyzy, opracowania czy artykuły. Tak dane, jak i teksty są dziś gromadzone przez instytucje publiczne, firmy, organizacje, ale też przez zwykłych ludzie pasjonujący się danym tematem. Często ze względu na ilość lub złożoność zawartych w nich informacji, nie jesteśmy w stanie sami ich zanalizować. Dlatego analiza wykonywana jest automatycznie – z wykorzystaniem komputerów i specjalnie do tego stworzonych algorytmów. TDM jest dziś wykorzystywany w instytutach badawczych, na

uniwersytetach, w przedsiębiorstwach i instytucjach publicznych – idea jest jednak taka, aby każdy badacz mógł bez obaw analizować nie tylko swoje dane, ale przede wszystkim (i o to chodzi w TDM) dane dostępne z różnych zewnętrznych źródeł, w tym te publikowane w internecie.

2. Co to jest TDM?

TDM jest szczególnie przydatny w badaniach naukowych i biznesie, ale nie tylko - jest pomocny wszędzie tam, gdzie praca polega na analizie wielkich zbiorów tekstów i danych:

- **TDM to postęp w nauce.** Przykładowo, w medycynie[1] komputerowa analiza danych pomaga w leczeniu raka - m.in. poprzez obniżenie kosztów i upowszechnienie przeprowadzania pełnej analizy ludzkiego genomu. To szansa dla chorych na Parkinsona, bo dziś, dzięki specjalnym urządzeniom, można monitorować parametry fizjologiczne pacjentów i od razu je analizować, tak aby wykryć symptomy i dostosować leczenie. TDM był też wykorzystywany przez naukowców w badaniach nad rozprzestrzenianiem się groźnych wirusów[2]: eboli, dengi i ostatnio wirusa ZIKA[3]. Z kolei Global Forest Watch[4] analizuje dane tworząc interaktywne mapy celu monitorowania m.in. zmian klimatu wskutek deforestacji czy wpływu pożarów lasów na stan powietrza.
- **TDM to szansa dla przedsiębiorców.** TDM jest coraz częściej stosowany przez sektor technologiczny. Dzięki analizie danych o klientach - która nawet najlepszemu analitykowi zabrałaby mnóstwo czasu i energii, a algorytmowi kilka chwil - firmy mogą lepiej dopasowywać oferty i ograniczać koszty. Kiedyś stworzenie odpowiednich algorytmów, które analizowałyby dla nas dane było bardzo drogie, a ich wykorzystanie wymagało zaawansowanego i bardzo kosztownego sprzętu. Dziś proste programy do analizy danych, modelowania i optymalizacji są dostępne za darmo i nie wymagają superkomputerów, dlatego stać na nie niemal wszystkich. Przykładowo, miasta mogą na podstawie danych o pasażerach usprawniać lokalną komunikację[5], a dzięki danym o trasach przejazdu

[1] Tak analiza big data pomaga ratować ludzkie życie, benchmark.pl (30.07.2015) http://www.benchmark.pl/aktualnosci/tak-analiza-big-data-pomaga-ratowac-ludzkie-zycie.html.
[2] How big data will help fight global epidemics, blog ITU4U (13.10.2015) https://itu4u.wordpress.com/2015/10/13/how-big-data-will-help-fight-global-epidemics/.
[3] Can big data help fight the ZIKA virus, Forbes (10.02.2016) http://www.forbes.com/sites/bernardmarr/2016/02/10/can-big-data-help-fight-the-zika-virus/#21a98b2ad7d9.
[4] Global Forest Watch http://www.globalforestwatch.org/.
[5] "Algorytm w wielkim mieście", Puls Biznesu (13.05.2015) http://pulsinnowacji.pb.pl/4060682,91864,algorytm-w-wielkim-miescie.

i ilości wypożyczeń, planować rozmieszczenie kolejnych stacji roweru miejskiego. Coraz częściej łączy się dane z bardzo różnych źródeł, publicznych i prywatnych, takich jak np. The Weather Company[6]. To największe na świecie prywatne przedsiębiorstwo zajmujące się zagadnieniami związanymi z pogodą. Dziennie przygotowuje nawet 26 miliardów (!) spersonalizowanych prognoz, które trafiają do indywidualnych konsumentów i firm, w tym np. linii lotniczych, gdzie informacje pogodowe są kluczowe dla bezpieczeństwa pasażerów i efektywności biznesu.

- **TDM stymuluje innowacje.** Testowanie wytrzymałości nowych materiałów czy zaawansowane prognozowanie pogody - to wszystko jest możliwe właśnie dzięki masowej analizie danych. W czerwcu 2016 roku na Uniwersytecie Warszawskim[7] ruszył jeden z najnowocześniejszych ośrodków analizy danych. Zainstalowane tam superkomputery analizują duże zasoby danych w czasie rzeczywistym m.in. na potrzeby sektora energetycznego. Prowadzone tam projekty na styku biznesu i technologii dotyczą m.in. projektowania kształtu śmigieł turbin wiatrowych.
- **Z TDM stykamy się wszyscy**, choć nie zdajemy sobie z sprawy, jak bardzo ułatwia nam codzienne życie. Korzystając z wyszukiwarek w cyfrowych archiwach (komercyjnych, ale też niekomercyjnych jak np. POLONA lub NINateka) mamy do czynienia z niczym innym jak eksploracją tekstu. Nawet kiedy wpisujemy zapytanie w wyszukiwarkę internetową to przecież specjalny algorytm analizuje za nas zasoby sieci i na tej podstawie prezentuje wyniki.

3. Diagnoza - wyzwania związane z TDM

Aby wykorzystać potencjał, jaki drzemie w TDM, trzeba zniwelować bariery, które utrudniają analizę danych i tekstu, dostępnych poprzez internet. Zmiany są potrzebne w następujących obszarach:

- **kwestie prawno-autorskie, dotyczące zasad korzystania z tekstu i danych** – zwłaszcza jeśli chodzi o zasoby dostępne w internecie. Brak jasności co do autorstwa i zasad na jakich te zasoby są udostępniane zniechęcają do ich przetwarzania z obawy przed łamaniem prawa. Dotyczy to nie tylko samych tekstów, ale również filmów, muzyki, grafik czy bazy danych. Niejasne są tutaj zasady samego korzystania z utworów jak i wprowadzania w nich modyfikacji czy kopiowania (np. ściągania na dysk) - co ma kluczowe znaczenie, kiedy mowa

[6] The Weather Company http://www.theweathercompany.com/company/worlds-largest-private-weather-enterprise.
[7] Nowe superkomputery na UW będą analizować dane m.in. dla energetyki, PAP (16.06.2016) http://naukawpolsce.pap.pl/aktualnosci/news,410164,nowe-superkomputery-na-uw-beda-analizowac-dane-min-dla-energetyki.html.

o TDM. Problem mają np. biblioteki, które nawet jeśli mają licencję pozwalającą na dostęp do komercyjnie bazy danych, nie mogą na niej wykonywać TDM.

- **ochrona baz danych** - prawo dodatkowo chroni uporządkowane zbiory danych, na których stworzenie poniesiony został istotny nakład inwestycyjny. W praktyce eksplorując daną bazę trudno stwierdzić, czy jest ona objęta ochroną czy nie. Ponadto twórca bazy danych ma wyłączne prawo pobierania danych i ich ponownego wykorzystania i może tego zabronić osobom trzecim. Jednak żaden z tych zapisów nie daje pełnej jasności w kontekście legalności TDM.
- **bariery techniczne w dostępie do danych** - obecnie dane są bardzo często publikowane w nieprzeszukiwalnych formatach i o zamkniętym dostępie (np. format PDF dla plików tekstowych i danych liczbowych, skany dokumentów), udostępniane są też bez wcześniejszego odpowiedniego przygotowania, co negatywnie wpływa na ich jakość (np. brak odpowiedniej struktury pliku, nieczyszczenie danych, brak ich weryfikacji). Udostępniane zasoby nie mają też odpowiednich opisów, w tym metadanych. Brakuje świadomości, że nie wystarczy plik udostępnić, trzeba to robić też w odpowiedniej formie (otwartej), tak aby można było z niego korzystać.
- **dobre praktyki** - brakuje wzorców, przykładów udanych przedsięwzięć, które motywowałyby do dzielenia się danymi, ich analizowania i pokazywania korzyści. Działają takie portale jak np. danepubliczne.gov.pl, to jednak w Polsce nadal instytucje tak publiczne, jak i niepubliczne (przedsiębiorstwa) unikają dzielenia się informacjami i danymi. W administracji publicznej wciąż dominuje przekonanie, że nadmierna otwartość nie jest niczym dobrym. Przedsiębiorstwa z kolei wolą nie opowiadać o tym, z jakich korzystają danych i jak to robią, dlatego że obawiają się się po pierwsze reakcji klientów, a po drugie utraty przewagi konkurencyjnej. Brakuje również wiedzy na temat tego, jak można wykorzystywać TDM, oraz że są do tego odpowiednie narzędzia, dostępne za darmo.

4. Rekomendacje

- **uporządkowanie kwestii prawno-autorskich, dotyczących zasad dostępu do tekstu i danych**
    - **zapewnienie swobody prowadzenia TDM z wykorzystaniem zasobów** objętych prawem autorskim na potrzeby TDM - na przykład w ramach dozwolonego użytku, tak aby TDM mógł być wykonywany bez zgody autora, nieodpłatnie i również do celów komercyjnych.
    - **Wolne licencje jako standard w przypadku zamawiania utworów przez instytucje publiczne.** Postulujemy, by w przypadku publikowania zasobów publicznych zamówionych u zewnętrznych podmiotów (np.

ekspertów spoza administracji) standardem było wolne licencjonowanie (np. CC-BY bądź CC-BY-SA). To automatycznie umożliwi TDM danych publicznych i wyznaczy dobry kierunek. Rekomendujemy również określenie standardów zapewniających jasne oznaczanie stanu prawnego zasobów. W przypadku treści nieobjętych prawem autorskim rekomendujemy stosowanie opracowanego przez Creative Commons Oznaczenia Domeny Publicznej.

- **uporządkowanie kwestii związanych z ochroną baz danych**
  - **potwierdzenie dopuszczalności metod pozwalających na nieodpłatne wykorzystanie zasobów** objętych prawem do baz danych na potrzeby TDM - **na** przykład w ramach dozwolonego użytku (jeśli dane nie stanowią istotnej części zbioru, ani nie wiąże się to z rażącym naruszeniem prywatności), tak aby TDM mógł być wykonywany bez zgody autora, nieodpłatnie i również dla celów komercyjnych.
- **likwidacja barier technicznych w dostępie do danych**
  - **Jednolite standardy udostępniania** - postulujemy upowszechnianie stosowania jednolitych standardów udostępniania danych i tekstu (np. Five Stars Open Data) które zawierałyby wytyczne dotyczące formatów plików. Celem jest, aby dane i teksty były publikowane w otwartych formatach dostępnych dla wszystkich i umożliwiających maszynowy odczyt[8].

5. Materiały Centrum Cyfrowego

- Future TDM strona projektu, którego partnerem jest Centrum Cyfrowe, a jego celem jest zidentyfikowanie przeszkód (prawnych, politycznych i organizacyjnych) utrudniających wykorzystania pełnego potencjału TDM[9].
- Eksploracja danych - artykuł na stronie Centrum Cyfrowego na temat tego, czym jest TDM[10].
- Eksploracja tekstu i danych – bariery prawne w Europie i Polsce - artykuł na stronie Centrum Cyfrowego na temat barier prawnych w wykorzystaniu TDM[11].

[8] http://5stardata.info/en/.
[9] http://project.futuretdm.eu/.
[10] http://centrumcyfrowe.pl/projekty/future-tdm/.
[11] http://centrumcyfrowe.pl/eksploracja-tekstu-i-danych-bariery-prawne-w-europie-i-polsce/.

6. Materiały zewnętrzne

- OpenMinted.eu - strona projektu poświęconego stworzeniu otwartej, zorientowanej na usługi infrastruktury (platformy) pozwalającej na TDM treści naukowych i edukacyjnych[12].
- Deklaracja Haska o dostępie wiedzy w środowisku cyfrowym, której jesteśmy sygnatariuszami[13].

[12] http://openminted.eu/about/overview/.
[13] http://thehaguedeclaration.com/.

Centrum Cyfrowe pracuje na rzecz zmiany społecznej wykorzystując potencjał technologii cyfrowych. Skupiamy się na edukacji i kulturze, promując otwartość: współpracę opartą na dzieleniu się zasobami i wiedzą. Przełączamy społeczeństwo na cyfrowe. www.centrumcyfrowe.pl

Publikacja jest dostępna na licencji Creative Commons Uznanie Autorstwa 4.0 Międzynarodowa pewne prawa zastrzeżone na rzecz Centrum Cyfrowego.
Pełna treść licencji jest dostępna na stronie https://creativecommons.org/licenses/by/4.0/legalcode.pl.
Zezwala się na dowolne wykorzystywanie treści publikacji pod warunkiem wskazania autorstwa Centrum Cyfrowego oraz podania informacji o licencji.