# WebSty - otwarty webowy system do analiz stylometrycznych

**Maciej Piasecki, Tomasz Walkowiak**
Politechnika Wrocławska
**Maciej Eder**
Instytut Języka Polskiego PAN i Uniwersytet Pedagogiczny im KEN w Krakowie

Metody analizy stylometrycznej mogą być pomocne nie tylko przy ustalaniu autorstwa lub wyznaczaniu tekstów, które zostały napisane przez tego samego autora. Można je stosować do określania grup tekstów, które wykazują wspólne cechy, np. cechy stylu literackiego, autorskiego, gatunkowego, cechy wynikające ze specyficznego pochodzenia tekstu itp. Metody tego typu mogą znaleźć zastosowanie poza tradycyjnymi obszarami zastosowania w literaturoznawstwie lub językoznawstwie sądowym, np. kulturoznawstwie, socjologii, politologii lub historii. Przeszkodą w szerszym stosowaniu może być konieczność doboru odpowiedniego oprogramowania do analizy statystycznej, jego instalacja oraz jego powiązanie z programami do analizy wstępnej tekstu, np. w zakresie analizy morfo-syntaktycznej.
WebSty (http://ws.clarin-pl.eu/demo2/websty.shtml) to otwarty, sieciowy system stylometryczny o interfejsie użytkownika w postaci strony WWW (tzw. aplikacja webowa). WebSty został zbudowany i jest utrzymywany w ramach infrastruktury CLARIN-PL (www.clarin-pl.eu). WebSty nie wymaga instalacji, może być użytkowany poprzez dowolną przeglądarkę WWW, udostępnia szereg gotowych procedur działania jak i możliwość szczegółowej konfiguracji sesji przetwarzania. Posiada budowę modułową umożliwiającą wewnętrzne wykorzystanie szeregu gotowych narzędzi językowych do przetwarzania tekstu jak i systemów do analizy statystycznej i grupowania tekstów, np. Stylo (Edera i Rybickiego), Cluto, Orange, czy też SciPy.

W trakcie wystąpienia zostanie zaprezentowana koncepcja systemu WebSty, proces przetwarzania tekstu, przewidziane scenariusze użycia, przegląd dostępnych cech opisu dokumentów, metod analizy danych statystycznych oraz wizualizacji wyników analizy. W obecnej wersji WebSty dokumenty opisywane są za pomocą częstości: wystąpień wyrazów, lematów (form podstawowych), cech gramatycznych (części mowy, klasy gramatyczne, kategorie gramatyczne, np. przypadek lub liczba), klas semantycznych nazw własnych czy też dziedzin tematycznych słów w tekście. W celu ustalenia wartości cech tekst jest przetwarzany poprzez automatycznie zestawiany potok narzędzi językowych. System umożliwia przetwarzanie tekstów polskich, angielskich i niemieckich. Pierwotne wartości cech w postaci częstości wystąpień mogą następnie zostać poddane filtrowaniu (np. po progu minimalnej wartości) oraz przekształceniu za pomocą funkcji wagi cechy dla dokumentu (np. w oparciu o statystyczne miary asocjacyjne lub miary oparte na teorii informacji). Podobieństwo otrzymanych wektorów danych dla dokumentów może zostać

obliczone jedną z wielu funkcji podobieństwa (np. miarą kosinusową lub deltą Burrowsa). Następnie WebSty umożliwia wykorzystanie szeregu narzędzi do grupowania (np. Stylo lub Cluto) oraz oferuje szereg metod wizualizacji (np. dendrogramy, mapy cieplne, dynamiczne grafy aglomeracyjne). Dla ustalonych grup dokumentów można wyznaczyć ich cechy charakterystyczne: opisujące i różnicujące. Wyniki analizy można pobrać do pliku (np. w formacie CSV).

WebSty umożliwia załadowanie plików w wielu znanych formatach. Większe zbiory mogą być ładowane poprzez repozytorium CLARIN-PL.

W dalszej części wystąpienia omówimy dalsze plany rozwoju systemu WebSty. Między innymi pełniejszą obsługę metod opartych na maszynowym uczeniu się i klasyfikacji. Rozszerzony zostanie zestaw cech, np. o cechy oparte na sekwencjach liter, ale też o wyniki analizy leksykalno-semantycznej oraz parsowania. Wprowadzone zostaną automatyczne mechanizmy optymalizacji zestawu cech i redukcji wymiarowości. Funkcjonalność umożliwiająca identyfikacji cech istotnych dla wyników grupowania zostanie rozszerzona umożliwiając badanie wpływu cech na strukturę podobieństwa w zbiorze tekstów. W znacznym zakresie zostaną rozbudowane metody wizualizacji wyników w oparciu o interaktywne drzewa oraz skalowanie wielowymiarowe na przestrzeń 2D i 3D (z obrazem ruchomym i stereoskopią).

Całość systemu będzie dążyła do integracji mechanizmów analizy stylometrycznej wraz z klasyfikacją semantyczną tekstów.