Instytut Fizyki Jądrowej
im. Henryka Niewodniczańskiego
Polskiej Akademii Nauk

ul. Radzikowskiego 152
31-342 Kraków
centrala: +48 12 6628000
dyrektor: +48 12 6628200
fax: +48 12 6628458
www.ifj.edu.pl
dyrektor@ifj.edu.pl

Kraków, 11 kwietnia 2019

Tekst jak sieć: Ile wyrazów wystarczy, by rozpoznać autora?

Jesteśmy bardziej oryginalni niż sądzimy, sugerują analizy tekstów literackich przeprowadzone nową metodą stylometrii, zaproponowaną przez naukowców z Instytutu Fizyki Jądrowej PAN w Krakowie. Indywidualność autora widać już w powiązaniach między zaledwie kilkunastoma wyrazami tekstu angielskiego. W językach słowiańskich do identyfikacji twórcy wystarcza nawet mniejsza liczba wyrazów, a na dodatek wynik jest pewniejszy.

Ustalenie, kto jest autorem tekstu, na ogół nie jest trudne – wystarczy przeczytać podpis. Zdarza się jednak, że podpisu nie ma, ponieważ się nie zachował lub został przez autora z premedytacją pominięty. Nierzadko też zamiast imienia i nazwiska widzimy pseudonim. Jak więc zweryfikować, spod czyjego pióra wyszedł historyczny tekst znany jedynie z fragmentów? Jak ustalić rzeczywistego twórcę internetowego paszkwilu? Jak naprawdę wiarygodnie stwierdzić, czy tekst pracy magisterskiej bądź doktorskiej nie jest plagiatem? Tradycyjne metody stylometryczne w wielu przypadkach zawodzą lub nie prowadzą do dostatecznie pewnych wniosków. Na łamach czasopisma „Information Sciences" naukowcy z Instytutu Fizyki Jądrowej Polskiej Akademii Nauk (IFJ PAN) w Krakowie przedstawili własne narzędzie statystyczne do analizy stylometrycznej. Skonstruowane z użyciem grafów, pozwala ono spojrzeć na strukturę tekstów w jakościowo nowy sposób.

„Wnioski płynące z naszych badań z jednej strony są budujące. Wskazują bowiem, że indywidualność każdej osoby przejawia się wyraźnie w sposobie używania już zaskakująco małej liczby wyrazów. Ale jest i druga, ciemniejsza strona medalu. Skoro bowiem okazujemy się tak oryginalni, będzie nas można łatwiej identyfikować po wypowiedziach", mówi prof. dr hab. Stanisław Drożdż (IFJ PAN, Politechnika Krakowska).

Stylometria – czyli nauka zajmująca się wyznaczaniem statystycznych charakterystyk stylu tekstów – opiera się na spostrzeżeniu, że każdy z nas nieco inaczej używa nawet tego samego języka. Jedni mają szerszy zasób słownictwa, inni węższy, ktoś lubi stosować pewne sformułowania i popełnia błędy, ktoś inny unika powtórzeń i jest purystą językowym. A gdy piszemy, różnimy się też sposobem stosowania znaków interpunkcyjnych. W typowym podejściu stylometrycznym zazwyczaj bada się podstawowe cechy tekstu, np. częstotliwość występowania poszczególnych wyrazów, interpunkcję zaś się ignoruje. Analizy są przeprowadzane dla badanego tekstu oraz dla tekstów napisanych przez potencjalnych, dobrze znanych autorów. Za twórcę uznaje się tę osobę, której dzieła mają parametry o wartościach najbardziej zbliżonych do otrzymanych dla identyfikowanego materiału.

„My zaproponowaliśmy, żeby charakterystycznych cech stylu szukać w sieciowej reprezentacji tekstu, za pomocą grafów", wyjaśnia Tomasz Stanisz, doktorant IFJ PAN i pierwszy autor

publikacji, po czym precyzuje: „Graf to zbiór punktów, czyli wierzchołków grafu, połączonych liniami, czyli krawędziami grafu. W najprostszym przypadku – w tak zwanej sieci nieważonej – wierzchołki odpowiadają poszczególnym wyrazom i są połączone krawędziami wtedy i tylko wtedy, gdy w tekście dane dwa wyrazy przynajmniej raz wystąpiły obok siebie. Na przykład dla zdania 'Ala ma kota' graf miałby trzy wierzchołki, po jednym dla każdego wyrazu, ale krawędzie byłyby tylko dwie, jedna między 'Ala' a 'ma', druga między 'ma' a 'kota'".

Podczas konstruowania swoich narzędzi stylometrycznych badacze z IFJ PAN testowali różne rodzaje grafów. Najlepsze wyniki otrzymano dla grafów ważonych, a więc takich, w których każda krawędź niesie informację o liczbie wystąpień odpowiadającego jej połączenia między wyrazami. W takich sieciach najbardziej przydatne okazały się dwa parametry: krotność węzłów i współczynnik gronowania. Pierwszy z nich opisuje liczbę krawędzi wychodzących z danego węzła i bezpośrednio wiąże się z liczbą wystąpień danego wyrazu w tekście. Z kolei współczynnik gronowania opisuje prawdopodobieństwo tego, że dwa wyrazy połączone krawędzią z danym wyrazem są połączone krawędzią także między sobą.

Za pomocą tak przygotowanych narzędzi statystycznych krakowscy fizycy przyjrzeli się 96 książkom: po sześciu powieściom ośmiu znanych autorów angielskich (Austen, Conrad, Defoe, Dickens, Doyle, Eliot, Orwell, Twain) i ośmiu polskich (Korczak, Kraszewski, Lam, Orzeszkowa, Prus, Reymont, Sienkiewicz, Żeromski). W gronie autorów było dwóch laureatów literackiej Nagrody Nobla (Władysław Reymont i Henryk Sienkiewicz). Wszystkie teksty pobrano z serwisów Project Gutenberg, Wikisources i Wolne Lektury. Grupa z IFJ PAN sprawdzała następnie, z jaką wiarygodnością można w ramach jednego języka stwierdzić autorstwo 12 losowo wybranych dzieł, traktując pozostałą część puli utworów jako materiał do porównań.

„W przypadku tekstów angielskich identyfikowaliśmy autorów poprawnie w niemal 90% przypadków. Na dodatek by osiągnąć sukces należało prześledzić powiązania między zaledwie 10-12 wyrazami badanego tekstu. Wbrew naiwnej intuicji, dalsze zwiększanie liczby badanych wyrazów nie podnosiło znacząco skuteczności metody", mówi Tomasz Stanisz.

W języku polskim ustalenie autorstwa okazało się jeszcze prostsze: wystarczało prześledzić powiązania zaledwie 5-6 wyrazów. Co szczególnie ciekawe, mimo dwukrotnie mniejszej niż w języku angielskim puli istotnych wyrazów, prawdopodobieństwo poprawnej identyfikacji wzrastało – nawet do 95%! Tak wysoka poprawność diagnoz była jednak osiągana tylko wtedy, gdy jako osobne wyrazy traktowano także znaki interpunkcyjne. W obu językach pominięcie interpunkcji skutkowało wyraźną redukcją liczby poprawnych odgadnięć. Zaobserwowana rola interpunkcji to kolejne potwierdzenie wniosków z publikacji grupy prof. Drożdża z 2017 roku, gdzie wykazano, że interpunkcja pełni w języku rolę równie ważną jak same wyrazy.

„W porównaniu z językiem angielskim język polski wydaje się dawać większe możliwości ujawniania się stylu autora. Sądzimy, że podobną cechą charakteryzują się również pozostałe języki słowiańskie. Angielski jest bowiem językiem pozycyjnym, co oznacza, że istotna jest w nim kolejność wyrazów w zdaniu. Taki język pozostawia mniej miejsca na indywidualny styl wypowiedzi niż języki słowiańskie, w których o roli słowa czy wyrazu w zdaniu decyduje fleksja, czyli odmiana. Dopuszcza ona bowiem większą swobodę organizacji kolejności wyrazów w zdaniu przy niezmienionym jego znaczeniu", podsumowuje prof. Drożdż.

KONTAKT:

prof. dr hab. **Stanisław Drożdż**
Instytut Fizyki Jądrowej Polskiej Akademii Nauk, Politechnika Krakowska
tel: +48 12 6628220
email: stanislaw.drozdz@ifj.edu.pl

PUBLIKACJE NAUKOWE:

1. „Linguistic data mining with complex networks: A stylometric-oriented approach”
T. Stanisz, J. Kwapień, S. Drożdż
Information Sciences 482 (2019) 301–320
DOI: https://doi.org/10.1016/j.ins.2019.01.040

2. „In narrative texts punctuation marks obey the same statistics as words”
A. Kulig, J. Kwapień, T. Stanisz, S. Drożdż
Information Sciences 375 (2017) 98–113
DOI: http://dx.doi.org/10.1016/j.ins.2016.09.051

POWIĄZANE STRONY WWW:

http://www.ifj.edu.pl/
Strona Instytutu Fizyki Jądrowej Polskiej Akademii Nauk.

http://press.ifj.edu.pl/
Serwis prasowy Instytutu Fizyki Jądrowej PAN.

MATERIAŁY GRAFICZNE:

IFJ190411b_fot01s.jpg **HR:** http://press.ifj.edu.pl/news/2019/04/11/IFJ190411b_fot01.jpg
Autora niepodpisanego tekstu można zidentyfikować analizując zależności między zaledwie kilkoma wyrazami tekstu, wykazali fizycy-statystycy z Instytutu Fizyki Jądrowej Polskiej Akademii Nauk w Krakowie. (Źródło: IFJ PAN)