

Przetwarzanie tekstu naturalnego i słowosieci

Na Wydziale Zarządzania i Inżynierii Produkcji powstał zespół, którego celem jest integracja danych o działalności przedsiębiorstw i ich wykorzystanie do uczenia maszynowego. Wykonanie projektu wymaga połączenia teorii finansów z językoznawstwem oraz wykorzystania komputerowych metod automatycznego przetwarzania tekstu.

Przedsiębiorstwa co roku udostępniają sądom rejestrowym i inwestorom ogromną ilość danych w sprawozdaniach: sprzedaż, marże, majątek czy zadłużenie. Na ich podstawie można ocenić spółkę lub sprawdzić sytuację w branży, a nowoczesne techniki pozwalają na analizy dużych zbiorów takich danych. Szczególnym wyzwaniem jest analiza danych niestrukturyzowanych, takich jak teksty zawarte w sprawozdaniach.

Finansowanie badań w projekcie *Ewolucja języka sprawozdawczości spółek giełdowych: analiza tekstu z wykorzystaniem metod komputerowych* zapewniło Narodowe Centrum Nauki w konkursie OPUS.

W zespole kierowanym przez dr. hab. Karola Marka Klimczaka, prof. PŁ z Katedry Zarządzania uczestniczą inż. Jan Makary Fryczak i mgr inż. Artur Kaużyński

otrzymujący stypendia NCN. Przy publikacjach współpracuje dr Dominika Hadro z Uniwersytetu Ekonomicznego we Wrocławiu, specjalistka w zakresie informacji na rynkach finansowych i analizy korpusowej (polega ona na poszukiwaniu typowych użyciu słów oraz informacji o ich znaczeniu i funkcji). Wsparcie w przetwarzaniu języka naturalnego zapewnia Politechnika Wrocławska w ramach infrastruktury naukowej CLARIN-PL, która umożliwia badaczom wygodną pracę z bardzo dużymi zbiorami tekstów.

Duże zbiory danych o zarządzaniu

Zespół tworzy programy do wyszukiwania, przetwarzania i analizy danych o przedsiębiorstwach, bazując na języku programowania Python. Jednym z efektów projektu będą obszerne zbiory danych udo- ►

- steepnione w repozytorium naukowym „Polski Korpus Komunikacji Finansowej”. Według dr. hab. Klimczaka korpus ułatwi dostęp do informacji nt. polskiej gospodarki – *Na świecie powstają zbiory dokumentów, korpusy, które różni badacze mogą łatwo wykorzystać w systemach uczenia maszynowego czy sztucznej inteligencji. Korzystają z nich też programiści rozwijający systemy tłumaczeń czy analizy gospodarcze. Chcemy, żeby mieli także łatwy dostęp do polskich danych.*

Przygotowując pracę magisterską o zmianach strategicznych w zarządzaniu inż. Fryczak stanął przed ciekawym problemem – jak zidentyfikować i zmierzyć różne cechy przedsiębiorstw. Weryfikując dostępne studia przypadku zauważył, że dobrze opisują one pojedyncze organizacje, ale nie zawierają danych potrzebnych do analiz statystycznych. – *Metody kwestionariuszowe mogą nie być miarodajne, bo polegają na opinii jednego pracownika. Sprawozdania finansowe mógłbym łatwo wykorzystać do badań statystycznych, ale brakuje w nich informacji o organizacji i zarządzaniu. Postanowiłem więc wzbogacić dane finansowe o nieustrukturyzowane dane tekstowe pochodzące z raportów, komunikatów, mediów, czy sieci społecznościowych.*

Słowosieć

Słowosieć (<http://plwordnet.pwr.wroc.pl>) to opracowana na Politechnice Wrocławskiej baza danych złożona ze słów pogrupowanych według znaczenia w zbiór synonimów. Każda grupa znaczeniowa jest



Od lewej: inż. Jan Makary Fryczak, dr hab. inż. Karol Marek Klimczak, prof. PŁ i mgr inż. Artur Kaużyński
foto: arch. prywatne

rzutowana na odpowiadającą jej grupę synonimów w Princeton WordNet, skąd można przejść na kolejne języki. Jak wyjaśnia dr hab. Klimczak, prof.

► c.d. na str. 54

Przetwarzanie tekstu naturalnego i słowosieci

PŁ – Pracujemy na danych polsko-angielskich i francusko-angielskich pochodzących z Kanady. Przechodząc z angielskiego na inny język poszerzamy stosowanie najlepszych metod analizy tekstu na nowe kraje.

Transformacja cyfrowa

W przyszłości metody opracowane przez zespół PŁ mają pomóc w badaniach nad rozwojem gospodarki opartej na danych. Unia Europejska ogłosiła inicjatywę „Cyfrowa Dekada”, planując finansowanie badań, strategicznej produkcji i rozwoju cyfrowych usług dla

ludności. Przedstawiono również projekt pierwszego na świecie prawa regulującego wykorzystanie sztucznej inteligencji. Horyzont Europa przewiduje szereg strumieni finansowania na badania w klastrze „Digital, Industry and Space”. Zespół planuje uczestniczyć w międzynarodowych badaniach nad metodami zarządzania dostosowanymi do wyzwań cyfrowej rzeczywistości.

■ Artur Kaużyński
Katedra Zarządzania