Maciej Tanaś, Mariusz Kamola
Rafał Lange, Mariusz Fila

BigData w edukacji

CONTENT 1.0 – prototyp aplikacji do analizy treści internetu

WYDAWNICTWO AKADEMII PEDAGOGIKI SPECJALNEJ

NASK
PAŃSTWOWY INSTYTUT BADAWCZY

BigData w edukacji

CONTENT 1.0 – prototyp aplikacji do analizy treści internetu

NASK Naukowa i Akademicka Sieć Komputerowa
Państwowy Instytut Badawczy

Maciej Tanaś, Mariusz Kamola
Rafał Lange, Mariusz Fila

BigData w edukacji

CONTENT 1.0 – prototyp aplikacji do analizy treści internetu

Słowo wstępne – Marcin Bochenek
Dyrektor Pionu Rozwoju Społeczeństwa Informacyjnego NASK PIB

WYDAWNICTWO AKADEMII PEDAGOGIKI SPECJALNEJ

Recenzenci:
dr hab. Barbara Galas, prof. UKSW
dr hab. Jan Łaszczyk, prof. APS

Zespół badawczy:
dr hab. Maciej Tanaś, prof. APS – kierownik
dr inż. Mariusz Kamola
dr Rafał Lange
mgr Mariusz Fila

Projekt okładki
Anna Gogolewska

Ilustracja na okładce
lightwise/123RF

Redakcja
Monika Bielska-Łach

Korekta
Zespół

Wydanie I

Warszawa 2019

ISBN: 978-83-66010-29-1

SPIS TREŚCI

SŁOWO WSTĘPNE

Marcin Bochenek
Dyrektor Pionu Rozwoju Społeczeństwa
Informacyjnego NASK PIB

Rozwój cyfrowego świata wpływa na nasze codzienne życie. Zmiany cywilizacyjne, społeczne postępują z nieznaną dotąd prędkością. Mamy dostęp do ogromnej, stale rosnącej, liczby danych. W ciągu krótkiego okresu internet, będący teraz siecią naukową, a wcześniej systemem przygotowywanym na potrzeby militarne, stał się medium i kreatorem naszej rzeczywistości. Jego obecność w naszym życiu publicznym, naukowym, zawodowym, prywatnym, powoduje, że stał się światem wymykającym się z ram opisu, których używamy do tworzenia obrazu świata.

Dzisiejsza rzeczywistość, dynamicznie zmieniająca się, charakteryzująca się ciągłym ruchem, stanowi wyzwanie dla naukowców. Także dla specjalistów w dziedzinie badań społecznych. Opisanie, dostrzeżenie, a wreszcie zrozumienie współczesności ma kluczowe znaczenie dla naszej teraźniejszości i przyszłości. Nie jest to znaczenie *stricte* poznawcze. To również materiał mogący wspomagać konkretne decyzje i rozwiązania dotyczący przyszłości, a być może nawet być albo nie być naszej cywilizacji. Bo przecież już dziś sztuczna inteligencja i internet to nie twory *science fiction*, a realne

rozwiązania istniejące w naszym świecie. W tych dziedzinach jesteśmy na początku drogi, ale jej kierunek powinny wyznaczać nie tylko możliwość techniczne, lecz także, a może przede wszystkim, zasady kreowane w ramach życia społecznego, w polityce. Nauki społeczne mają w tym procesie szczególne znaczenie. Z jednej strony oczywiście uczeni nie są powołani do samodzielnego kształtowania świata, z drugiej zaś ich wiedza i badania powinny pomagać w budowaniu i realizacji koncepcji rozwoju.

Obecnie nauki społeczne to także analityka, badania oparte na rozwiązaniach *stricte* technologicznych. Prezentowane opracowanie stanowi przyczynek w realizacji tego procesu. Jest to propozycja rozwiązania, które wprowadzane w świat analiz społecznych, może dać konkretne rezultaty. To droga do analizy danych, których sprawdzanie, analizowanie metodami znanymi od wieków byłoby w tej skali niemożliwe. To także otwarcie drogi do dalszych prac badawczych właśnie nad analizą wielkich zbiorów danych i informacji.

Dziś definiujemy problemy, opisujemy środowisko, ekosystem, w którym funkcjonujemy, wskazujemy wstępne rozwiązania i stawiamy kolejne pytania. Opisywany projekt jasno i klarownie wskazuje, że w naukach społecznych dziś potrzebne są na pewno dwa elementy. Rzetelna analiza, przetwarzanie dużych ilości danych, ale także praca naukowców nad otrzymanymi wynikami, stawianie hipotez, ich weryfikacja, wreszcie precyzowanie wniosków, prognoz i zaleceń. Dziś badacze procesów społecznych otrzymują do ręki narzędzia dające ogromne możliwości, ale w ostatecznym rachunku to po ich stronie leży wykorzystanie efektów ich działania i opisanie świata.

1

BIG DATA – INFORMATYKA W METODOLOGII NAUK PEDAGOGICZNYCH

Maciej Tanaś

Współczesny człowiek funkcjonuje w dwóch, przenikających się wzajem przestrzeniach: realnej i wirtualnej. Pierwsza – fizyczna, świat życia i śmierci, ale i bogactwa doznań polisensorycznych, to przestrzeń tętniąca kolorami i kształtami, świat zapachów, smaków i dotyku, łez i miłości. Druga – wirtualna, zrodzona z odwiecznych marzeń człowieka o likwidacji barier czasu, odległości i nadmiernego trudu, to świat dźwięków oraz barwnych, statycznych i ruchomych obrazów. To przestrzeń oplatająca ziemski glob siecią ludzkich konfliktów i twórczości, ale też pole wojen gospodarczych, ideologicznych i politycznych, agora społecznego dyskursu, świat marzeń, bazar handlu ludzkimi organami, globalny rynek przedsięwzięć ekonomicznych i – pole człowieczych podłości.

O ile pierwsza z owych przestrzeni, ta fizyczna – od wieków jest legislacyjnie kodyfikowana, o tyle druga – wirtualna, występując pod złudnym parasolem nieograniczonej wolności, jest miejscem kontroli człowieka i zbiorowości społecznych, które on tworzy.

Jest przestrzenią ludzkiej aktywności, ale też pracy służb policyjnych oraz bezwzględnych, lokalnych i globalnych działań bandytów. Zbyt często ginie w niej człowiek, zaplątany i zniewolony w rynsztoku informacji oraz migotliwych, cyfrowych obrazów. Zbyt często…

Ta przestrzeń wymaga mądrych działań prawnych. Działań prowadzonych nie w imię kontroli człowieka, zwiększania cenzury i uprawnień policji oraz służb specjalnych, nie w imię ograniczenia jego twórczej i społecznej aktywności, lecz w imię odwiecznych praw do ludzkiej godności i bezpieczeństwa, do życia i ochrony zdrowia, do wolności myśli, sumienia i wyznania, do podmiotowości prawnej i szczęścia. Tymczasem demokratycznej idei powszechnego dostępu do dobra wspólnego: informacji, wiedzy i innych osiągnięć, przeciwstawia się toczona w cyberprzestrzeni bezwzględna gra interesów, jakże często naruszająca te i inne prawa człowieka.

Sieć internetowa jest globalnym medium, przez które przepływają niezliczone informacje. Jak ujawnił roczny raport NASK: *z internetu korzysta 3,8 miliarda ludzi, czyli mniej więcej połowa ludzkości. Co roku przybywa na świecie kolejne 83 miliony ludzi, ale użytkowników internetu zdecydowanie więcej, bo ponad 354 milionów rocznie*[1].

Za utrzymanie bezpieczeństwa i stabilności internetu w sensie spójnej adresacji usług i urządzeń *odpowiedzialna jest ICANN* (ang. *Internet Corporation for Assigned Names and Numbers*), która zarządza parametrami technicznymi sieci, decyduje o transporcie cyfrowym

[1] Anna Gniadek, Weronika Rakowska, Tomasz Szladowski, *Rynek nazw domeny.pl. Raport roczny.* Wersja elektroniczna zob.: https://www.dns.pl/NASK-raport-rynek-nazw-domeny-pl-2017.pdf, dn. 10.07.2018.

i odpowiada za adresacje ruchu[2]. Nie jest to jednak jedyna organizacja, dbająca o zasady funkcjonowania sieci. Do tej roli pretendują także globalne korporacje, związane z rynkiem cyfrowym i dysponujące olbrzymim kapitałem. Zabieganie przez nie o wpływ na kształtowanie zasad regulujących funkcjonowanie internetu oraz na sprawowanie nad nim kontroli jest ich żywotnym interesem i nie powinno to nikogo dziwić, że podejmują je Facebook, Google i inni cyfrowi giganci.

Permanentnie rosnące przypływy informacji pochodzące z różnych źródeł, a zatem o różnej charakterystyce, a także ich rosnąca użyteczność dla różnych obszarów nauki, zarządzania, administracji, usług i produkcji wywołują pilną potrzebę tworzenia nowych technik analizy danych oraz rozwiązań technologicznych i sprawiają, że Big Data stały się jednym z najważniejszych współcześnie wyzwań informatycznych. Rodzą się problemy równoległego przetwarzania danych oraz odejścia od klasycznego schematu ich przechowywania, a także zróżnicowania danych, ich wolumenu, redukcji wymiaru i jakości oraz możliwości wnioskowania.

Współczesne urządzenia mobilne: laptopy, tablety oraz smartfony i coraz liczniejsze urządzenia przenośne (*Wearable Computers*), wzrost pamięci masowej w chmurze, jak również rozwijające się pola zastosowań: rozszerzona rzeczywistość (*Augmented Reality*), sztuczna inteligencja (*Artificial Intelligence*) oraz internet rzeczy (*Internet of Things*), przynoszą dane o coraz większej złożoności, o nowych formach i źródłach pochodzenia. Do analizy bardzo dużych, różnorodnych zbiorów danych semistrukturalnych, prawiestrukturalnych i niestrukturalnych, pochodzących z różnych źródeł

[2] Anna Gniadek: *Internet? Kto tu rządzi?* [w:] Anna Gniadek, Weronika Rakowska, Tomasz Szladowski, *Rynek nazw domeny.pl. Raport roczny…*, *op. cit*., s. 18.

i w różnych rozmiarach (od terabajtów do zettabajtów, tj. od 10^{12} do 10^{21} bajtów[3]), stosuje się coraz bardziej zaawansowane techniki analityczne.

Big Data to termin stosowany do takich zestawów danych, których rozmiar lub typ wykracza poza zdolność do przechwytywania, zarządzania i przetwarzania za pomocą tradycyjnych algorytmów i relacyjnych baz danych. Dane te posiadają jedną lub więcej z następujących cech: dużą objętość (*high volume*), dużą intensywność strumienia (*high velocity*), dużą różnorodność (*high variety*) lub zróżnicowaną wiarygodność (*high veracity*)[4]. Big Data pochodzą z czujników, urządzeń multimedialnych, dzienników aktywności programów komputerowych, aplikacji transakcyjnych, stron internetowych i mediów społecznościowych – większość z nich generowana jest w czasie rzeczywistym i na bardzo dużą skalę.

Definicja Big Data przez lata ewoluowała od takich, które koncentrowały się na desygnatach nazwy, do tych, które odwoływały się do jej konotacji. Przykładem definicji pierwszego typu jest ta, którą zaproponowali Michael Cox i David Ellsworth. Ich zdaniem Big Data to po prostu duże dane, których liczbę należy maksymalizować dla wydobycia w trakcie analizy ich wartości informacyjnych[5]. Podobnie

[3] Jednostki używane do określania rozmiaru największych pamięci masowych, zasobów plików i baz danych dawno przekroczyły kilobajt (10^3) i megabajt (10^6). Po gigabajtach(10^9), nastąpiły terabajty (10^{12}), petabajty (10^{15}), eksabajty (10^{18}) i zettabajty (10^{21}). Kolejne to jottabajty (10^{24}), xenottabajty (10^{27}) i shilentnobajty (10^{30}).

[4] Zob. szerzej https://www.ibm.com/analytics/hadoop/big-data-analytics, dn. 13.07.2018.

[5] Michael Cox i David Ellsworth, *Managing Big Data for Scientific Visualization*, 1997, ACM SIGGRAPH '97 Course #4, Exploring Gigabyte Datasets in Real-Time: Algorithms, Data Management, and Time-Critical Design, Los Angeles, zob.: https://www.researchgate.net/profile/David_Ellsworth2/

Avita Katal, Mohammad Wazid i R.H. Goudar wyjaśniali pojęcie przez wskazanie, że to *duża liczba danych, która wymaga zastosowania nowych technologii i architektur, tak by możliwa była ekstrakcja wartości płynącej z tych danych poprzez uchwycenie i analizę procesu*[6]. Dobrym przykładem definicji drugiego typu jest ta przyjęta w tej pracy za IBM i zacytowana wcześniej. Sformułował ją Doug Laney już w 2001 roku[7].

Analiza Big Data pozwala podejmować decyzje na podstawie danych, które wcześniej były niedostępne lub nieużyteczne. Dzięki zaawansowanym technikom analitycznym, takim jak uczenie maszynowe, analiza predykcyjna, eksploracja danych, statystyki i przetwarzanie języka naturalnego, można analizować wcześniej niewykorzystywane źródła danych niezależnie lub razem z istniejącymi i tradycyjnie dostępnymi (badania sondażowe itd.). Dzięki temu pozyskuje się nowe informacje, niezwykle użyteczne w procesie wnioskowania i podejmowania decyzji i to na wielu polach. Owa użyteczność ma swe źródło w stosunkowo niskich kosztach oraz w szybkości

---

publication/238704525_Managing_big_data_for_scientific_visualization/links/54ad79d20cf2213c5fe4081a/Managing-big-data-for-scientific-visualization.pdf, pobrane dn. 13.07.2018.

[6] Avita Katal, Mohammad Wazid, R.H. Goudar, *Big Data: Issues, Challenges, Tools and Good Practices*, 2013, Sixth International Conference on Contemporary Computing (IC3), IEEE, Noida, s. 404–409, za: Marta Tabakow, Jerzy Korczak, Bogdan Franczyk, *Big Data – definicje, wyzwania i technologie informatyczne*, „Informatyka Ekonomiczna. Business Informatics" 2014, nr 1(31), s. 141.

[7] Wspomniany autor sformułował ją jako 3V, a nie 4V, pominął bowiem zróżnicowaną wiarygodność (*high veracity*). Por. Doug Laney, *3D DataManagement: Controlling Data Volume, Velocity, and Variety*, „Application Delivery Strategies" 2001, META Group Inc. Zob.: https://blogs.gartner.com/doug-laney/files/2012/01/ ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf, pobrane dn. 13.07.2018.

pozyskiwania wiarygodnych informacji z olbrzymiej masy danych, niemożliwych do zdobycia w tej liczbie i w takim tempie przy użyciu metod tradycyjnych.

Bez wątpienia badanie zachowań i opinii użytkowników portali internetowych, a zwłaszcza sieci społecznościowych otwiera zupełnie nowe perspektywy poznawcze i to nie tylko przed globalnymi korporacjami (korzystającymi już przecież z tej drogi pozyskiwania informacji), ale też przed naukami społecznymi, w tym naukami pedagogicznymi. Możliwe jest oczywiście także wykrywanie źródeł zagrożeń społecznych i ekonomicznych, działań terrorystycznych, konfliktów politycznych i militarnych, stosunkowo precyzyjne śledzenie ich rozprzestrzeniania, skali, odbioru społecznego itd.

Jeśli teza *sieci społecznościowe mają ogromną wiedzę o naszej rzeczywistości i tym wszystkim, co się wokół nas dzieje* jest prawdziwa, a nie ma powodu, żeby w nią wątpić, to nie wystarczy prosta konstatacja, ale konieczne jest też opisanie dostępnych metod, technik i narzędzi poznania. Niezbędna jest metodologiczna refleksja nad wartością poznawczą źródeł, sposobów i dróg pozyskiwania danych, metod analizy zbieranego materiału empirycznego oraz interpretacji wyników badań. Wyzwania badawcze dotyczą także sfery technologicznej: opracowanie innowacyjnej architektury, identyfikacja źródeł danych, określenie filtrów danych, automatyczne generowanie metadanych, niezwłoczna (bieżąca) obsługa napływu nowych, strumieniowych danych i ich aktualizacja, zarządzanie stosem danych w szybkich i skalowalnych warstwach przechowywania i przetwarzanie zapytań, integracji pochodzących z różnych źródeł danych pojawiających się w różnych formatach i modelach. Jedno jest pewne – skrzynie pełne skarbów czekają na swoich odkrywców.

Według prognoz CISCO System Inc.[8] roczny globalny ruch w sieci do roku 2021 osiągnie 3,3 zettabajtów (ZB) rocznie, czyli 278 eksabajtów (EB) miesięcznie. W 2016 roku stopa realizacji dla globalnego ruchu w internecie wynosiła 1,2 ZB rocznie, czyli 96 EB miesięcznie. W ciągu najbliższych lat wzrośnie on kilkakrotnie. Miesięczny ruch w internecie z 13 GB na osobę w 2016 roku sięgnie 35 GB w 2021 roku.

Smartfony okażą się bardziej użyteczne od komputerów. Jeszcze w 2016 roku poprzez komputery odbywało się 46 procent całkowitego przepływu informacji, ale w 2021 roku będzie to już jedynie 25 procent ruchu. Równocześnie smartfony przejmą 33 procent całkowitego ruchu w internecie. Wprawdzie przepływ informacji wzrośnie do 2021 roku także na komputerach, ale dla telewizorów, tabletów, smartfonów i modułów M2M (*Machine-to-Machine*) wskaźnik wzrostu ruchu w 2021 roku będzie większy i wyniesie odpowiednio 21 procent, 29 procent, 49 procent i 49 procent. W latach 2016–2021 wzrosną też 20-krotnie współczynniki udziału w sieci wirtualnej rzeczywistości (VR) i rozszerzonej rzeczywistości (AR). Warto sobie uświadomić, że tylko obejrzenie wszystkich filmów, które będą przesyłane w sieci w każdym miesiącu 2021 roku musiałoby trwać ponad 5 milionów lat.

Ilość danych powstających i wędrujących po sieci jest porażająca. Stanowi to poważne wyzwanie technologiczne, ale jest też obiecującym polem badań. Truizmem jest stwierdzenie, że media społecznościowe są ważnym czynnikiem wpływającym na zachowania ludzkie w sieci i w świecie realnym. Jeśli tak, to z istoty rzeczy powinny stać

[8] *The Zettabyte Era: Trends and Analysis,* White Papers, Cisco,https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.html, pobrane dn. 13.07.2018.

się przedmiotem eksploracji i wnioskowania. I dodajmy – nie tylko one. Polem niezwykle interesujących poznawczo badań mogą być portale internetowe, sieć blogów itp., a nawet gwałtownie rozwijający się internet rzeczy (IoT). Już dziś liczba urządzeń podłączonych do sieci i komunikujących się ze sobą jest większa niż liczba ludzi na świecie[9]. Tymczasem internet rzeczy może być równie groźny co pomocny. Z jednej strony jest podstawą tzw. inteligentnego domu i pozwalając urządzeniom na wzajemną komunikację, poprawia komfort życia i pozorne bezpieczeństwo jego mieszkańców. Z drugiej strony jednak, na skutek niedostatecznych działań rewidujących jakość technologii, dużej liczby urządzeń i taniej przepustowości, staje się źródłem ataków cyberprzestępców, naraża na inwigilację i zagrożenie także fizycznego bezpieczeństwa[10].

Przykładem niech służy groźny, globalny atak z wykorzystaniem urządzeń IoT, który miał miejsce w 2016 roku. Złośliwe oprogramowanie umożliwiło wówczas stworzenie za pomocą botneta Mirai „armii" kamerek internetowych oraz Smart TV i sparaliżowanie takich serwisów, jak Reddit, Twitter, Spotify, Netflix, New York Times czy PayPal. W Polsce zaobserwowano wtedy nawet 14 054 przejętych urządzeń dziennie.

Warto przywołać w tym miejscu również projekt OSE – Ogólnopolskiej Sieci Edukacyjnej – jednej z najważniejszych inwestycji

[9] *Według instytutu badawczego Gartner, liczba inteligentnych urządzeń – z grupy IoT – może wzrosnąć z 8,4 mld w 2017 r. do 20,4 mld w 2021 r.* Zob. *Inteligentne urządzenia wokół nas. A co z naszym bezpieczeństwem?*, „Interia Biznes" 17.02.2018, www.biznes.interia.pl, pobrano dn. 13.07.2018.

[10] Por. *Krajobraz bezpieczeństwa polskiego internetu 2016. Raport roczny z działalności CERT Polska,* NASK/CERT Polska 2016, s. 23–29.

technologicznych i oświatowych w Polsce[11]. Koszty tego niezwykłego projektu, realizowanego przez NASK Państwowy Instytut Badawczy, zostaną pokryte z budżetu Państwa oraz Programu Operacyjnego Polska Cyfrowa. Polska wpisuje się tym projektem w trendy europejskie. Na przykład, unijny program WiFi4EU ma zapewnić obywatelom oraz osobom przybywającym do Unii Europejskiej dostęp do bezpłatnego WiFi w przestrzeniach publicznych, takich jak parki, place, budynki publiczne, biblioteki, ośrodki zdrowia. Projekt polski jest jednym z największych programów cyfryzacji szkół, podejmowanych w UE i powinien przyczynić się nie tylko do wspomagania edukacji tradycyjnej – cyfrową, lecz także zwiększyć arsenał środków, metod i treści kształcenia, umożliwić rzeczywisty rozwój kompetencji cyfrowych uczniów oraz wyrównać ich szanse edukacyjne.

Budowana błyskawicznie przez NASK sieć szkolna powinna stać się nade wszystko polem badań dla tych osób i instytucji, dla których los dzieci i młodzieży, poprawa jakości procesu nauczania-uczenia się, szerzej otwierania przed człowiekiem świata kultury i nauki oraz prowadzenia ku życiu wartościowemu i twórczemu, stanowi cel i przedmiot działania. Te badania powinny być prowadzone zgodnie z przepisami prawa i bezwzględnie respektować kodeks etyczny badań naukowych.

[11] Inicjatywa OSE została przyjęta przez Radę Ministrów 13.06.2017 r. jako Uchwała „100 Mega na 100-lecie" i ma na celu zapewnienie powszechnego i równego dostępu szkół do bardzo szybkiego (co najmniej 100 Mb/s), bezpiecznego oraz bezpłatnego internetu. Zgodnie z założeniami projektu wszystkie szkoły podstawowe i ponadpodstawowe zostaną do 2021 roku podłączone do OSE. Por. Marcin Bochenek, *Rok pilotażu OSE*, [w:] *Akademia NASK, O OSE*, https://akademia.nask.pl/projekt-48/o-projekcie.html, pobrano dn. 17.07.2018. *Ustawa o Ogólnopolskiej Sieci Edukacyjnej* została jednogłośnie przyjęta przez Senat RP 10.11.2017, a następnie podpisana przez Prezydenta RP i ogłoszona 28 listopada w Dzienniku Ustaw 2017, poz. 2184, tom 1.

Dopóki nie było stosownych i dostępnych programów informatycznych, pozwalających na intencjonalne prowadzenie badań, na ekonomicznie uzasadnione zbieranie i skuteczną selekcję Big Data, ani też metod pozwalających na analizę tak wielu danych, było to po prostu zadanie niewykonalne. Działaniom w tym zakresie nie sprzyjały także: brak powszechnej świadomości wartości poznawczej tego typu badań oraz równoczesna dominacja tradycyjnych, już sprawdzonych metod empirycznych.

Nie znaczy to, że nie było wyprzedzających prób teoretycznego opisu i wyjaśnienia problemu Big Data oraz pionierskich badań w naukach społecznych. Z pewnością największy wkład poznawczy, ale też organizacyjny i popularyzatorski w Polsce wniósł prof. dr hab. inż. Włodzimierz Gogołek. Liczne prace naukowe profesora, porywające wykłady na sympozjach oraz konferencjach naukowych, a także prekursorskie eksperymenty są kamieniami milowymi naukowych odkryć w tym zakresie[12]. Włodzimierz Gogołek jest również autorem pojęć określających w języku polskim proces i autorską metodę badań Big Data. Proces ten określił mianem rafinacji sieciowej przez analogię do procesu *oczyszczania i uszlachetniania substancji naturalnych lub produktów przemysłowych w celu nadania im odpowiedniej czystości, barwy, zapachu*[13]. Przyjęta definicja, zaczerpnięta ze

[12] Włodzimierz Gogołek, *Big Data. Sieciowe źródło informacji dla edukacji,* [w:] *Cyfrowa przestrzeń kształcenia, Seria Cyberprzestrzeń – Człowiek – Edukacja*. Tom 1. Praca zbiorowa pod red. Macieja Tanasia i Sylwii Galanciak, Oficyna Wydawnicza „Impuls", Kraków 2015, s. 97–104; tenże, *Rafinacja informacji sieciowej*, [w:] *Informatyka w dobie XXI wieku. Nauka, Technika, Edukacja a nowoczesne technologie informatyczne*. Praca zbiorowa pod red. Aleksandra Jastriebowa, Beaty Kuźmińskiej-Sołśnia, Marii Raczyńskiej, Politechnika Radomska, Radom 2011. Zob. też przywoływane w tym art. inne prace tego autora.

[13] Mieczysław Szymczak, *Słownik języka polskiego*, Państwowe Wydawnictwo Naukowe, Warszawa 1978.

Słownika języka polskiego, trafnie opisuje istotę i sposób procesu analizy Big Data, pozyskiwanych z sieci lub z dużych zbiorów informacyjnych dostępnych poza siecią[14].

Przebieg procesu rafinacji Big Data sprowadza się do kilku etapów. Po określeniu typu i zakresu materiałów źródłowych z sieci lub innego źródła, a także czasu i częstotliwości ich pobierania należy ustalić hasła, związane z badanym zjawiskiem i występujące w obsługiwanych przez system źródłach danych. Takimi hasłami mogą być słowa wraz z ich formami fleksyjnymi, wyrażenia czy też całe zestawy słów. Hasła noszą nazwę słupów. Specyfika języka polskiego powoduje, że słup może obejmować wybrane lub wszystkie możliwe odmiany słowa lub wyrażenia przez osoby, liczby, rodzaje, przypadki, czasy, tryby, strony, imiesłowy, formy bezosobowe i nieregularne, a nawet możliwe błędy ortograficzne, neologizmy i synonimy. Może też obejmować tzw. hashtagi, czyli pojedyncze słowa lub wyrażenia poprzedzone symbolem # (z ang. *hash, hashtag,* ale też *octothorp, octothorpe*, *octathorp*, *octatherp, fence, mesh*, w Singapurze *hex*, a w muzyce *sharp*), bez użycia spacji. Pełnią one funkcję nieustrukturyzowanych metadanych, ułatwiających znajdowanie wiadomości o określanym temacie lub zawartości i są używane w sieciach społecznościowych, takich jak Twitter oraz w innych usługach mikroblogowania.

Pobierane dane mogą występować w postaci artykułu (artykuł na stronie, komunikat, post), bloku (tytuł, autor, pod- i śródtytuł, tekst, podpisy, także treść komentarzy), a także pojedynczego zdania lub słowa. Kolejnym krokiem jest określenie tzw. sentymentów, będących wyrażeniami niosącymi pozytywny, neutralny lub

[14] Włodzimierz Gogołek, Paweł Kuczma, *Rafinacja informacji sieciowych na przykładzie wyborów parlamentarnych. Część 1. Blogi, fora, analiza sentymentów*, „Studia Medioznawcze" 2013, nr 2(53).

negatywny ładunek emocjonalny. Sentymenty są ocenami słupów i występującymi w lub obok badanych tekstów, obrazów, plików audio czy video. Identyfikacja sentymentów powinna być poprzedzona tzw. obróbką przygotowującą materiał źródłowy[15]. Polega ona na odfiltrowaniu treści podlegających badaniu, oczyszczeniu danych oraz przekształceniu ich do postaci czytelnej dla programu. Zebrany materiał badawczy podlega następnie analizie ilościowej (statystycznej) i jakościowej oraz – co ważne – autorskiej interpretacji uzyskanych wyników.

Jest w procesie rafinacji sieciowej Big Data urok nowości, jest jednak nade wszystko potencjał poznawczy. Wiąże się on z wielkością zbiorów danych, szybkością ich napływu oraz olbrzymią różnorodnością. Równocześnie pojawiają się coraz doskonalsze technologie ich zbierania i rafinacji, a także wspomagane informatycznie coraz skuteczniejsze i precyzyjniejsze metody ich analizowania i wnioskowania. Bez wątpienia użyteczna okazała się w tym względzie chmura obliczeniowa (*cloud computing*). Pojęcie to oznacza zazwyczaj skalowalną platformę, zawierającą sprzęt IT wraz z oprogramowaniem, dostępną u zewnętrznego operatora jako usługa internetowa. Dodajmy, że *cloud computing* oznacza również system rozproszenia, zdolność uruchamiania programu lub aplikacji na wielu połączonych komputerach w tym samym czasie lub dynamiczną obsługę danego żądania, polegającą na przydzieleniu zadania do jednego z dostępnych serwerów. Jeśli chodzi o informatyczne narzędzia użyteczne w rafinacji Big Data, to należy koniecznie wspomnieć o projekcie Stratoshpere[16]

[15] Włodzimierz Gogołek, Dariusz Jaruga, *Z badań nad systemem rafinacji sieciowej. Identyfikacja sentymentów*, „Studia Medioznawcze" 2016, nr 4(67), s. 104–105.

[16] Zob. szerzej. http://stratosphere.eu/.

oraz Apache Hadoop[17] i innych technologiach Big Data, takich jak: Apache Storm[18], Apache Kafka[19] i Apache Impala[20].

Stratosphere to projekt badawczy, którego celem było stworzenie platformy Big Data Analytics następnej generacji. Podjęły go niemieckie ośrodki akademickie: Technische Universität Berlin, Humboldt-Universität oraz Hasso-Plattner-Institut. Dzięki projektowi Stratosphere opracowano i przyczyniono się do powstania platformy, która w 2014 roku stała się projektem Apache pod nazwą Apache Flink[21].

Najczęściej wykorzystywaną stała się jednak Apache Hadoop – otwarta platforma programistyczna, napisana w języku Java, a przeznaczona do rozproszonego składowania i przetwarzania wielkich zbiorów danych przy pomocy klastrów komputerowych. Zapewne jej popularność wynika właśnie z faktu, że wspomniana platforma jest zbiorem narzędzi open-source. Projekt obejmuje obecnie: Hadoop Common, Hadoop Distributed File System, Hadoop Yarn, Hadoop MapReduce oraz inne projekty, jak: Ambari™, Avro™, Cassandra™, Chukwa™, HBase™, Hive™, Mahout™, Pig™, Spark™, Tez™ czy ZooKeeper™[22].

Obok wcześniej wymienionych tworzone są także inne architektury informatyczne, dedykowane Big Data. Opisany w prezentowanym raporcie projekt CONTENT 1.0 jest przykładem poszukiwań badawczych, które zakończyły się, zdaniem autorów, pierwszym, jeszcze wstępnym sukcesem.

---

[17] Oficjalna strona https://hadoop.apache.org/.
[18] Zob. szerzej http://storm.apache.org/.
[19] Zob. https://kafka.apache.org/.
[20] Zob. https://impala.apache.org/.
[21] Strona projektu: https://flink.apache.org/.
[22] http://hadoop.apache.org/.

Inspiracją do prac nad projektem i jego rozpoczęciem była towarzyska i jakże cenna rozmowa, dotycząca poznawczego i gospodarczego znaczenia Big Data, podjęta niegdyś z dr Agnieszką Wrońską – kierownikiem Działu Akademia NASK. Niezwykle pomocna okazała się również sugestia prof. dr hab. inż. Ewy Niewiadomskiej-Szynkiewicz – Dyrektora Pionu Naukowego NASK-PIB, żeby do zespołu zaprosić dr inż. Mariusza Kamolę, absolwenta nauk technicznych w zakresie automatyki oraz robotyki i – co było szczególnie istotne w tym przypadku – specjalistę od sieci społecznych i technologicznych. Bez rady i pomocy Pani Profesor projekt spaliłby na panewce. Ponadto do zespołu zostali włączeni także dr Rafał Lange i mgr Mariusz Fila z kierowanej przeze mnie Pracowni Edukacyjnych Zastosowań Technologii Informacyjno-Komunikacyjnych NASK-PIB. Życzliwe i mądre wsparcie Marcina Bochenka – Dyrektora Pionu Rozwoju Społeczeństwa Informacyjnego NASK-PIB spowodowało, że projekt nabrał realnych kształtów i możliwe stało się jego ostateczne opracowanie i złożenie. Wkrótce, dzięki pozytywnej ocenie Komisji Konkursowej NASK-PIB można było rozpocząć realizację pierwszego, a później drugiego etapu.

Raz jeszcze okazało się, że dzięki świadomym decyzjom ludzi dobrej woli, rzeczywistej zarządczej i organizacyjnej sprawności Dyrekcji NASK, umiejętności współpracy w zespole oraz bardzo wysokim kompetencjom jego członków w zakresie informatyki i nauk społecznych, projekt CONTENT 1.0 został zaakceptowany przez Komisję oraz podjęty i zrealizowany. Stało się tak dzięki tym wszystkim niezwykłym osobom, które rozumiały ideę, jaką się kierowaliśmy i nie pozwoliły zgasić pokładanej w projekcie nadziei. Dlatego też tym, którzy nas inspirowali i umożliwili skuteczną realizację planów, słowem, okazali nieocenioną pomoc i wsparcie, składam w imieniu własnym i wszystkich członków zespołu wyrazy głębokiej wdzięczności.

Metody analizy Big Data są poważnym wyzwaniem informatycznym. Ich opis i wyjaśnienie wydają się być istotne dla rozwoju naukowego i gospodarczego. Dostęp do informacji i możliwości przetwarzania dużych zbiorów danych o różnym typie i złożoności oraz źródłach pochodzenia jest bezcenny dla każdego przedsiębiorstwa. Przemawiających za tym i wystarczających argumentów dostarcza ekonomia, gospodarka oparta na wiedzy i praktyka społeczna. Już przecież w 2013 roku Kenneth Cukier i Viktor Mayer-Schönberger – współautorzy książki dotyczącej Big Data, dostrzegając olbrzymi wpływ tego zjawiska na gospodarkę, naukę i społeczeństwo, nadali jej znamienny tytuł: *Big Data: Rewolucja, która zmieni sposób naszego życia, pracy i myślenia*[23]. I dodajmy – już przeobraziła i zmieniać nadal będzie.

Z całą pewnością technologie Big Data tworzą też nowe, atrakcyjne perspektywy poznawcze. Przy czym nie chodzi wyłącznie o liczbę danych, ale też o ich wiarygodność, unikatowość oraz możliwość podejmowania pionierskich badań naukowych, na dotychczas nieeksplorowanych polach. Coraz bardziej prawdopodobne empirycznie staje się zatem intencjonalne wykorzystanie Big Data tak w badaniach wysokospecjalistycznych, jak i inter- czy transdyscyplinarnych.

Big Data bezsprzecznie już udowodniły swoją znaczną, naukową przydatność. Analizy obszernych zbiorów danych przyniosły atrakcyjne owoce na wielu polach: od eksplozji w biologii, wraz z jej rozrastającymi się bazami danych genomów i białek, poprzez astronomię, z petabajtami płynącymi z obserwacji nieba, do nauk społecznych, z miliardami postów i tweetów krążących w Internecie. Potok danych jest zbyt duży, by mógł go precyzyjnie analizować „nieuzbrojony"

[23] Viktor Mayer-Schönberger, Kenneth Cukier, *A Revolution that will transform how we live, work andthink*, Boston–New York 2013.

ludzki umysł, ale rozwój nauk informatycznych oraz postęp technologiczny, które pomogły w dostarczeniu tych danych, stworzyły także nowe, potężne narzędzia, które już dziś okazują się niezwykle użyteczne nie tylko w procesie zbierania i przesyłania, lecz także – analizy i zrozumienia. Nadszedł czas na podjęcie badań z wykorzystaniem Big Data także w naukach pedagogicznych. *Terra incognita* czeka na swych odkrywców. Drogę do wysp nieznanych otworzyła informatyka.

2

APLIKACJA

Wydawać by się mogło, że obecnie rynek usług analitycznych obecności i percepcji zadanych pojęć, wyrażających się w cyberprzestrzeni, jest zapełniony w stopniu odpowiadającym zupełnie potrzebom użytkowników. Istnieją na nim serwisy ukierunkowane na analizę określonych portali społecznościowych […], agregację i selekcję istotnych doniesień […] – a także aplikacje uniwersalne, dokonujące łącznej analizy wzmianek na temat zadanego pojęcia występujących w wielu różnych źródłach […]. Użytkownik otrzymuje wyniki analiz na żądanie, w atrakcyjnej wizualnie formie, albo *ad hoc*, w sytuacji pojawienia się nowego zjawiska lub istotnej zmiany jego dynamiki.

Wielość i różnorodność dostępnych aplikacji może sprawiać wrażenie, ich umiejętny wybór a następnie świadome z nich korzystanie zaspokajają obecne potrzeby analizy obecności interesujących użytkownika pojęć w internecie. W istocie tak nie jest, co najmniej z trzech powodów. Po pierwsze, istnieje potrzeba elastyczniejszej i precyzyjniejszej parametryzacji algorytmów wyszukiwania i przetwarzania danych surowych tak, odpowiadającej rzeczywistym potrzebom świadomego i wymagającego użytkownika. Po drugie, sam sposób działania i wynik algorytmów powinien być jawny (większość obecnych usług, chociażby szeregowania wyników wyszukiwania taka nie

jest). Po trzecie, łańcuch przetwarzania wyników składa się wyłącznie z algorytmów komputerowych, nie pozostawiając miejsca na ingerencję ekspertów dziedzinowych w kluczowych etapach analizy.

Funkcjonalność i związana z nią architektura systemu CONTENT 1.0 usuwają wszystkie powyższe niedostatki i stanowią jego cechy wyróżniające spośród innych istniejących rozwiązań.

2.1. Funkcjonalność

Działanie systemu można przedstawić najczytelniej, omawiając typowe scenariusze korzystania z niego przez użytkownika, czyli tzw. przypadki użycia. Użytkownik, tj. klient końcowy albo wspierający go i doradzający mu ekspert dziedzinowy, definiuje *zlecenie* analizy obecności określonego hasła w obsługiwanych przez system źródłach danych. Ponieważ jakakolwiek analiza danych wymaga ich uprzedniego zgromadzenia, system rozpoczyna okresowe skanowanie określonych w zleceniu źródeł danych i gromadzenie tych, które będą przydatne do dalszej analizy. Obecnie obsługiwanymi źródłami danych są portale twitter.com, facebook.com oraz onet.pl.

Aby udostępnić możliwość formułowania precyzyjnych i elastycznych kryteriów wyszukiwania, zaproponowano ustalony podział pobieranych danych na następujące jednostki:

- *artykuł* – odpowiada pojedynczemu artykułowi na stronie onet.pl, komunikatowi (*tweet*) w serwisie twitter.com oraz wpisowi (*post*) w serwisie facebook.com;
- *blok* – podjednostka artykułu, odpowiadająca części artykułu lub pojedynczemu komentarzowi do artykułu;

- *zdanie* – pojedyncze zdanie;
- *słowo* – pojedyncze słowo.

Podstawowym parametrem zlecenia jest *kwerenda*, czyli wyrażenie, którego wartość jest wyznaczana dla każdego napotkanego artykułu w skanowanych źródłach. Składnia kwerendy jest następująca (nawiasy kwadratowe oznaczają element opcjonalny, kreska pionowa oznacza alternatywę, [...] oznacza dowolną liczbę powtórzeń bezpośrednio poprzedzającego elementu wyrażenia):

hasło [[*op_logiczny*] *hasło* [...]]]

gdzie *hasło* ma postać:

[*id_typu_bloku*[*id_typu_bloku*[...]]]*słowo*[*końcówka*[|*końcówka*[...]][|.|**?**|*]

op_logiczny ma postać:

|

lub ma postać:

[**&**[+|-]*liczba*[**w**|**s**]]]

Wyjaśnienie oznaczeń:

- *hasło*–pojedyncze słowo wraz z jego formami fleksyjnymi,
- *op_logiczny*– złożony operator logiczny,
- *id_typu_bloku*– jednocyfrowy specyfikator, precyzujący typ bloku dokumentu, w obrębie którego poszukiwane będą hasła,
- *słowo*– część nieodmienna szukanego terminu (niekoniecznie temat gram.),
- *końcówka*– końcówka fleksyjna (dowolny ciąg znaków),
- . – wystąpienie zera lub jednego znaku,
- **?**– wystąpienie dokładnie jednego znaku,
- *– wystąpienie dowolnej liczby znaków (do separatora słowa),
- |– alternatywa (wystąpienie jednego z haseł jest wystarczające),

- **&**– koniunkcja (wystąpienie obu haseł jest konieczne),
- **+**–- następujące hasło musi występować po poprzednim,
- **-** – następujące hasło musi występować przed poprzednim,
- *liczba*– liczba słów lub zdań, w zakresie których ma nastąpić wystąpienie określone przez + lub -,
- **w** – określona liczba powyżej dotyczy słów,
- **s** – określona liczba powyżej dotyczy zdań.

Domyślne działanie polega na wyszukaniu koniunkcji haseł w formie dokładnie podanej przez użytkownika, w całym artykule i towarzyszącym mu komentarzach, bez uwzględniania kolejności występowania haseł.

Opracowana i przedstawiona tu składnia wywodzi się ze składni wyrażeń regularnych. Została ona istotnie zmodyfikowana, aby umożliwić wygodne, intuicyjne specyfikowanie wariantowego zakończenia haseł, filtrować hasła ze względu na ich położenie w artykule oraz ze względu na wzajemne oddalenie haseł w tekście. W przypadku tej ostatniej opcji, wystarczy poprzedzić wyszukiwane hasło ciągiem cyfr, np. 145pies, aby ograniczyć wyszukiwanie wystąpienia słowa pies do trzech typów bloków, identyfikowanych cyframi 1, 4 i 5. W odniesieniu do wszystkich rodzajów źródeł, przyjęto podział artykułu na bloki następujących typów:

0. Tytuł artykułu
1. Autor artykułu
2. Podtytuły i śródtytuły
3. Tekst zasadniczy (pomiędzy tytułami)
4. Podpisy pod infografikami
5. Treści komentarzy

Oszacowanie wartości kwerendy dla konkretnego artykułu zwraca wartość całkowitą. Jeśli struktura zapytania powoduje, że ostatnim

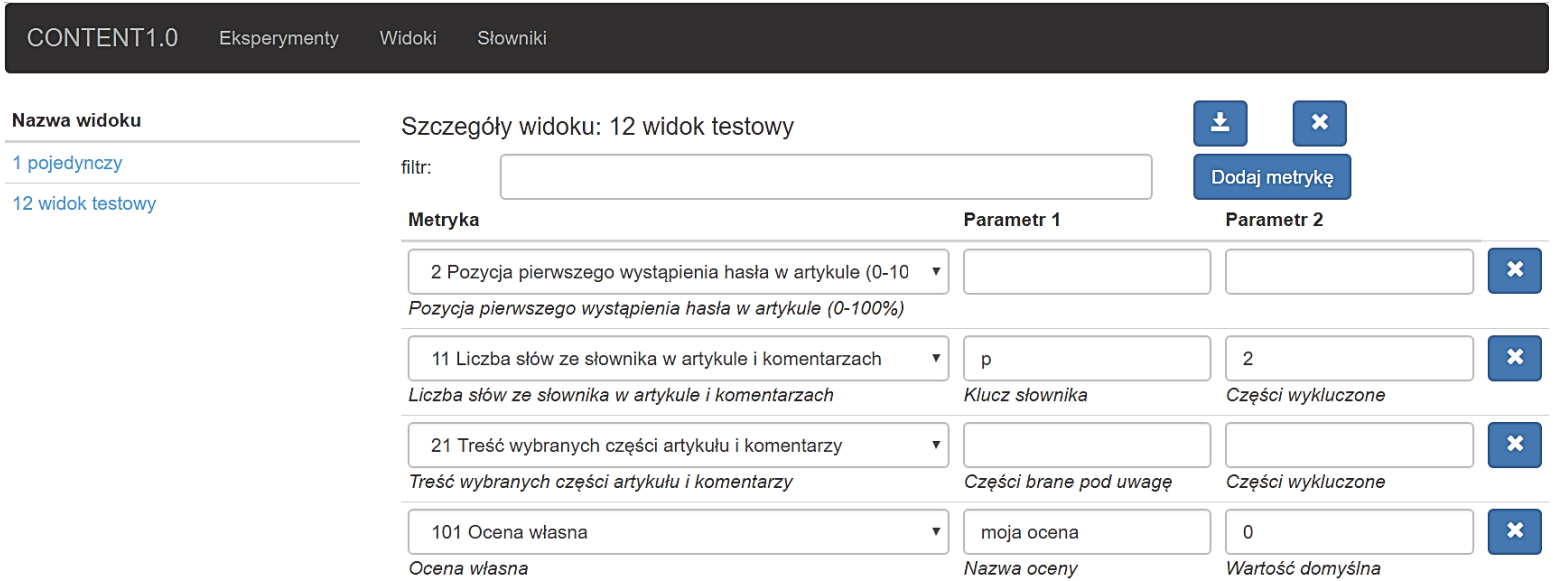
oszacowywanym operatorem jest koniunkcja (&), wówczas wynik zapytania może przyjmować wartość zero (fałsz, treść artykułu nie pasuje do kwerendy) lub jeden (prawda). Jeśli ostatnim oszacowywanym operatorem jest alternatywa, wartość kwerendy może być większa od jedności. W taki przypadku odpowiada on liczbie wszystkich wystąpień w artykule obu argumentów alternatywy.

Istotną innowacją w stosunku do standardowych wyrażeń regularnych jest umożliwienie wyspecyfikowania maksymalnej odległości w tekście pomiędzy wyszukiwanymi hasłami. Obsługiwanymi jednostkami odległości są słowo i zdanie. Jeśli kwerenda dotyczy tylko niektórych typów bloków artykułu, przyjmuje się roboczo, że pozostałe bloki nie istnieją, w związku z czym przeszukiwane bloki są traktowane tak, jakby następowały bezpośrednio po sobie.

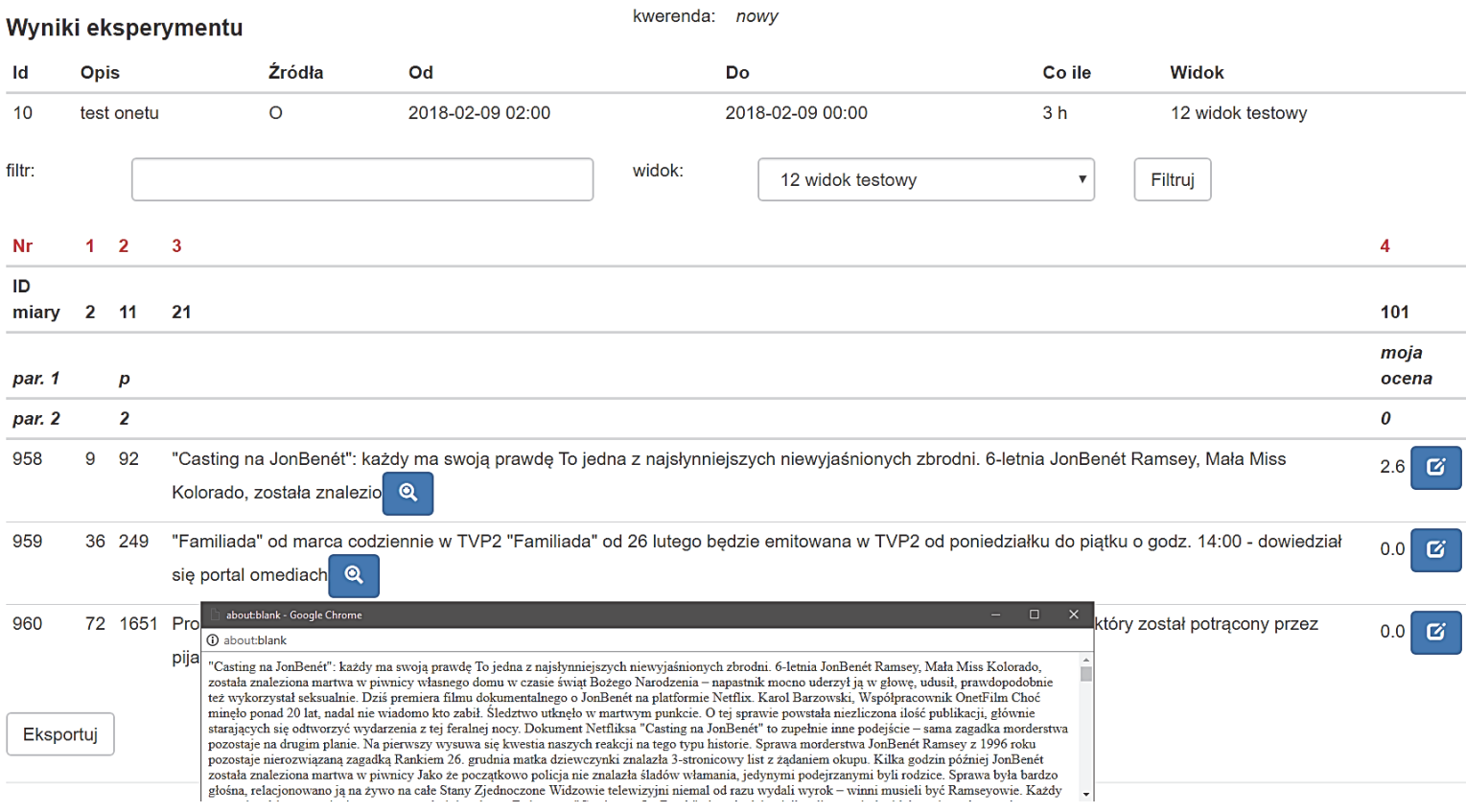
W systemie CONTENT 1.0 wprowadzono szereg predefiniowanych *metryk*, tj. algorytmów wyznaczających określone statystyki dla pojedynczego artykułu. Większość z nich może być parametryzowana przez użytkownika, jak to przedstawia tabela 1. Użytkownik może zdefiniować *widoki*, czyli zestawy metryk użyte do przedstawienia wyników eksperymentu. Dzięki uniwersalności metryk, można wykorzystywać te widoki wielokrotnie, w odniesieniu do różnych eksperymentów, traktując je jako swoistą perspektywę badawczą stanowiącą punkt wyjścia do dalszej, subiektywnej lub obiektywnej analizy szczegółowej wyników. Dzięki zaś parametryzacji można wykorzystywać większość metryk wielokrotnie, nawet w obrębie pojedynczego widoku, np. zestawiając liczbę znaków przestankowych w zasadniczym tekście artykułu oraz w komentarzach. Ekran definiowania widoku przestawiono na rys. 1; natomiast rys. 2 prezentuje wyniki zlecenia ukazane w tymże widoku. Zauważmy, że każdej definicji metryki odpowiada pojedyncza kolumna tabeli.

Tabela 1. Zestawienie metryk

Id. metryki	Wartość	Parametr 1	Parametr 2
1	Liczba wystąpień hasła w artykule i komentarzach	Typy bloków uwzględnionych	Typy bloków pominiętych
2	Pozycja względna pierwszego wystąpienia hasła z kwerendy (0–100%)	j.w.	j.w.
3	Liczba zdań w artykule i komentarzach	j.w.	j.w.
4	Średnia liczba znaków w zdaniu	j.w.	j.w.
9	Liczba znaków przestankowych	j.w.	j.w.
16	Liczba ilustracji	j.w.	j.w.
19	Liczba *hashtagów*	j.w.	j.w.
21	Treść wybranych bloków	j.w.	j.w.
23	Ilustracje	j.w.	j.w.
11	Liczba słów ze słownika	Identyfikator słownika	j.w.
5	Źródło artykułu		
6	Względna pozycja artykułu na portalu (0–100%)	Moment pomiaru (0–100% ogólnego czasu trwania zlecenia)	
10	Średnia liczba znaków przestankowych w komentarzu		
12	Średnia liczba emotikonów w komentarzu		
101	Ocena subiektywna	Identyfikator oceny	Wartość początkowa oceny



Rys. 1. Formularz wyboru metryk tworzących widok



Rys. 2. Ekran wyników analizy wg określonego widoku, wraz z okienkiem inspekcji pełnej treści artykułu

Aby umożliwić elastyczną ekspercką ocenę wyników, wprowadzono specjalny typ metryki (101) pozwalający użytkownikowi wprowadzać własne oceny poszczególnych artykułów. Ocena ma postać liczby z częścią ułamkową; takie ograniczenie umożliwia późniejsze, jednolite przetwarzanie ocen. Definicja typów ocen ma charakter opisowy; można wprowadzić dowolną liczbę typów ocen.

Kolejnym szczególnym typem metryki, powiązanych podobnie jak oceny z dodatkowym słownikiem danych, jest liczba słów należących do określonego, nazwanego zbioru. System CONTENT 1.0 wyposażono w zbiory słów polskich w formach podstawowych, mających wydźwięk pozytywny, negatywny, a także kojarzących się z emocjami podstawowymi (radość, zaufanie, cieszenie się na coś oczekiwanego, smutek, złość, strach, wstręt, zaskoczenie czymś nieprzewidywanym) oraz wartościami uniwersalnymi (użyteczność, dobro drugiego człowieka, prawda, wiedza, piękno, szczęście, nieużyteczność, krzywda, niewiedza, błąd, brzydota, nieszczęście). Zbiory te pochodzą ze Słowosieci+ emo, czyli polskiego odpowiednika słownika Wordnet[24].

Konsekwentna reprezentacja wyników zlecenia w postaci widoku w układzie tabelarycznym umożliwia eksport wstępnie przetworzonych danych do dalszej obróbki. Wyniki ujęte w konkretnym widoku można zapisać do pliku w formacie Microsoft Excel (.xls). Dla ilustracji powiązanych z artykułem (metryka typu 23) zapisywane są wyłącznie adresy URL, dla zapewnienia przenośności i redukcji rozmiaru pliku wynikowego.

[24] Słowosieć, TBC.

2.2. Architektura

Odpowiadając na współczesne potrzeby i trendy, a także perspektywy dalszego rozwoju, system CONTENT 1.0 został zaprojektowany z użyciem obecnie stosowanych, nowoczesnych technologii informatycznych. System składa się z szeregu powiązanych *mikrousług*, tj. wielu komponentów realizujących ściśle zdefiniowane, stosunkowo niewielkie fragmenty aplikacji. Możemy więc wyróżnić mikrousługę obróbki dokumentów, realizującą centralnie algorytm wykonywania kwerend, trzy mikrousługi skanujące odpowiednie źródła sieciowe oraz usługę koordynującą działanie wszystkich pozostałych i w szczególności odpowiedzialną za terminowe wykonywanie poszczególnych zleceń.

Graficzny interfejs użytkownika zaimplementowano w formie aplikacji sieciowej, w której formularze budowane są dynamicznie z wykorzystaniem biblioteki *Angular JS* po stronie przeglądarki. Kod aplikacji i definicje formularzy serwowane są przez statyczny serwer WWW; natomiast za kontrolę nad danymi do wyświetlenia odpowiada dedykowana mikrousługa. W ten sposób, realizując współczesne paradygmaty projektowania, rozdzielono logikę aplikacji, definicje wyglądu poszczególnych ekranów użytkownika, oraz manipulację właściwymi danymi. W szczególności odseparowano logikę aplikacji od bazy danych. Podobny zabieg wykonano po stronie usług skanowania.

Dekompozycja systemu na szereg możliwie bezstanowych usług oraz wprowadzenie warstwy abstrakcji dla przechowywania danych stanowią cenny kapitał – są bowiem bardzo dobrym punktem wyjścia do zadania skalowania wydajności systemu, niezbędnego w miarę wzrostu przetwarzanych danych.

Z tych samych powodów, system został od samego początku uruchomiony na maszynie wirtualnej dużego dostawcy usług hostingowych. Pozwala to mieć nadzieję na jego dalszy harmonijny wzrost, który wymagać będzie wdrożenia kolejnych rozwiązań właściwych dla systemów obsługi i analizy danych masowych (np. wdrożenia baz NoSQL i wprowadzenie kontenerowej architektury mikrousług).

3

EKSPERYMENT

3.1. Zbieranie danych

Eksperyment 1: szukamy artykułów zawierających gdziekolwiek słowa zaczynające się od „bezpieczeństw" oraz „cyfrow" (tj. bezpieczeństwo cyfrowe z uwzględnieniem końcówek fleksyjnych). Skanowano wszystkie źródła od 10 do 30 maja, powtarzając zbieranie danych co 6 godzin. Znaleziono łącznie zaledwie 35 artykułów, z czego trzy pochodzące z serwisu onet.pl, a pozostałe z następujących profili Facebooka popularnych witryn branżowych: Technowinki oraz niebezpiecznik.

Niewielka liczba wyników wynika z dynamicznych zmian w strukturze stron serwisu onet.pl, która spowodowała niedomagania w działaniu modułów skanujących ten serwis bezpośrednio, jak również jego bliźniaczy profil na Facebooku. Niestety, odświeżenie układu stron i wprowadzanie nowych funkcjonalności przez dostawców treści powodują najczęściej konieczność natychmiastowego dostosowania do nich programów skanujących. Dlatego pozyskiwanie danych poprzez web scraping jest uznawane za bardzo kosztowne w utrzymaniu w porównaniu z korzystania z API, i stosuje się je w ostateczności.

Eksperyment 2: szukamy artykułów zawierających słowo NASK (wielkość liter bez znaczenia). Skanowano te same źródła co powyżej, od 21 marca do 30 kwietnia, co trzy godziny.

3.2. Analiza statystyczna

Do projektu wybrano przetwarzanie wsadowe, które wymaga skompletowania pełnego/zamkniętego zbioru danych wejściowych. Każdy rekord musi być zapisany w postaci ilościowej (lub zrekodowanej do takiej formy).

Podstawą analizy ilościowej są miary tendencji centralnej oraz miary rozproszenia (w zależności od skali pomiarowej).

Przetwarzanie danych zostanie przeprowadzone metodą funkcji podobieństwa (metodą liniową). Transformacja danych uzupełniona zostanie ekstrakcją wstępną, czyli sprowadzeniem zbioru danych do możliwe optymalnego podzbioru cech, które dają jak największe możliwości eksploracyjne. Transformacja i ekstrakcja wstępna zostanie przeprowadzona za pomocą statystycznej analizy skupień (przy wykorzystaniu SPSS).

Analiza skupień to zbiór metod wielowymiarowej analizy statystycznej, służących wyodrębnianiu jednorodnych podzbiorów obiektów badanej populacji obiektów. Metody analizy skupień są stosowane wówczas, gdy nie dysponujemy hipotezami a priori, a badania są w fazie eksploracyjnej. Dzięki analizie skupień można wykryć, czy otrzymane skupienia wskazują na jakąś prawidłowość, dokonać redukcji dużego zbioru danych do średnich poszczególnych grup,

Tabela 2. Statystyki – miary tendencji centralnej i miary rozproszenia.

		Liczba wystąpień hasła	Liczba zdań w artykule	Średnia liczba znaków w artykule i komentarzach	Liczba znaków przestankowych w artykule	Liczba znaków przestankowych w komentarzach
N	Ważne	325	325	323	325	17
	Braki	0	0	2	0	308
Średnia		2,1969	10,6338	60,38427	19,8615	3,6341
Mediana		2,0000	5,0000	57,40000	6,0000	4,0909
Dominanta		2,00	3,00	57,400	6,00	1,00
Odchylenie standardowe		1,25388	23,97603	25,927020	62,20612	1,37328
Wariancja		1,572	574,850	672,210	3869,601	1,886

potraktować rozdzielenie na grupy jako wstęp do dalszych wielowymiarowych analiz[25].

Statystyczna analiza skupień będzie zatem dla naszego zbioru surowego (ilościowego) algorytmem selekcji, gdzie filtrem wbudowanym do wyboru podzbiorów cech będzie podobieństwo/niepodobieństwo obiektów a kryterium stopu: kompletność przeszukania, specyficzna granica ilości iteracji lub ilości cech, brak przyrostu nowych związanych obiektów w klastrze, określony błąd pomiaru.

Istnieją dwa sposoby aglomeracji danych: metody hierarchiczne oraz grupowanie metodą *k*-średnich. W projekcie zostanie zastosowana metoda hierarchiczna, która jest nieparametryczna, niewrażliwa na występowanie szumu i braków danych oraz nie wymaga apriorycznej konieczności ustalenia dokładnej, zamkniętej struktury zbioru zmiennych[26]. Dodatkowo, zaletą wykorzystania hierarchicznych metody aglomeracyjnej jest zastosowanie jednej, centralnej procedury aglomeracyjnej, podczas której proces grupowania można śledzić a wyniki kontrolować.

Do realizacji metody hierarchicznej najczęściej wykorzystywane są techniki aglomeracyjne, w których początkowo każdy obiekt stanowi osobne skupienie, następnie obiekty leżące najbliżej siebie są łączone w nowe skupienie aż do uzyskania jednego skupienia. Problemem jest określenie odległości (czyli zasady wiązania) między nowymi skupieniami, powstającymi z połączonych obiektów. Istnieje szereg różnych zasad wiązania, które między sobą różnią się jedynie sposobami

[25] Brian S. Everitt, Sabine Landau, Morven Leese, Daniel Stahl, *Cluster analysis*, 5th edition, John Wiley & Sons, Chichester 2011.

[26] Kamila Migdał-Najman, Krzysztof Najman, *Samouczące się sztuczne sieci neuronowe w grupowaniu i klasyfikacji danych. Teoria i zastosowania w ekonomii*, Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk 2013.

obliczania odległości między skupieniami (*single linkage metod*, *complete linkage*, UPGMA – *unweighted pair-group metod using arithmetic averages*, WPGMA – *weighted pair-group metod using arithmetic averages*, UPGMC – *unweighted pair-group metod using the centroid average*, *weighted pair-group metod using the centroid average*, *Ward's method*). Do projektu została wybrana metoda Warda[27]. Ta metoda różni się od wszystkich pozostałych, ponieważ do oszacowania odległości między skupieniami wykorzystuje podejście analizy wariancji – zmierza do minimalizacji sumy kwadratów odchyleń dowolnych dwóch skupień, które mogą zostać uformowane na każdym etapie. Metoda ta zmierza do minimalizacji sumy kwadratów odchyleń wewnątrz skupień. Miarą zróżnicowania skupienia względem wartości średnich jest ESS (*Error Sum of Squares*), zwane również błędem sumy kwadratów. ESS jest określone wzorem:

$$ESS = \sum_{i=1}^{k} (x_i - \overline{x})^2$$

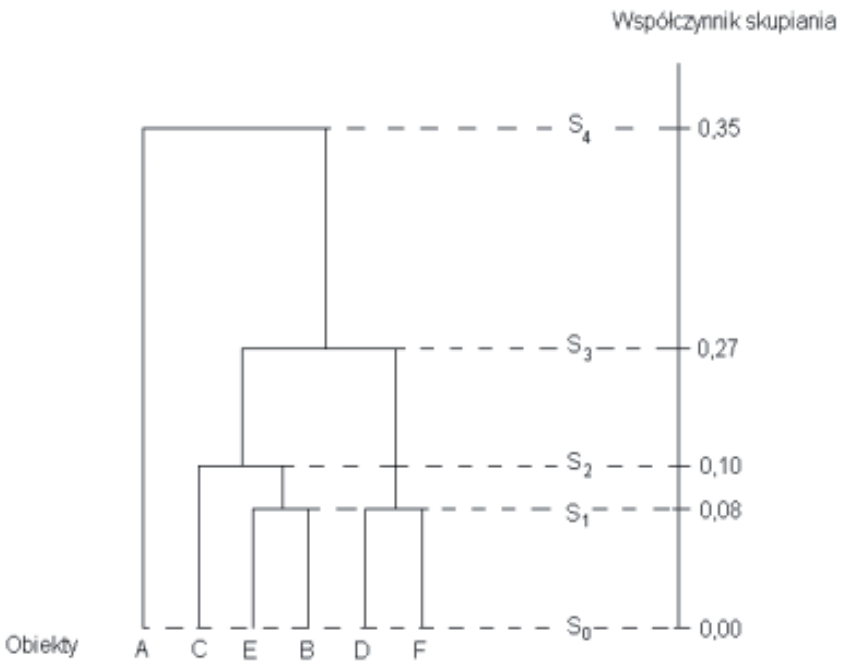
x_i – wartość zmiennej będącej kryterium segmentacji dla i-tego obiektu,
k – liczba obiektów w skupieniu.

Wybór tej metody wynika z jej ponadprzeciętnej efektywności, tzn. tworzy skupienia statystycznie homogeniczne.

Algorytmy aglomeracyjne są uniwersalne, mogą być stosowane dla danych ilościowych i jakościowych (zrekodowanych do postaci numerycznych). Ponadto charakteryzuje je szybkość działania. Niemniej jednak, do ustalenia końcowej liczby skupień konieczna jest analiza dendrogramu, który jest ilustracją graficzną procesu

[27] Joe H. Ward, *Hierarchical Grouping in Optimize an Objective Function*, „Journal of the American Statistical Association" 1963, vol. 58.

łączenia obiektów. Procedura łączenia zostaje wstrzymana po przekroczeniu ustalonej, progowej wartości miary odmienności pomiędzy skupieniami.



Rys. 3. Dendrogram – wynik agregacji metodą Warda

W zależności od przyjętych założeń badania, w tym zwłaszcza akceptowanej odległości taksonomicznej między obiektami ze względu na zaproponowany zestaw cech, możemy wyróżniać większe lub mniejsze skupienia, a co za tym idzie – mniejszą lub większą ich liczbę.

Celem obserwacji kolejności połączeń grup z miernikami charakteryzującymi odległość oraz wykluczenia wiązań pozornych (wynikłych np. z powodu wystąpienia outlayersów), interpretacja i określenie granic zbioru cech zostaną przeprowadzone (dla każdej operacji agregowania) przez badacza.

Dodatkowo, na podstawie hierarchicznej analizy skupień, zostaną skonstruowane numeryczne, zagregowane zmienne czynnikowe, pozwalające na dalszą analizę data mining i końcową interpretacje wyników pomiaru. Wybrany statystyczny algorytm przetwarzania danych jest optymalny, gdyż zapewnia reprezentację dużych ilości danych, a także agreguje te dane, przez co przyspiesza proces przeszukiwania, przetwarzania, klasyfikacji, oraz dyskryminacji wzorców.

Dobór próby do analizy jakościowej.
W sytuacji pomiarów, gdzie wystąpi duża liczba rekordów, zostanie zastosowanydobór systematyczny losowania próby do analizy jakościowej. Dobór systematyczny polega na wyborze z uporządkowanego zbioru odpowiedniej liczby jednostek w równych odstępach (interwałach). Najpierw ustala się liczebność (N) całej zbiorowości, a następnie liczebność (n) próby i na tej podstawie ustala się interwał losowania $k = N/n$. Poczynając następnie od losowo obranej jednostki pierwszego interwału dobiera się kolejno co k jednostek z każdego interwału po jednej jednostce, aż osiągnie się pożądaną wielkość próby losowej.

Wielkość próby dla takiego losowania będzie liczona ze wzoru:

$$n_b = \frac{N}{1 + \frac{d^2(N-1)}{z_\alpha^2 pq}}$$

N – liczność populacji;
p – spodziewany rząd wielkości szacowanej frakcji;
q – $1 - p$;
z_α – 1,64 dla $\alpha = 0{,}10$;
1,96 dla $\alpha = 0{,}05$;
2,58 dla $\alpha = 0{,}01$;
d – dopuszczalny błąd szacunku frakcji p.

3.3. Analiza jakościowa

Projekt miał na celu stworzenie aplikacji umożliwiającej gromadzenie danych oraz realizację analizy jakościowej zgodnie z założeniami metodologicznymi teorii ugruntowanej opracowanej przez Glasera i Straussa[28].

Filarami teorii ugruntowanej są trzy zasady:

- Badania należy rozpoczynać bez przyjmowania wstępnej hipotezy, dzięki temu unikamy sytuacji, w której istniejące teorie wpłyną na spostrzeganie badanego zjawiska.
- Druga zasada polega na nieustannym porównywaniu ze sobą zebranych fragmentów materiału empirycznego. To porównanie prowadzi do określenia kodów służących do porządkowania i zinterpretowania materiału w celu wyróżnienia najważniejszych kategorii, z których zostanie zbudowana teoria dotycząca badanego zjawiska.
- Trzecia zasada to teoretyczne pobieranie próbek. Polega na tym, że materiał do badania wybieramy w taki sposób, by poszerzyć naszą znajomość problemu, a nie by uzyskać jedynie próbkę reprezentatywną.

Teoria ugruntowana wymaga od badacza przestrzegania wyznaczonych reguł postępowania. Zgodnie z założeniami metody należy podchodzić do badanego przedmiotu w sposób otwarty, bez przywiązywania większej wagi do tworzenia hipotez już w początkowym stadium badania. Jednak oczywiste jest, że każdy badacz wnosi do procesu badawczego swój sposób myślenia, przekonania i założenia,

[28] Barney Glaser i Anselm L. Strauss, *Odkrywanie teorii ugruntowanej. Strategie badania jakościowego*, Zakład Wydawniczy Nomos, Kraków 2009.

które nabył w trakcie życia[29]. Ważne jest, żeby badacz miał świadomość, w jakim stopniu jego sposób interpretacji wynika z badanej rzeczywistości, a w jakim z jego uprzedzeń, przekonań i preferencji.

Proces badania zgodnie z założeniami teorii ugruntowanej składa się z trzech rodzajów działań:

- zbierania danych;
- kodowania i identyfikowania idei lub koncepcji;
- generowania teorii.

Przy zbieraniu danych w badaniach prowadzonych zgodnie z zaleceniami teorii ugruntowanej należy kierować się zasadą teoretycznego pobierania próbek. Dane należy zbierać tak długo, aż osiągniemy stan nasycenia teoretycznego, co oznacza, że dalsze zbieranie danych nie wzbogaci już wiedzy o badanym zjawisku i nie pomoże w dalszym rozwijaniu tworzonej przez badacza teorii.

Oprogramowanie zostało tak zaprojektowane, żeby w wymaganym stopniu umożliwiać tworzenie właściwego zbioru danych. Dane składają się ze zbioru artykułów pozyskiwanych według zadanego przez badacza zapytania zbudowanego z interesującego go hasła bądź kilku haseł. Progi nasycenia teoretycznego mogą być ustalone na dwa sposoby: pierwszy to liczba artykułów w zbiorze, drugi to czas zbierania artykułów. Czas zbierania artykułów jest szczególnie istotny przy badaniu dynamiki zjawisk, zwłaszcza tych pojawiających się nagle i szybko przemijających.

[29] Constance T. Fischer, *Bracketing in qualitative research: Conceptual and practical matters*, „Psychotherapy Research" 2009, 19(4–5), s. 583–590. doi:10.1080/10503300902798375.

Zwiększenie wiarygodności badań zapewnia triangulacja danych, która w praktyce realizowana jest poprzez sięganie po dane z różnych źródeł. Stworzona aplikacja umożliwia w każdym uruchomionym eksperymencie pobieranie danych z wielu źródeł, np. portali twitter.com, facebook.com oraz onet.pl. Ponadto dane te mogą być pobieranie w różnym, określonym przez badacza czasie.

Zgromadzone w ten sposób dane powinny być poddane kodowaniu. Kodowanie to jeden z najważniejszych etapów projektu badawczego prowadzonego zgodnie z zaleceniami teorii ugruntowanej. W tej fazie badania przechodzimy od danych do kategorii abstrakcyjnych, z których w końcowym etapie powstanie teoria średniego zasięgu.

Badacze stosują różne strategie kodowania materiału empirycznego: słowo po słowie, wiersz po wierszu, zdarzenie po zdarzeniu[30]. Wszystkie trzy strategie mają na celu dostrzeżenie nowych zjawisk w dobrze znanym na pozór materiale[31]. Kodowanie słowo po słowie pozwala skoncentrować uwagę na niuansach. Kodowanie wiersz po wierszu narzuca spojrzenie na kodowany tekst przez pryzmat podziału na wiersze. Najbardziej zbliżoną do naturalnego sposobu spostrzegania narracji wydaje się być analiza i kodowanie zdarzenie po zdarzeniu. Jednakże wybór strategii kodowania jest uzależniony od wielu czynników, między innymi od długości analizowanego tekstu.

Zaprojektowane oprogramowanie posiada możliwość quasi kodowania, które może być przeprowadzone przezzastosowanie specjalnego typu metryki pozwalający użytkownikowi wprowadzać własne oceny

[30] Kathy Charmaz, *Teoria Ugruntowana. Praktyczny przewodnik po analizie jakościowej*, WN PWN, Warszawa 2009.

[31] Judith A. Holton, *The Coding Process and Its Challenges*, „The Grounded Theory Review" 2010, vol. 9, nr 1, s. 21–38.

poszczególnych artykułów. Ocenama postać liczby z częścią ułamkową; takie ograniczenie umożliwia późniejsze, jednolite przetwarzanie ocen. Definicja typów ocen ma charakter opisowy; użytkownik może wprowadzić dowolną liczbę typów ocen.

Kodowanie i dalsze etapy badania mogą być realizowane poprzez wykorzystanie specjalistycznego oprogramowania zewnętrznego, takiego jak MAXQDA, Nvivo lub Atlas. Programy te nie tylko ułatwiają kodowanie, ale także oferują graficzną wizualizację struktury badanego materiału. Stworzone oprogramowanie nie będzie stwarzało badaczowi ograniczeń w korzystaniu z zewnętrznych programów do analizy pogłębionej, dzięki eksportowi danych do pliku w popularnym formacie xlsx.

W kolejnym etapie zakodowane opisy powinny być grupowane w kategorie, co ułatwia porzucenie myślenia o konkretnych zdarzeniach na rzecz analizy w kategoriach na wyższym poziomie abstrakcji. Analiza kategorii może doprowadzić do tworzenia teorii odnośnie do badanego zjawiska.

4

WYNIKI

4.1. Analiza statystyczna danych

Eksperyment „NASK"

Małe zróżnicowanie źródeł obserwacji, które wynikało z dynamicznych zmian w strukturze stron serwisu Onet.pl, ma swój wyraz w otrzymanej strukturze rekordów. Dominują obserwacje z Twittera (89,5%), rekordy z Facebooka stanowią 8,3%, a z Onet.pl jedynie 2,2% (patrz tabela 3).

Tabela 3. Rozkład procentowy i częstości źródeł rekordów w eksperymencie „NASK"

	Częstość	Procent
Twitter	291	89,5
Facebook	27	8,3
Onet	7	2,2
Ogółem	325	100,0

Miary tendencji centralnej i miary rozproszenia

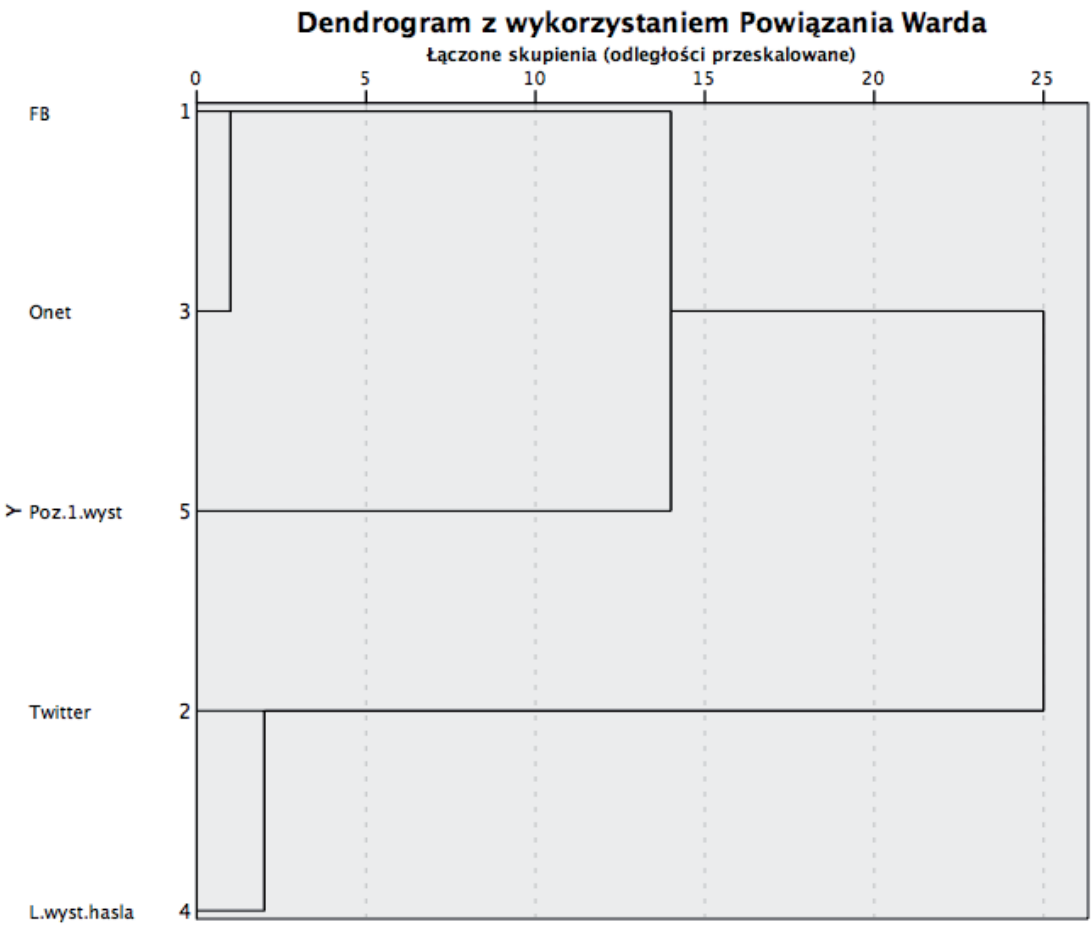
Średnio liczba wystąpień hasła „NASK" w artykule wyniosła 2 razy, średnia (mediana) pozycja pierwszego wystąpienia hasła to 19,0% w stosunku do pierwszego słowa w artykule, średnia (mediana) liczba

zdań w artykule i komentarzach – 5, średnia (mediana) liczba znaków w artykule i w komentarzach – ok. 58, a znaków przestankowych – 6 (w tym w komentarzach – 4) oraz hashtagów – 0, natomiast średnia (mediana) liczba emotikonów w komentarzach – 0,26, średnia liczba słów pozytywnych i negatywnych lub z kategorii: błąd, zaufanie, użyteczność, nieużyteczność, wiedza, niewiedza – 0 (patrz tabela 4).

Transformacja i ekstrakcja – analiza skupień

Analiza skupień pozwala na eksploracje danych i poszukiwanie zależności całych grup zmiennych. Przykładowo odnotowano następujące korelacje:

- Częściej hasło „NASK" występowało na Twitterze niż Facebooku czy Onet.pl, natomiast na FB i Onet.pl, „NASK" jest częściej pozycjonowane w tytule i lub na początku wpisu/artykułu (patrz rys. 4).
- Emocjonalny charakter wypowiedzi (negatywnej lub pozytywnej) mocniej jest związana z liczbą wystąpień samego hasła „NASK" niż z jego pozycją w artykule (patrz rys. 5).
- Emocjonalnanaturatreści (negatywna lub pozytywna) mocniej jest także związana z liczbą zdań w artykule, im większa liczba zdań w artykule, tym częściej występowały narracje uczuciowe (patrz rys. 6).
- Treści o charakterze pozytywnym lub negatywnym zdecydowanie częściej występują na Facebook i Onecie niż na Twitterze (patrz rys. 7).
- Artukuły/wpisy na Facebooku i Onet.pl zdecydowanie częściej zawierają słowa o wydźwięku – 'błąd', niewiedza', 'nieużyteczność', natomiast w mniejszym stopniu słowa o wydźwięku – 'zaufanie', 'wiedza'i 'użyteczność'. Wpisy na Twitterze nie korelują z żadnymi zbiorami słów o wspomnianym wcześniej wydźwięku (patrz rys. 8).

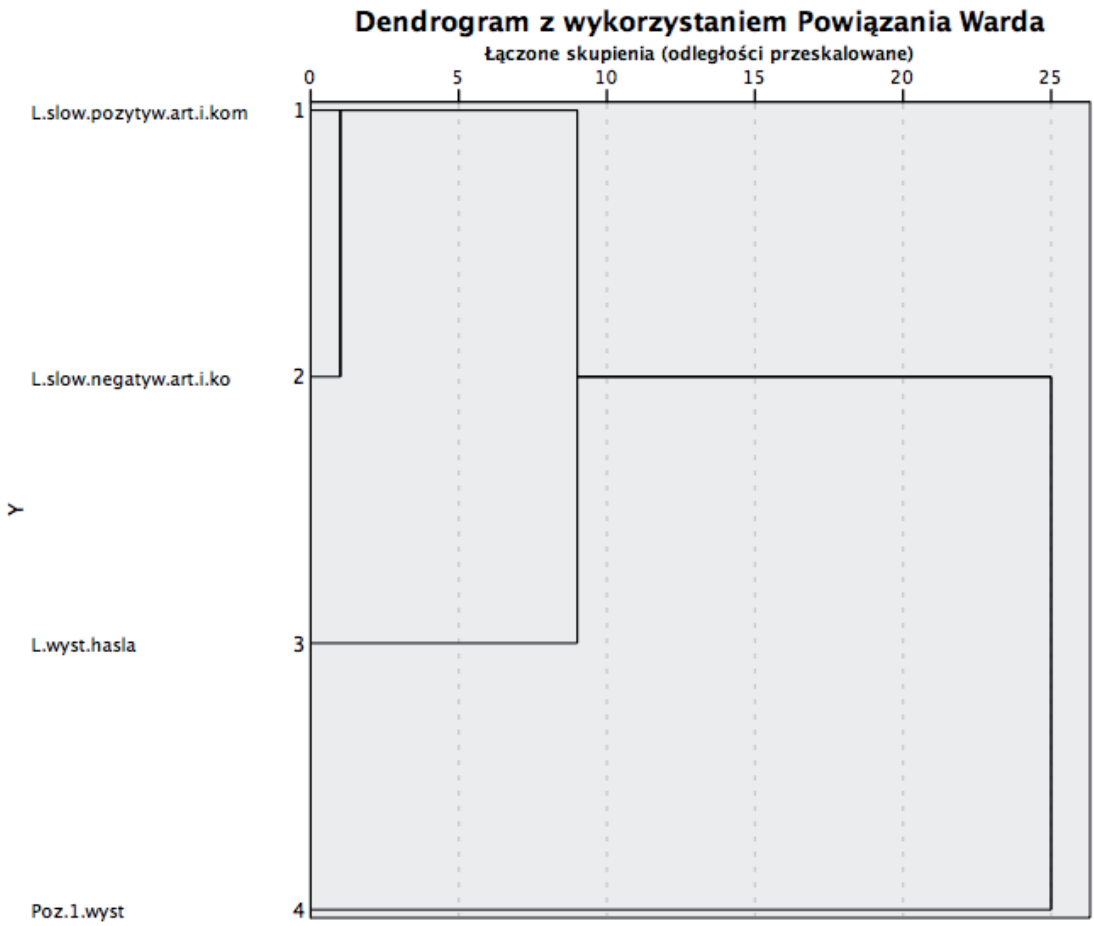


Rys. 4. Dendrogram – wynik agregacji zmiennych „źródło", „pozycja hasła w artykule", „liczba wystąpień hasła"

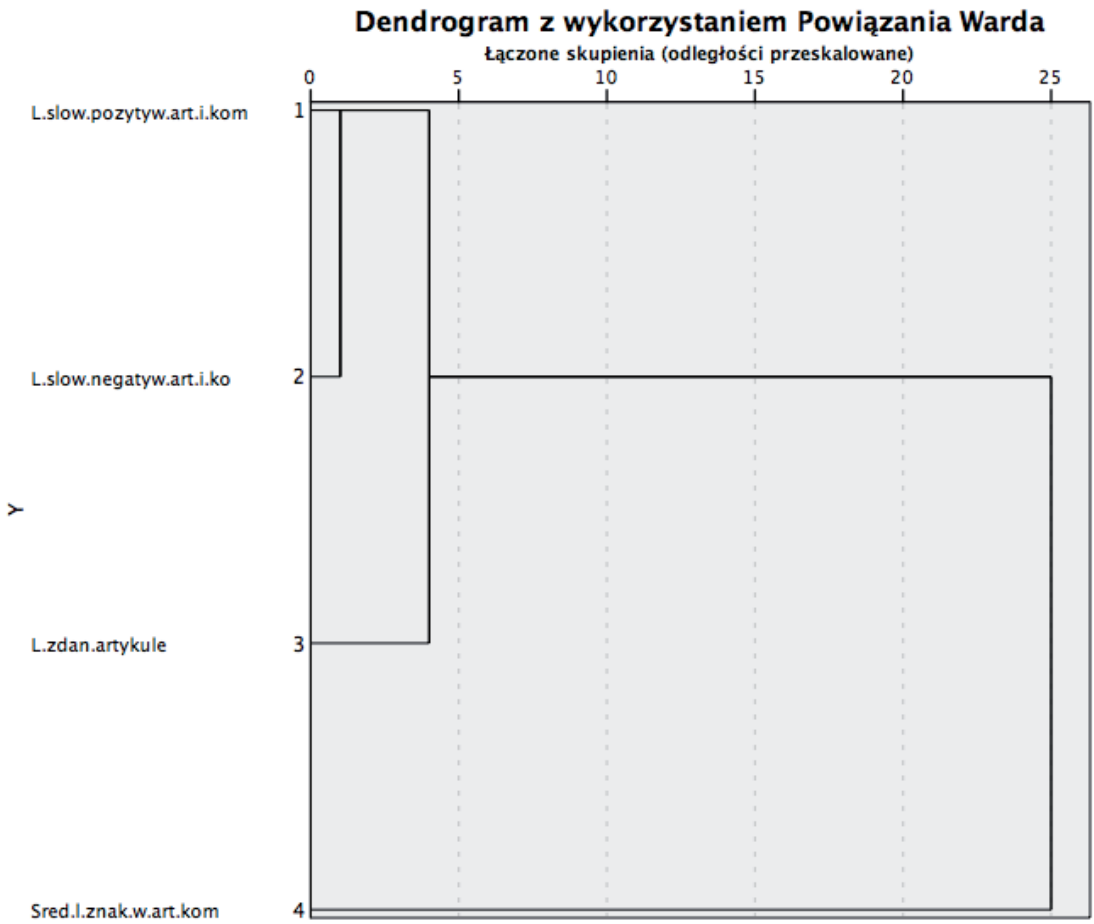
Tabela 4. Statystyki – miary tendencji centralnej i miary rozproszenia – eksperyment „NASK"

		L. wyst. hasła	Poz.1.wyst	L.zdan.artykule.i.kom.	Sred.l.znak.w.art.kom	Znaki.przest	S.l.znaków.kom	S.l.emotikon.kom	L.hash.w.art.i.kom	L.slow.pozytyw. art.i.kom
N	325	325	325	323	325	17	24	325	325	325
	bd	0	0	2	0	308	301	0	0	0
M		2,197	23,499	10,6338	60,384	19,862	3,634	0,266	1,132	2,846
Me		2,000	19,000	5,000	57,400	6,000	4,091	0,261	0,000	0,000
D		2,00	2,00	3,00	57,400	6,00	1,00	0,00	0,00	0,00
Σ		1,254	20,685	23,976	25,927	62,206	1,373	0,178	1,781	13,834
Var		1,572	427,850	574,850	672,210	3869,601	1,886	0,032	3,171	191,377

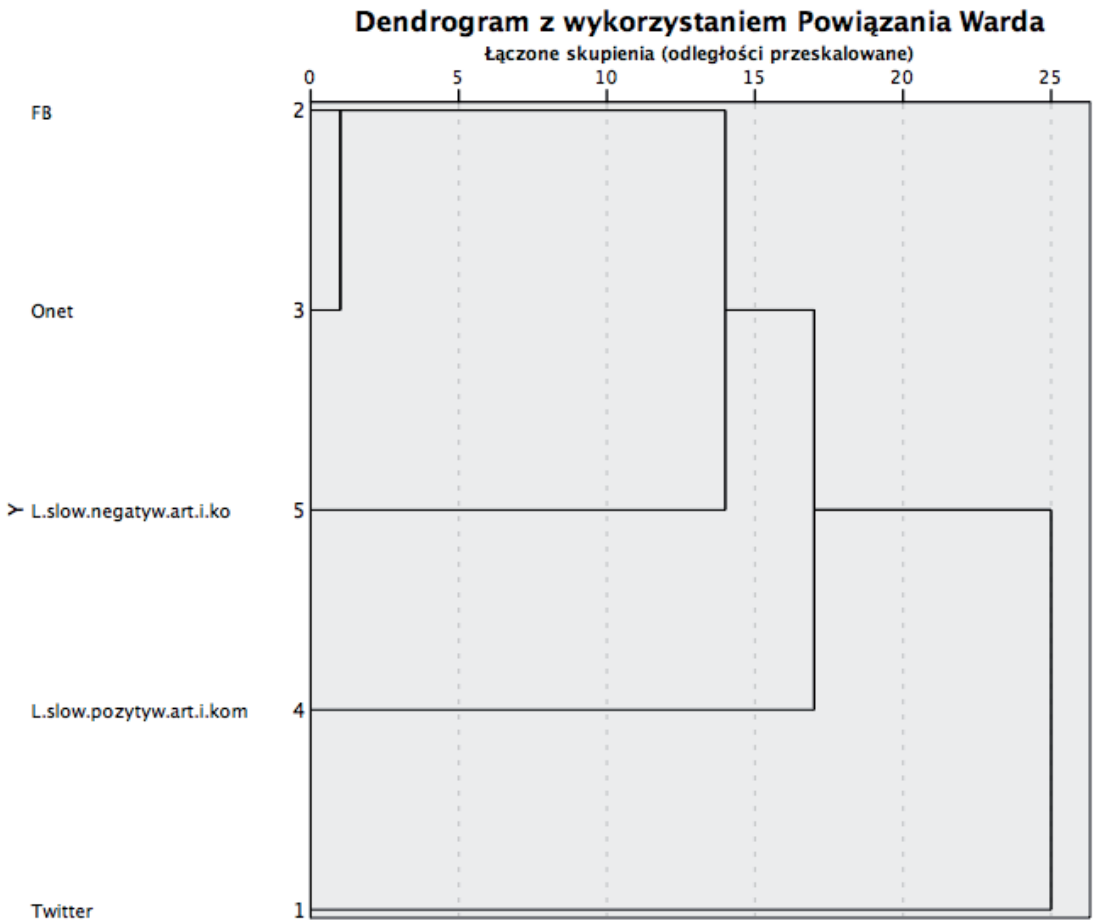
		L.slow.negatyw.art.i.ko	L.slow.błąd.art.i.kom	L.slow.zaufanie.art.i.kom	L.slow.uzytecznosc.art.i.kom	L.slow.nieuzyteczn0sc.ar.kom	L.slow.wedza.art.i.kom	L.slow.niewiedza.art.i.kom
N	N	325	325	325	325	325	325	325
	bd.	0	0	0		0	0	0
M		4,0400	1,4308	1,9477	2,5908	1,3385	0,7692	0,4985
Me		0,00	0,00	0,00	0,00	0,00	0,00	0,00
D		0,00	0,00	0,00	0,00	0,00	0,00	0,00
Σ		15,51647	6,51108	7,54822	8,98019	6,60312	3,72417	2,38129
Var		240,761	42,394	56,976	80,644	43,601	13,869	5,671



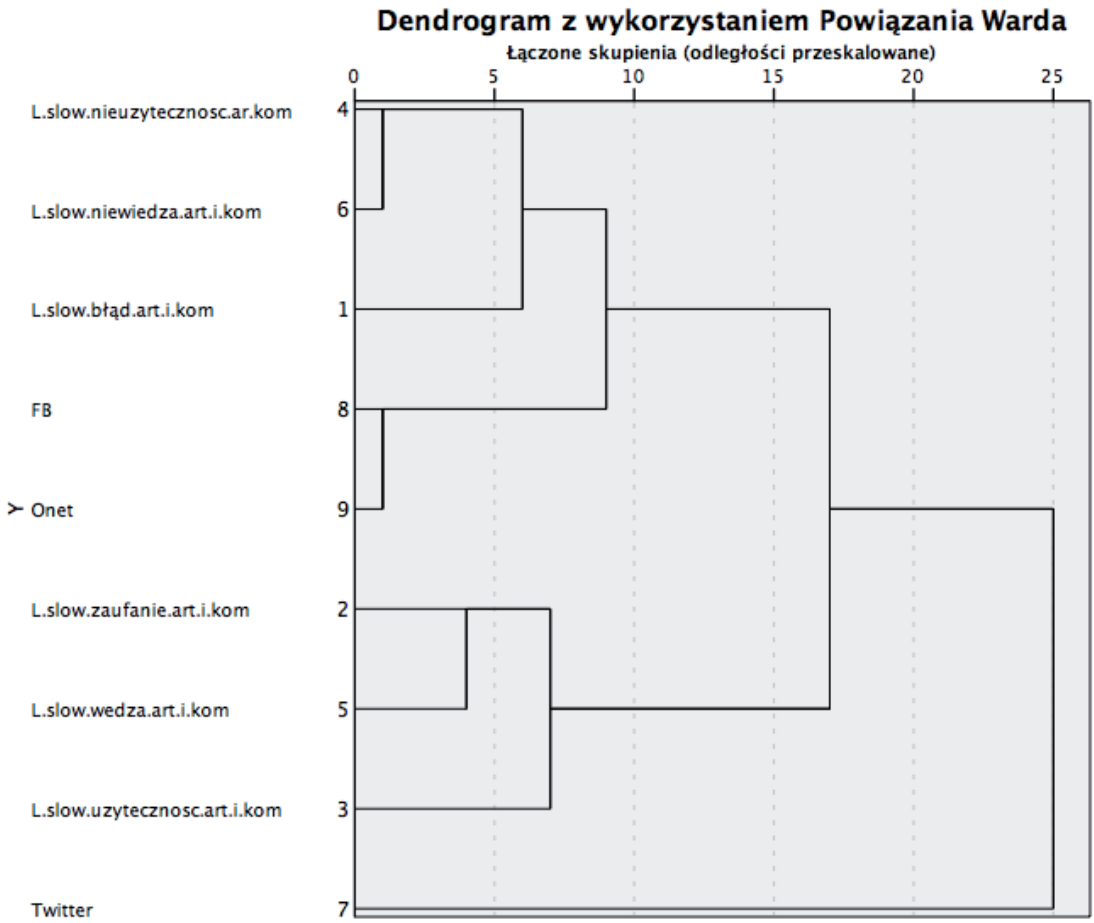
Rys. 5. Dendrogram – wynik agregacji zmiennych „liczba wystąpień słów pozytywnych", „liczba wystąpień słów pozytywnych", „liczba wystąpień hasła"



Rys. 6. Dendrogram – wynik agregacji zmiennych „liczba wystąpień słów pozytywnych", „liczba wystąpień słów pozytywnych", „liczba zdań w artykule", „średnia liczba znaków w artykule"



Rys. 7. Dendrogram – wynik agregacji zmiennych „źródło", „liczba wystąpień słów pozytywnych", „liczba wystąpień słów negatywnych"



Rys. 8. Dendrogram – wynik agregacji zmiennych „źródło", „liczba wystąpień słów 'błąd'", „liczba wystąpień słów 'zaufanie'", „liczba wystąpień słów 'wiedza", „liczba wystąpień słów 'użyteczność'", „liczba wystąpień słów 'niewiedza", „liczba wystąpień ", „liczba wystąpień słów 'nieużyteczność'"

Pokazana tu rafinacja nie ma charakteru reprezentatywności, ponieważ eksperyment został przeprowadzony w momentach czasowych wybranych przypadkowo, a samo zbieranie danych zostało obarczone błędem zmian w strukturze stron serwisu Onet.pl. Jednakże, powyższe analizy są przykładami obrazującymi możliwości prototypu aplikacji.

Eksperyment „bezpieczeństwo cyfrowe"

W eksperymencie „Bezpieczeństwo cyfrowe", małe zróżnicowanie próby badawczej jest jeszcze większe niż w eksperymencie „NASK", przyczyny tego zostały już wyjaśnione wcześniej. Dominują tutaj obserwacje z Facebooka (91,2%), rekordy z Onet.pl stanowią jedynie 8,9%, a z Twittera jest ich w próbie brak (patrz tabela 5).

Tabela 5. Rozkład procentowy i częstości źródeł rekordów w eksperymencie „Bezpieczeństwo cyfrowe"

	Częstość	Procent
Twitter	0	0,0
Facebook	31	91,2
Onet	3	8,9
Ogółem	34	100,0

Miary tendencji centralnej i miary rozproszenia

Średnio (mediana) liczba wystąpień hasła „Bezpieczeństwo cyfrowe" w artykule wyniosła 1, średnia (mediana) pozycja pierwszego wystąpienia hasła to 34,0% w stosunku do pierwszego słowa w artykule, średnia (mediana) liczba zdań w artykule i komentarzach – 88, średnia (mediana) liczba znaków w artykule i w komentarzach – 271, a znaków przestankowych – 4, hashtagów – 0, natomiast średnia

(mediana) liczba emotikonów w komentarzach – 0,30, średnia liczba słów pozytywnych – 35 i negatywnych – 45, a z kategorii: błąd – 21, zaufanie – 22, użyteczność – 27, nieużyteczność – 20, wiedza – 10, niewiedza – 9 (patrz tabela 6).

Transformacja i ekstrakcja – analiza skupień

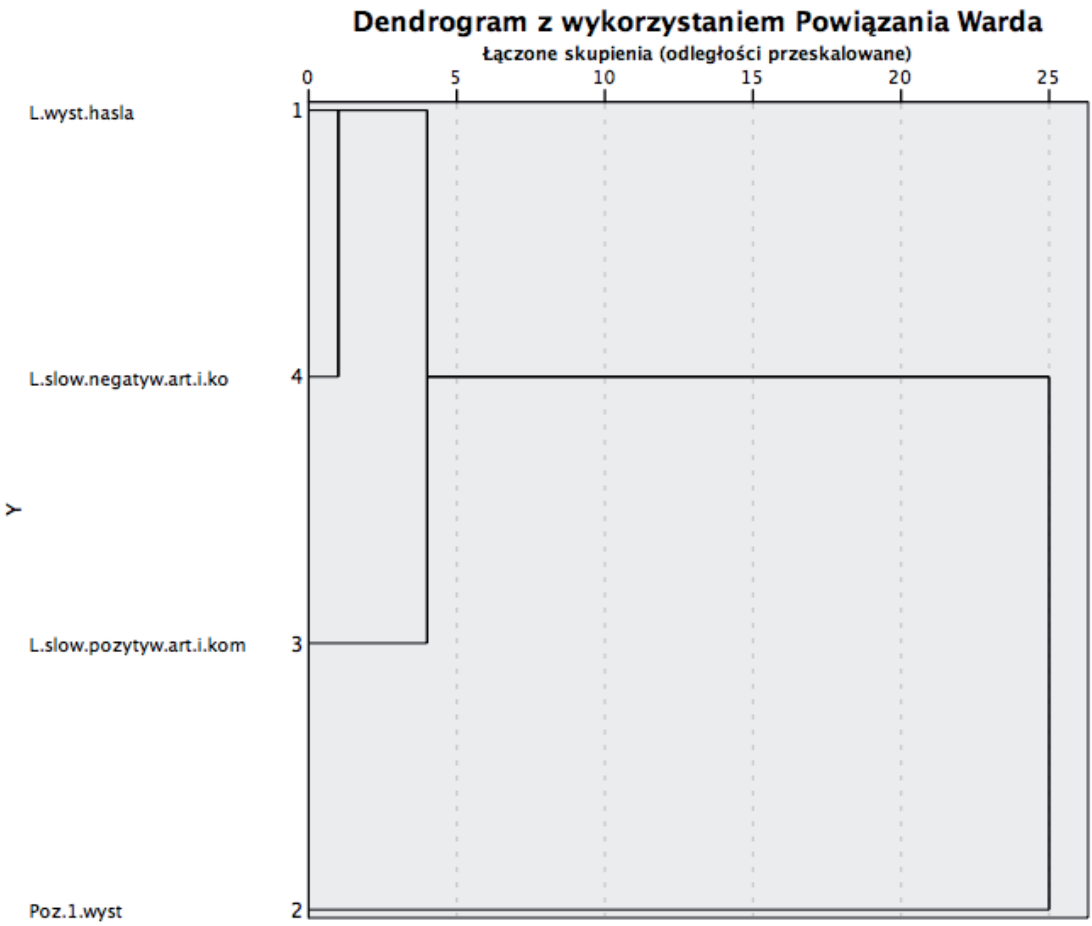
Również tutaj, celom poglądowym została wykonana analiza skupień, celem wstępnej ekstrakcji zmiennych o najwyższym ładunku predykcji. Przykładowo odnotowano następujące korelacje:

- Zarówno negatywny, jak i pozytywny wydźwięk koreluje pozytywnie z liczbą wystąpień hasła. W przypadku pozycji pierwszego hasła w artykule brak jest zależności z wydźwiękiem (patrz rys. 9).
- Pozytywny wydźwięk artykułu/wpisu koreluje pozytywnie z liczbą zdań i znaków w artykule/wpisie (patrz rys. 10).
- Średnia liczba emotikonów i znaków przestankowych nie ma wpływu na wydźwięk artykułu/wpisu (patrz rys. 11).
- Wydźwięk pozytywny, dodatnio koreluje z słowami z kategorii: „błąd", „nieużyteczność", „niewiedza", natomiast negatywny ze słowami z kategorii: „zaufanie", „wiedza", „użyteczność" (patrz rys. 12).

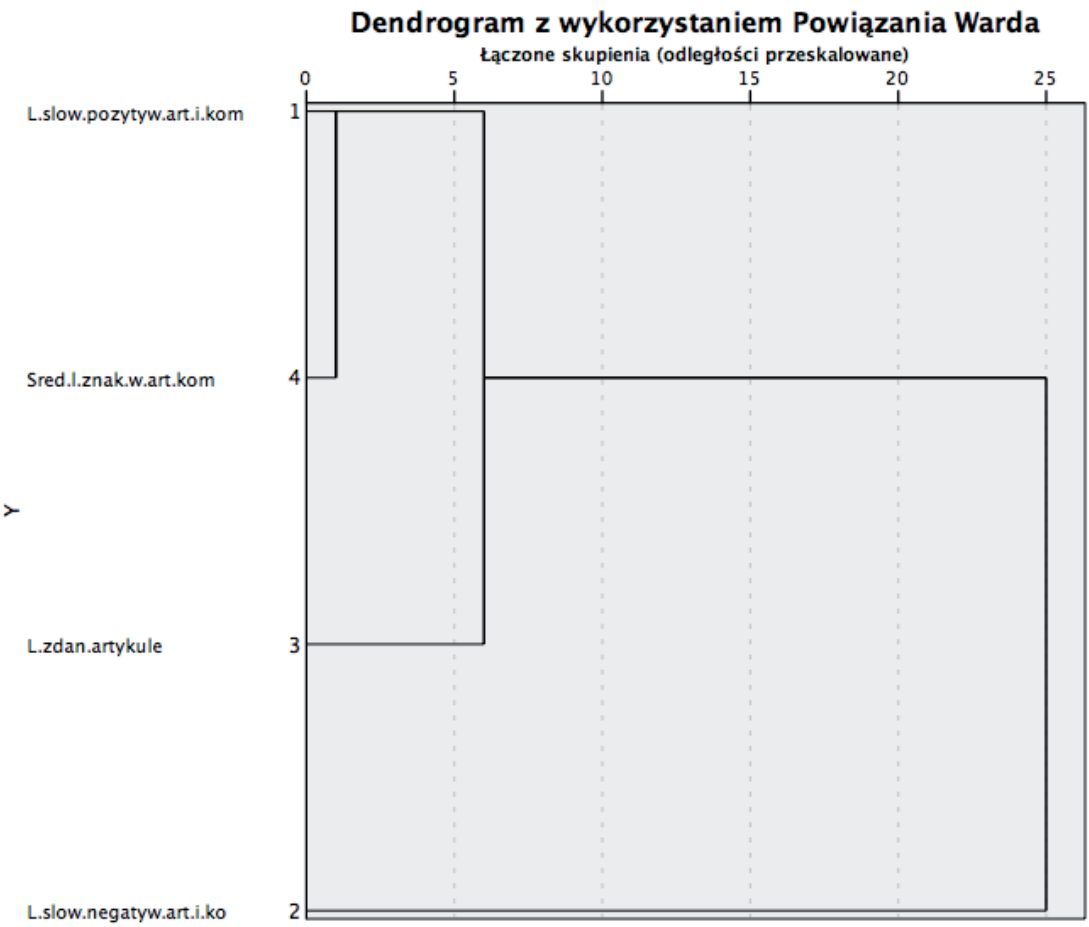
Tabela 6. Statystyki – miary tendencji centralnej i miary rozproszenia – eksperyment „Bezpieczeństwo cyfrowe"

		L.wyst.hasla	Poz.1.wyst	L.zdan.artykule.i.kom.	Sred.l.znak.w.art.kom	Znaki.przest	Sred.l.emot.w.kom	L.hash.w.art.i.kom	L.slow.pozytyw.art.i.kom
N	34	34	34	34	31	31	34	34	34
	0	0	0	0	3	3	0	0	0
M		1,2059	33,9706	97,9706	300,3824	4,03800	,35639	,6176	55,2941
Me		1,0000	34,0000	87,5000	271,0000	3,98148	,30435	,0000	34,5000
D		1,00	34,00	64,00[a]	10,00[a]	2,562[a]	,050[a]	,00	,00[a]
Σ		,59183	24,61520	73,23209	239,89215	1,076111	,193556	1,53770	51,20105
Var		,350	605,908	5362,939	57548,243	1,158	,037	2,365	2621,547

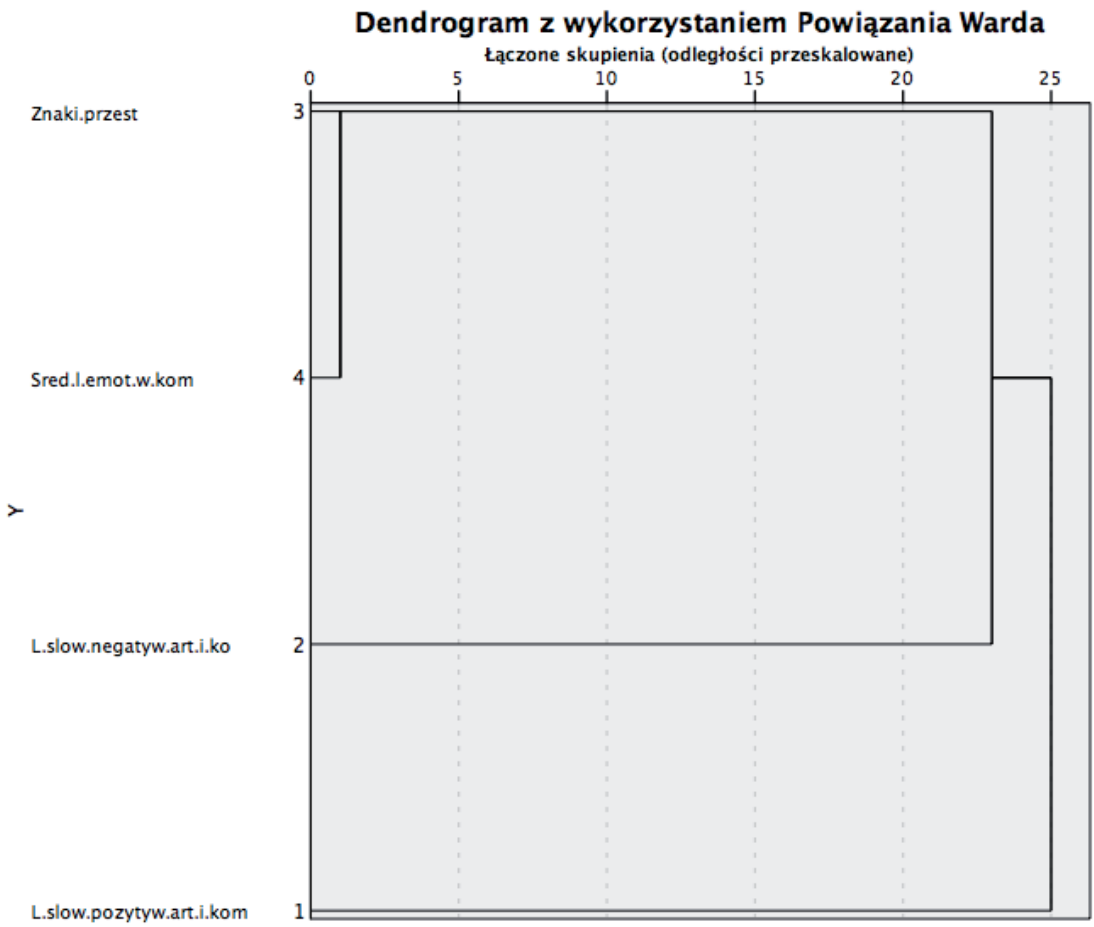
		L.slow.negatyw.art.i.ko	L.slow.błąd.art.i.kom	L.slow.zaufanie.art.i.kom	L.slow.uzytecznosc.art.i.kom	L.slow.nieuzyteczno sc.ar.kom	L.slow.wedza.art.i.kom	L.slow.niewiedza.art.i.kom
N		34	34	34	34	34	34	34
		0	0	0	0	0	0	0
M		68,3235	27,7941	34,6765	42,0588	25,2059	14,1176	9,5294
Me		45,0000	20,5000	22,0000	27,0000	19,5000	10,0000	8,5000
D		1,00[a]	,00[a]	22,00	1,00[a]	,00	8,00	,00
Σ		67,75210	25,58856	36,01071	42,07853	24,61680	13,59092	9,21210
Var		4590,347	654,775	1296,771	1770,602	605,987	184,713	84,863



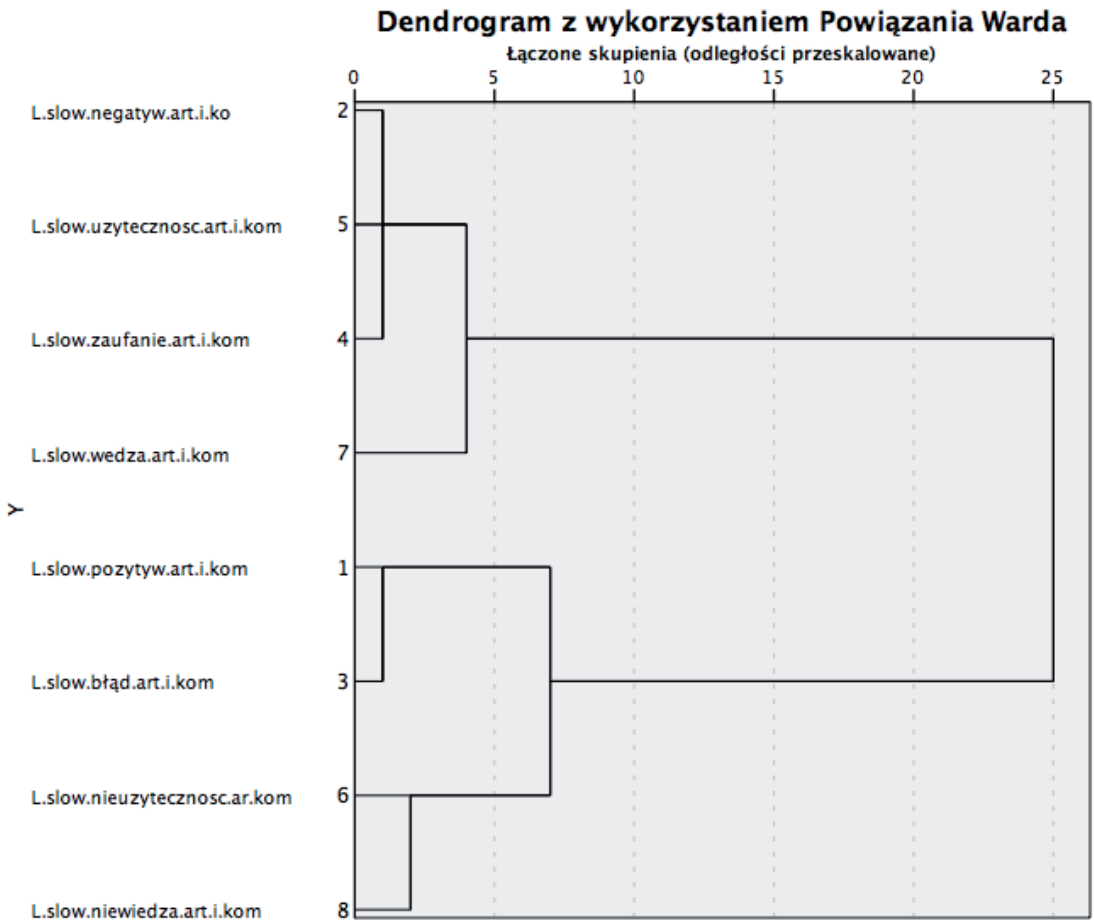
Rys. 9. Dendrogram – wynik agregacji zmiennych: „liczba wystąpień hasła", „liczba wystąpień słów pozytywnych", „liczba wystąpień słów negatywnych", „pozycja wystąpienia 1 hasła"



Rys. 10. Dendrogram – wynik agregacji zmiennych „liczba zdań w artykule", „średnia liczba znaków a w artykule i komentarzach", „liczba wystąpień słów pozytywnych", „liczba wystąpień słów negatywnych"



Rys. 11. Dendrogram – wynik agregacji zmiennych „liczba znaków przestankowych", „średnia liczba emotikonów w artykule i komentarzach", „liczba wystąpień słów pozytywnych", „liczba wystąpień słów negatywnych"



Rys. 12. Dendrogram – wynik agregacji zmiennych „liczba wystąpień słów pozytywnych", „liczba wystąpień słów negatywnych", „liczba wystąpień słów «błąd»", „liczba wystąpień słów «zaufanie»", „liczba wystąpień słów «wiedza»", „liczba wystąpień słów «użyteczność»"

Reasumując, przedstawione wcześniej ekstrakcje przy użyciu hierarchicznej analizy skupień są jedynie przykładem możliwości tworzenia pogłębionej analizy danych zebranych za pomocą prototypu CONTENT 1.0. Aby dokonać selekcji zmiennych (istotnych z punktu widzenia poprawienia efektywności wyboru), należy przeprowadzić dodatkowe eksperymenty na rozszerzonych zbiorach treści internetowych. Niemniej jednak, już na tym etapie analizy, głębokość zbierania danych prototypu (wraz z zakładanym algorytmem ekstrakcji zmiennych) pozwala stwierdzić, że w chwili obecnej nie ma polskim rynku tak zaawansowanego oprogramowania do analizy ilościowej tekstów internetowych.

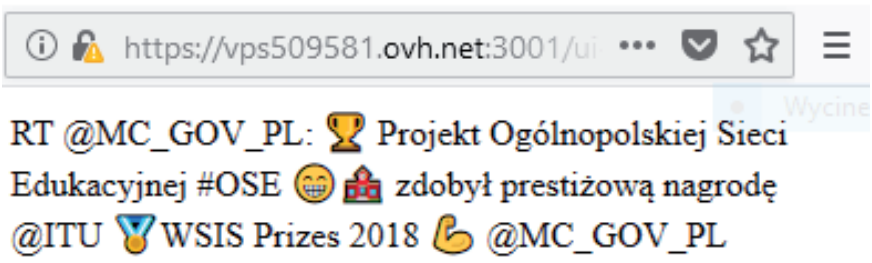
4.2. Analiza jakościowa danych

Głównym celem przeprowadzonej analizy danych było sprawdzenie, czy zaprojektowane i wyprodukowane oprogramowanie umożliwia taką analizę i na ile jest ona funkcjonalna.

Zbieranie danych i ich jakość

Zbieranie danych zostało przetestowane w trakcie realizacji głównego eksperymentu „NASK", który zebrał 325 artykułów z portali Twitter (n = 291), Facebook (n = 27) i Onet (n = 7). W trakcie analizy okazało się jednak, że 45 artykułów nie ma związku z badanym hasłem, które wystąpiło jedynie jako część innego wyrazu, np. **nask**oczyć, **nask**órek. Pozostałe 280 artykułów spełniało wymagania i zostało poddane dalszej analizie.

Program umożliwiał wstępne przeglądanie danych i czytanie całych artykułów w oknach typu pop-up. Co ważne prawidłowo wyświetlały się również wprowadzone w twittach emotikony (rys. 13).



Rys. 13. Przykładowe okno z wpisem z widocznymi emotikonami

Przetestowano wprowadzanie ocen w zdefiniowanych metrykach. Z wykorzystaniem przygotowanej metryki „zgodność" została zrealizowana wstępna ocena związku artykułu z wyszukiwanym hasłem. Zastosowano kodowanie 0-brak związku; 1-jest związek.

Eksport danych

W kolejnym kroku przeprowadzono export danych do pliku w formacie xlsx. Eksport przebiegł pomyślnie, plik wynikowy zawierał wszystkie kategorie danych: artykuły oraz oceny. Artykuły zawierały pełne treści z emotikonami i hiperłączami.

Następnie w programie Excel, plik z pozyskanymi danymi został przygotowany do exportu do specjalistycznego programu MAXQDA poprzez dodanie nagłówków kolumn zgodnie z wymaganiami programu MAXQDA. Dane z tak przetworzonego pliku zostały z powodzeniem zaimportowane do programu MAQDA, w którym zostały poddane dalszemu procesowi nadawania kodów oraz ich kategoryzacji.

Rezultaty analizy jakościowej

W trakcie analizy artykułów ujawniły się następujące kody:

- Informacje o OSE
- Nagroda dla OSE
- Akademia NASK
- Dzień Nowych technologii w Edukacji
- Mistrzowie kodowania/programowania
- Konkurs dla studentów
- Badania
- Edukacja
- Innowacja
- Cyberbezpieczeństwo
- NASK jako dostawca Internetu
- NASK rejestracja domen
- Konferencja SECURE
- Wysokie kompetencje
- Nowy minister
- EDZ

W kolejnym kroku dokonano połączenia kodów w kategorie:

- Profesjonalizm informatyczny
  - NASK jako dostawca Internetu
  - NASK rejestracja domen
  - Cyberbezpieczeństwo
  - Konferencja SECURE
  - Wysokie kompetencje

- Edukacja w społeczeństwie informacyjnym
  - OSE
    - Informacje o OSE
    - Nagroda dla OSE

Tabela 7. Kategoria: Edukacja w społeczeństwie informacyjnym – częstości wystąpień kodów składowych

Kod	Częstość wystąpień
OSE	76
Edukacja	10
Innowacja	10
Mistrzowie kodowania/programowania	8
Akademia NASK	4
Dzień Nowych Technologii w Edukacji	4
Konkurs dla studentów	4
Badania	2
Razem	118

Tabela 8. Profesjonalizm informatyczny – częstości wystąpień kodów składowych

Kod	Częstość wystąpień
Cyberbezpieczeństwo	58
NASK rejestracja domen	11
Konferencja SECURE	4
Wysokie kompetencje firmy	3
Dostawca Internetu	2
Razem	78

- Edukacja
- Innowacja
- Akademia NASK
- Dzień Nowych technologii w Edukacji

  - Mistrzowie kodowania/programowania
  - Konkurs dla studentów
  - Badania
- Ministerstwo
  - Nowy minister
  - EDZ

Na podstawie przeprowadzonej analizy można stwierdzić, że w przeszukiwanych źródłach „NASK" najczęściej występował w związku z realizacją projektu OSE oraz działań edukacyjnych na rzecz rozwoju społeczeństwa informacyjnego. Często również pojawiały się informacje o profesjonalnej działalności informatycznej firmy, szczególnie w zakresie cyberbezpieczeństwa i rejestracji domen.

Szczególnie popularne były twitty o nagrodzie dla projektu OSE (44 wystąpienia) o przykładowej treści:

„RT @MC_GOV_PL: 🏆 Projekt Ogólnopolskiej Sieci Edukacyjnej #OSE 😁👏! zdobył prestiżową nagrodę @ITU WSIS Prizes 2018 💪 @MC_GOV_PL @NASK_pl @M "

lub

„RT @NASK_pl: To wielki sukces! Dziękujemy internautom, ekspertom z @ITU @WSISprocess i wszystkim wspierającym program #OSE – to nas wielki…"

Wysoka częstość tego typu informacji wynikała głównie z tego, że była ona przesyłana dalej przez kolejnych użytkowników internetu, natomiast źródłowe informacje wychodziły z Ministerstwa Cyfryzacji oraz NASK PIB. Podobnie sytuacja wyglądała w przypadki ogólnych informacji odnośnie realizacji projektu OSE.

5

KONKLUZJE I POSTULATY

5.1. Bariery i szanse

Dalej przedstawiono wyniki analizy SWOT dla dostarczonego systemu informatycznego:

Silne strony:

- uwzględnienie struktury artykułu we wszelkich zadaniach analitycznych;
- złożona, elastyczna składnia kwerendy;
- szerokie możliwości parametryzacji widoków wyników;
- tabelaryczny układ wyników, ułatwiający współpracę z innymi narzędziami;
- najlepszy aktualnie dostępny słownik pojęć nacechowanych emocjonalnie;
- możliwość rozwijania i edycji słownika;
- możliwość wystawiania własnych ocen artykułów;
- nowoczesna, rozwojowa architektura informatyczna systemu;
- całkowita kontrola nad kodem źródłowym aplikacji;
- stosunkowo szybka prędkość edycji macierzy danych.

Słabe strony:

- konieczność zastosowania *web scrapingu*, podatnego na zmianę układu stron www skanowanego serwisu onet.pl;
- ograniczenie wydajności skanowania serwisów Facebook i Twitter z uwagi na politykę korzystania z API dostawców;
- ograniczona kontrola nad sposobem wyszukiwania w ww. serwisach;
- sens niektórych metryk ograniczony tylko do niektórych źródeł danych;
- złożona architektura kodu źródłowego, wymagająca zespołu o odpowiednich kwalifikacjach w celu dalszego rozwoju systemu;
- kod źródłowy nie został gruntownie przetestowany ani poddany audytom;
- brak pełnej kompatybilności do eksportu do SPSS Statistics;
- brak obsługi polskiej fleksji i synonimów;
- brak możliwości oznaczania kodem wybranych fragmentów artykułu.

Szanse:

- łatwość rozbudowy do w pełni funkcjonalnej aplikacji sieciowej i komercjalizacji w modelu SaaS;
- gotowość do bieżącego używania przez specjalistów w celu świadczenia komercyjnych usług analitycznych oraz prowadzenia badań naukowych;
- możliwość obudowania wtórnym API dla danych przetworzonych;
- skalowalność pozioma (skanowanie kolejnych, nowych źródeł) i pionowa (większa wydajność skanowania aktualnych serwisów poprzez zrównoleglenie);
- możliwość zawarcia umów partnerskich z dostawcami treści i skanowania na uprzywilejowanych zasadach;
- możliwość adaptacji do pogłębionej ilościowej analizy danych – nieobecnej na polskim rynku;

- łatwość wdrożenia mechanizmów głębszej analizy języka naturalnego.

Zagrożenia:
- restrykcje w polityce dostępu do obecnych i innych źródeł danych;
- rozwiązania konkurencyjne – ograniczona dostępność do dużych zbiorów danych;
- brak czytelnych, uzgodnionych i popartych zasobami ścieżek dalszego rozwoju.

5.2. Kierunki rozwoju

Pod względem architektonicznym, system CONTENT 1.0 umożliwia płynny dalszy rozwój. Już obecnie działa jako aplikacja sieciowa na dzierżawionej maszynie wirtualnej, co umożliwia dalsze skalowanie wydajności zarówno poprzez zwiększenie wydajności maszyny wirtualnej (tj. bez ingerencji w architekturę), jak i poprzez zwielokrotnienie instancji mikrousług i wprowadzenie narzędzi koordynujących (kontenery, kolejki).

Architektura, a zwłaszcza architektura mikrousługowa, wiąże się ściśle z funkcjonalnością, gdyż kluczowe operacje analityczne oraz interfejs do bazy danych realizowane są poprzez dedykowane komponenty. Oznacza to, że modyfikacje określonych aspektów funkcjonowania systemu (np. logiki filtrowania artykułów, zapisu do bazy, a w przyszłości np. uwierzytelnienia i płatności) dokonywane są zazwyczaj tylko w jednym, odpowiedzialnym komponencie. Ułatwia to modyfikacje i dodawanie nowych źródeł danych oraz narzędzi analitycznych.

Wykorzystany słownik „Słowosieć" daje możliwość edycji i tworzenia własnych sub-słowników na podstawie analizy semantycznej indukowanych eksperymentów, niemniej jednak, aby w pełni wykorzystać jego zalety, należy podjąć prace nad rozbudową słownika o katalog polskich fleksji i synonimów.

Osobnym zagadnieniem jest dostęp i agregacja dużych zbiorów danych (co wiąże się z dodatkowymi kosztami), dlatego też, należy podjąć kroki celem tworzenia NASK-owej bazy danych internetowych zarówno dla dominujących mediów, jak i mediów specjalistycznych, np.: edukacja, bankowość, telekomunikacja, technologie cyfrowe itp. Opracowany prototyp aplikacji może być także wykorzystany do analizy danych zastanych czyli archiwów cyfrowych.

BIBLIOGRAFIA

Bochenek Marcin, *Rok pilotażu OSE*, [w:] *Akademia NASK, O OSE*, https://akademia.nask.pl/projekt-48/o-projekcie.html, pobrane dn. 17.07.2018.

Charmaz Katchy, *Teoria Ugruntowana. Praktyczny przewodnik po analizie jakościowej*, WN PWN, Warszawa 2009.

Cox Michael i Ellsworth David, *Managing Big Data for Scientific Visualization*, 1997, ACM SIGGRAPH '97Course #4, Exploring Gigabyte Datasets in Real-Time: Algorithms, Data Management, and Time-Critical Design, Los Angeles, zob.: https://www.researchgate.net/profile/David_Ellsworth2/publication/238704525 _Managing_big_data_for_scientific_visualization/links/54ad79d20cf2213c5fe4081a/Managing-big-data-for--scientific-visualization.pdf, pobrane dn. 13.07.2018.

Everitt Brian S., Landau Sabine, Leese Morven, Stahl Daniel, *Cluster analysis*, 5th edition, John Wiley & Sons, Chichester 2011.

Fischer, Constance T., *Bracketing in qualitative research: Conceptual and practical matters*, „Psychotherapy Research" 2009, 19(4–5), s. 583–590.

Glaser Barney i Strauss Anselm L., *Odkrywanie teorii ugruntowanej. Strategie badania jakościowego*, Zakład Wydawniczy Nomos, Kraków 2009.

Gniadek Anna, Rakowska Weronika, Szladowski Tomasz, *Rynek nazw domeny.pl. Raport roczny.* Wersja elektroniczna zob.: https://www.dns.pl/NASK-raport-rynek-nazw-domeny-pl-2017.pdf, pobrane dn. 10.07.2018.

Gogołek Wodzimierz, *Big Data. Sieciowe źródło informacji dla edukacji,* [w:] *Cyfrowa przestrzeń kształcenia, Seria Cyberprzestrzeń – Człowiek – Edukacja*. Tom 1. Praca zbiorowa pod red. Macieja Tanasia i Sylwii Galanciak, Oficyna Wydawnicza „Impuls", Kraków 2015, s. 97–104.

Gogołek Wodzimierz, Kuczma Paweł, *Rafinacja informacji sieciowych na przykładzie wyborów parlamentarnych. Część 1. Blogi, fora, analiza sentymentów*, „Studia Medioznawcze" 2013, nr 2(53).

Gogołek Wodzimierz, *Rafinacja informacji sieciowej*, [w:] *Informatyka w dobie XXI wieku. Nauka, Technika, Edukacja a nowoczesne technologie informatyczne.* Praca zbiorowa pod red. Aleksandra Jastriebowa, Beaty Kuźmińskiej-Sołśnia, Marii Raczyńskiej, Politechnika Radomska, Radom 2011.

Gogołek Wodzimierz, Jaruga Dariusz, *Z badań nad systemem rafinacji sieciowej. Identyfikacja sentymentów*, „Studia Medioznawcze" 2016, nr 4 (67), s. 104–105.

Holton Judith A., *The Coding Process and Its Challenges*, „The Grounded Theory Review" 2010, vol. 9, nr 1, s. 21–38.

Inteligentne urządzenia wokół nas. A co z naszym bezpieczeństwem?, „Interia Biznes" 17.02.2018, www.biznes.interia.pl, pobrane dn. 13.07.2018.

Katal Avita, Wazid Mohammad, Goudar R.H., *Big Data: Issues, Challenges, Tools and Good Practices*, 2013, Sixth International Conference on Contemporary Computing (IC3), IEEE, Noida, s. 404–409.

Korczak J., Franczyk B., *Big Data – definicje, wyzwania i technologie informatyczne*, „Informatyka Ekonomiczna. Business Informatics" 2014, nr 1(31), s. 141.

Krajobraz bezpieczeństwa polskiego internetu 2016. Raport roczny z działalności CERT Polska, NASK/CERT Polska 2016, s. 23–29.

Laney Doug, *3D DataManagement: Controlling Data Volume, Velocity, and Variety*, „Application Delivery Strategies" 2001, META Group Inc. Zob.: https://blogs.gartner.com/doug-laney/files/2012/01/ ad949-3D-Data-Management-Controlling-Data--Volume-Velocity-and-Variety.pdf, pobrane dn. 13.07.2018.

Locke Karen (2001), *Grounded Theory in Management Research*, Sage, London 2001.

Mayer-Schönberger Wiktor, Cukier Kenneth, *A Revolution that will transform how we live, work and think*, Boston–New York 2013.

Migdał-Najman Kamila, Najman Krzysztof, *Samouczące się sztuczne sieci neuronowe w grupowaniu i klasyfikacji danych. Teoria i zastosowania w ekonomii*, Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk 2013.

Słowosieć, TBC.

Ustawa o Ogólnopolskiej Sieci Edukacyjnej została jednogłośnie przyjęta przez Senat RP 10.11.2017, a następnie podpisana przez Prezydenta RP i ogłoszona 28 listopada w Dzienniku Ustaw 2017, poz. 2184, tom 1.

Ward Joe H., *Hierarchical Grouping in Optimize an Objective Function*, „Journal of the American Statistical Association" 1963, vol. 58.

The Zettabyte Era: Trends and Analysis, White Papers, Cisco,https://www.cisco.com/c/en/us/solutions/ collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.html, pobrane dn. 13.07.2018.

Strony internetowe

http://hadoop.apache.org/.

http://storm.apache.org/.

http://stratosphere.eu/.
https://flink.apache.org/.
https://hadoop.apache.org/.
https://impala.apache.org/.
https://kafka.apache.org/.

O AUTORACH

Maciej Tanaś – kierownik Pracowni Edukacyjnych Zastosowań Technologii Informacyjno-Komunikacyjnych oraz przewodniczący Naukowego Kolegium Ekspertów NASK. Profesor APS i NASK PIB. Doktor *honoris causa* Winnickiego Państwowego Uniwersytetu Pedagogicznego im. Michała Kociubińskiego na Ukrainie (2017). Dziekan Wydziału Nauk Pedagogicznych Akademii Pedagogiki Specjalnej im. Marii Grzegorzewskiej (od 2012). Kierownik Zespołu Pedagogiki Medialnej przy Komitecie Nauk Pedagogicznych PAN, b. doradca społeczny i kierownik Zespołu ds. Bezpieczeństwa Dziecka w Cyberprzestrzeni przy Rzeczniku Praw Dziecka. Członek Rady Naukowej Muzeum Harcerstwa. Kierownik oraz uczestnik wielu międzynarodowych zespołów badawczych. Autor i współautor ponad 250 publikacji naukowych, redaktor naczelny międzynarodowego czasopisma naukowego „International Journal of Pedagogy, Innovation and New Technologies" oraz członek wielu komitetów naukowych polskich i zagranicznych czasopism. Członek Kapituły Medalu KNP PAN „Za Zasługi Dla Rozwoju Polskiej Pedagogiki" oraz Sekcji Pedagogiki Społecznej i Sekcji Pedagogiki Specjalnej przy KNP PAN, Polskiego Towarzystwa Naukowego Edukacji Internetowej, Polskiego Towarzystwa Technologii i Mediów Edukacyjnych, Sekcji Arteterapii

Polskiego Towarzystwa Psychiatrycznego, Jury Międzynarodowego Konkursu Fotograficznego „Matematyka w obiektywie". Naukowo zajmuje się dydaktyką ogólną, metodologią nauk społecznych, pedagogiką medialną i edukacją informatyczną oraz edukacją dla pokoju.

Mariusz Kamola – od 2002 r. stale związany zawodowo z Naukową i Akademicką Siecią Komputerową oraz z Politechniką Warszawską, na której w 2003 r. uzyskał stopień naukowy doktora w dziedzinie automatyki. Jest autorem lub współautorem ponad 50 publikacji naukowych i promotorem pond 40 prac dyplomowych. Prowadził prace badawcze z zakresu symulacji i optymalizacji numerycznej, inżynierii ruchu sieciowego, analizy danych i modelowania matematycznego. Brał udział i kierował projektami badawczymi finansowanymi w ramach 5. i 7. Programu Ramowego UE. Obecne zainteresowania naukowe dra Kamoli obejmują analizę języka naturalnego i Big Data, Internet Rzeczy oraz badania nad sztuczną inteligencją.

Rafał Lange – doktor socjologii; kierownik Pracowni Badań Społecznych w NASK PIB; zajmuje się przede wszystkim metodologią badań, analiza statystyczną, socjologią młodzieży i internetu.

Mariusz Fila – psycholog, pedagog twórczości, pracownik Pracowni Edukacyjnych Zastosowań Technologii Informacyjno-Komunikacyjnych NASK PIB oraz Zakładu Metodologii i Pedagogiki Twórczości Akademii Pedagogiki Specjalnej im. Marii Grzegorzewskiej. Prowadzi prace badawcze i wdrożeniowe z zakresu komputeryzacji kształcenia oraz metodologii badań. Kierował międzynarodowymi projektami, w tym: *Innovation Laboratories in the development of competences of special pedagogy teachers and people with special educational needs* (i-LAB3).

Informacje o NASK PIB

NASK
PAŃSTWOWY INSTYTUT BADAWCZY

NASK Państwowy Instytut Badawczy jest instytutem badawczym podległym Ministerstwu Cyfryzacji. Kluczowe obszary działalności NASK PIB obejmują zadania związane z zapewnieniem bezpieczeństwa internetu, a także z rozwojem polskiej cyberprzestrzeni. Instytut realizuje działania statutowe działając w różnych obszarach: naukowym, doradczym, edukacyjnym i gospodarczym.

W ramach NASK PIB działa Narodowe Centrum Cyberbezpieczeństwa (NC Cyber). Reagowaniem na zdarzenia naruszające bezpieczeństwo sieci zajmuje się zespół CERT Polska (Computer Emergency Response Team). W NC Cyber funkcjonuje także zespół Dyżurnet.pl, odpowiadający za przeciwdziałanie szkodliwym i nielegalnym treściom obecnym w internecie.

Instytut prowadzi badania w zakresie opracowywania rozwiązań zwiększających efektywność, niezawodność i bezpieczeństwo sieci teleinformatycznych oraz innych złożonych systemów sieciowych. Istotne miejsce w działalności instytutu zajmują badania dotyczące biometrycznych metod weryfikacji tożsamości w bezpieczeństwie usług. NASK PIB prowadzi także rejestr domeny.pl.

Funkcjonująca w strukturach instytutu Akademia NASK zajmuje się działalnością edukacyjną, popularyzatorską oraz szkoleniową.

Wieloletnia współpraca z ekspertami oraz przedstawicielami środowisk naukowych pozwoliła stworzyć szeroką gamę publikacji, poradników i materiałów edukacyjnych poruszających najbardziej aktualne zagadnienia związane z bezpieczeństwem dzieci i młodzieży online. Akademia NASK realizuje projekty adresowane do różnych grup społecznych, wiekowych oraz zawodowych. Od 2005 roku NASK

PIB jest koordynatorem Polskiego Centrum Programu Safer Internet – programu Komisji Europejskiej mającego na celu promocję bezpiecznego korzystania z nowych technologii i internetu wśród dzieci i młodzieży oraz przeciwdziałanie nielegalnym treściom online.

W Akademii NASK prowadzone są unikatowe szkolenia dla firm i instytucji ze szczególnym uwzględnieniem tematyki bezpieczeństwa ICT. Oferta szkoleniowa Akademii adresowana jest do sektora biznesu, administracji publicznej i instytucji akademickich. Posiadamy także kompleksową propozycję szkoleń społecznych dla samorządów oraz przedstawicieli sektora edukacyjnego.

W instytucie NASK PIB istotną rolę pełni Pracownia Edukacyjnych Zastosowań TIK. Pracownia zajmuje się prowadzeniem badań społecznych z obszaru społeczeństwa informacyjnego oraz implementacją technologii informacyjnych i komunikacyjnych w procesie edukacji. Istotnym zadaniem Pracowni jest diagnoza stanu bezpieczeństwa cyfrowego dzieci i młodzieży. Pracownia współpracuje z wiodącymi ośrodkami akademickimi i instytucjami naukowo-badawczymi oraz posiada zaplecze informatyczne i technologiczne w realizacji badań zleconych.

NASK – Państwowy Instytut Badawczy
ul. Kolska 12, 01-045 Warszawa
tel. 22 380 82 00, fax 22 380 82 01, nask@nask.pl
www.nask.pl

Wydanie pierwsze
Arkuszy drukarskich 5,25
Skład i łamanie: AnnGraf, Anna Szeląg
Druk ukończono w maju 2019
Druk i oprawa: Fabryka Druku

Raport badawczy NASK PIB ***CONTENT 1.0 – prototyp aplikacji do analizy treści internetu***, przygotowany przez zespół badawczy pod kierunkiem prof. Macieja Tanasia, wiąże się z nurtem poszukiwania nowych narzędzi do analizy i przetwarzania wielkich zbiorów danych, oraz ich zastosowania w metodologii badań społecznych i edukacyjnych, w czasie gdy gwałtownie rosną przypływy informacji z wielu różnych źródeł. Dane te mają ogromną użyteczność dla nauki, edukacji, gospodarki czy polityki, co rodzi pilną potrzebę tworzenia nowych metod i technik analizy Big Data, oraz nowych rozwiązań technologicznych, otwierających zupełnie nowe perspektywy poznawcze przed nauką i edukacją, pozwalające zdobyć bezcenną wiedzę o przestrzeni, w jakiej żyjemy.

Temu właśnie ma służyć m.in. przedstawiony w Raporcie projekt CONTENT 1.0, umożliwiający podejmowanie takich analiz z zastosowaniem wielowymiarowej analizy semantycznej treści zeskanowanych danych źródłowych ze stron internetowych i portali społecznościowych. Pierwsze eksperymenty z jego zastosowaniem pokazały, że stwarza on nie tylko duże możliwości analityczne, ale jego otwarty charakter pozwala też na wzbogacanie go o nowe elementy, co wydatnie umożliwia dalszy rozwój i zwiększa potencjalne możliwości jego wykorzystania w badaniach nad edukacją, czy szerzej – w obszarze nauk społecznych.

dr hab. Barbara Galas, prof. UKSW

Stanisław Lem przyrównał zjawisko internetu do biblijnego potopu, czyli nadmiaru wód, w którym można ze wszystkim utonąć, jeżeli nie zdołamy dla ratunku, jak Noe, zbudować sobie „Arki Noego Internetu". Człowiek potrafi takie łodzie budować, czego przykładem opiniowany produkt nazwany CONTENT 1.0, będący efektem pracy zespołu badawczego w składzie: mgr Mariusz Fila, dr inż. Mariusz Kamola, dr Rafał Lange oraz dr hab., prof. APS Maciej Tanaś – kierownik. Wynikiem pracy tego zespołu jest skonstruowana z myślą o przeszukiwaniu zasobów internetu aplikacja, stanowiąca swoiste narzędzie użyteczne w wyszukiwaniu haseł w postaci słów, pojedynczych zdań lub ciągu tych zdań.

Otrzymany rezultat charakteryzuje się zamierzoną adaptatywnością oraz wysoką efektywnością, co potwierdziły przeprowadzone testy. Testy te [...] dowodzą, że umiejętne wykorzystanie wytworzonego narzędzia może prowadzić do interesujących poznawczo wyników. Rzetelność oraz niezwykła wnikliwość przeprowadzonych eksperymentów potwierdza, znaną od dawna prawdę, że nietrywialne sposoby przetwarzania danych dają nietrywialne w swej treści syntezy. CONTENT 1.0 [...] posiada niezaprzeczalne cechy dokonania twórczego, jest bowiem egzemplifikacją jednej z definicji twórczości, mówiącej, że twórczość to także algorytmizacja niealgorytmizowalnego.

dr hab. Jan Łaszczyk, prof. APS

www.aps.edu.pl

ISBN 978-83-66010-29-1

9 788366 010291