FileStat (wersja 1.4) – program do porównywania plików tekstowych

Program FileStat służy porównywaniu wejściowych plików tekstowych przy pomocy zdefiniowanej miary podobieństwa. Przy ocenie stopnia podobieństwa plików istnieje możliwość zastosowania miary podobieństwa pomiędzy poszczególnymi formami wyrazowymi.

Program oblicza również szereg statystyk charakteryzujących pliki tekstowe:

- rozmiar pliku,
- liczba znaków (włącznie ze znakami końca linii),
- liczba linii tekstu,
- liczba wyrazów,
- liczba różnych form wyrazowych,
- średnia długość wyrazu,
- liczba zdań,
- średnia długość zdania,
- liczba paragrafów,
- średnia długość paragrafu.

Ponadto, program oblicza liczbę wystąpień form wyrazowych i bigramów a także współczynniki IPF oraz PMI, o których będzie jeszcze mowa poniżej.

Program napisany został w języku Java i uruchomiony może być na każdej platformie sprzętowej, na której zainstalowano Java Runtime Environment w wersji 7 lub wyższej.

1 Miary podobieństwa plików tekstowych

Głównym zadaniem programu FileStat jest porównywanie zawartości wejściowych plików tekstowych.

Pliki tekstowe D_X i D_Y są reprezentowane odpowiednio przez wektory $X = \{x_1, \ldots, x_n\}$ oraz $Y = \{y_1, \ldots, y_n\}$, gdzie x_i i y_i oznaczają liczbę wystąpień formy wyrazowej w_i ze zbioru $W = \{w_1, \ldots, w_n\}$ wszystkich form zawartych w plikach D_X i D_Y. Program stosuje dwie miary podobieństwa plików – miarę prostą araz miarę kosinusową.

Miara prosta $sim(D_X, D_Y)$ wyraża stosunek części wspólnej obu porównywanych plików D_X i D_Y do ich sumy, co określić można w następujący sposób:

$$sim_s(D_X, D_Y) = \frac{X \cap Y}{X \cup Y},$$

gdzie $X \cap Y$ i $X \cup Y$ są wyrażone odpowiednio wzorami:

$$X \cap Y = \sum_{i=0}^{n} min(x_i, y_i),$$

$$X \cup Y = \sum_{i=0}^{n} max(x_i, y_i).$$

Miara kosinusowa $sim_c(D_X, D_Y)$ jest określona przez kosinus kąta pomiędzy wektorami X i Y.

$$sim_c(D_X, D_Y) = \frac{X \circ Y}{|X| \cdot |Y|},$$

gdzie $X \circ Y$ i $|X| \cdot |Y|$ są wyrażone odpowiednio wzorami:

$$X \circ Y = \sum_{i=0}^{n} x_i y_i,$$

$$|X| \cdot |Y| = \sqrt{\sum_{i=0}^{n} x_i^2} \cdot \sqrt{\sum_{i=0}^{n} y_i^2}.$$

2 Miara podobieństwa wyrazów

Powyższe miary podobieństwa plików możemy zmodyfikować uwzględniając podobieństwo badanych form wyrazowych. Program stosuje miarę podobieństwa bazującą na zmodyfikowanej odległości edycyjnej Levenshteina. Odległość edycyjna Levenshteina wyraża minimalną liczbę podstawowych działań niezbędnych do

przkształcenia jednego ciągu znaków w drugi, przy czym przez podstawowe działanie rozumiemy wstawienie lub usunięcie znaku, względnie zamianę jednego znaku na drugi. Za odległość edycyjną można przyjąć sumę kosztów wszystkich wykonanych podstawowych działań edycyjnych. W podstawowej definicji odległości Levenshteina przyjmuje się, że koszt każdego podstawowego działania wynosi 1.

Program FileStat umożliwia modyfikację kosztu działania podstawowego takiego jak zamiana wielkości liter lub usunięcie (względnie wstawienie) znaku diakrytycznego. Ponadto, program umożliwia uzależnienie kosztu działania od pozycji w ciągu znaków, na której zostało ono wykonane, co szczególnie jest przydatne przy porównywaniu odmienionych przez przypadki form wyrazowych różniących się tylko końcówkami.

Miarą podobieństwa między dwoma ciągami znaków zastosowaną w programie jest stosunek kosztu przekształcenia jednego ciągu znaków w drugi do maksymalnej liczby znaków występujących w porównywanych ciągach. Innymi słowy, jeśli koszt przkształcenia ciągu a w ciąg b wynosi k, to miara ta wyrażona jest wzorem

$$sim(a, b) = \frac{k}{max(|a|, |b|)},$$

gdzie $|a|$ i $|b|$ oznaczają długości ciągów a oraz b.

W przypadku bigramów $A = (a_1, a_2)$ i $B = (b_1, b_2)$,

$$sim(A, B) = sim(a_1, a_2) \cdot sim(b_1, b_2).$$

Zanim program zastosuje miarę podobieństwa wyrazów w procesie porównywania plików, musi przeliczyć frekwencje wystąpień wszystkich form wyrazowych. Sam algorytm porównywania plików nie zmienia się, zmieniają się tylko liczby reprezentujące frekwencje występowania poszczególnych form wyrazowych. Jeśli dwie formy wyrazowe a i b są podobne w stopniu x, nie mniejszym niż pewna wartość progowa, forma a wystąpiła w badanym pliku n razy, zaś forma b wystąpiła m razy, to po przeliczeniu frekwencje te wynoszą odpowiednio $n+xm$ oraz $m+xn$.

Trzeba pamiętać, że algorytm przeliczania frekwencji jest kosztowny, ponieważ porównywane są między sobą wszystkie formy wyrazowe. Złożoność tego algorytmu to $\mathcal{O}(n^2/2)$, gdzie n oznacza liczbę wszystkich form wyrazowych.

3 Współczynnik IPF

Jedną z wartości wyliczanych dla form wyrazowych jest wsółczynnik IPF (ang. *inverse paragraph frequency*). Odpowiada on powszechnie znanemu współczynnikowi IDF (ang. *inverse documant frequency*), z tą tylko różnicą, że rolę dokumentów odgrywają poszczególne paragrafy teksu w pliku. Jeśli P jest zbiorem wszystkich

paragrafów w badanym pliku tekstowym, to dla danej formy wyrazowej w, współczynnik ten definiujemy następująco:

$$ipf(w) = \log\left(\frac{|P|}{|\{p \in P : w \in p|}\right)$$

Jest to zatem wartość logarytmu dziesiętnego ze stosunku ogólnej liczby paragrafów, do liczby tych paragrafów, w których występuje forma w.

Wzór ten można nieco zmodyfikować w przypadku, gdy interesować nas będą wystąpienia form podobnych do danej formy w.

$$ipf(w) = \log\left(\frac{|P|}{\|P_w\|}\right),$$

gdzie P_w jest zbiorem paragrafów p_w zawierających formy podobne do formy w, $\|P_w\| = \sum_{p_w \in P_w} \|p_w\|$, a $\|p_w\|$ oznacza maksymalną wartość podobieństwa wyrazów podobnych do wyrazu w występujących w paragrafie p_w.

4 Współczynnik PMI

Wartością wyliczaną dla bigramów jest współczynnik PMI (ang. *pointwise mutual information*), a właściwie jego odmiana (*normalized pointwise mutual information*). Współczynnik ten określa rozbieżność pomiędzy częstością występowania danego bigramu (x, y) w tekście, a hipotetyczną częstością jego występowania wynikającą z częstości występowania jego składowych x i y.

Niech $C(x, y)$ oznacza liczbę bigramów (x, y), zaś N liczbę wszystkich bigramów występujących w dokumencie D. Współczynnik PMI wyrażony jest następującym wzorem:

$$\begin{aligned} pmi(x,y) &= \left(\ln \tfrac{p(x,y)}{p(x)p(y)}\right) \div -\ln p(x,y), \\ p(x,y) &= \tfrac{C(x,y)}{N}, \\ p(x) &= \tfrac{\sum_{t \in D} C(x,t)}{N}, \\ p(y) &= \tfrac{\sum_{t \in D} C(t,y)}{N}. \end{aligned}$$

Jeśli będziemy brać pod uwagę podobieństwo bigramów, to liczba $C(x, y)$ oznaczać będzie liczbę wystąpień wszystkich bigramów podobnych do (x, y) w sensie miary podobieństwa bigramów opisanej w rozdziale 2. Wówczas za $C(x, y)$ przyjmować będziemy następującą liczbę:

$$C(x,y) = \sum_{Z \in B} sim(T, Z),$$

gdzie $T = (x, y)$ oraz $B = \{Z = (x', y') : Z$ jest podobny do $T\}$.

5 Interfejs użytkownika

Podstawowym elementem interfejsu użytkownika programu FileStat jest okno główne (Rysunek 1), w którym wyświetlane są statystyki dotyczące otwartych plików oraz wyniki ich porównania. Działanie programu kontroluje się przy pomocy menu oraz paska narzędzi znajdującego się na górze okna.

Pliki do analizy wybieramy poleceniem **Open** lub **Open from URL** znajdującymi się w menu **File**, albo przyciskiem  na pasku narzędzi, co odpowiada poleceniu **Open**.

Oprócz plików tekstowych (z rozszerzeniem TXT), program obsługuje także pliki w formacie PDF oraz pliki utworzone przez programy Microsoft Word (DOC, DOCX) i Open Office (ODT).

Wprzypadku plików o rozszerzeniu TXT można wybrać jedną z metod kodowania znaków:

- UTF8 (Eight-bit Unicode Transformation Format),
- Cp1250 (Windows Eastern European),
- Cp852 (MS-DOS Latin-2),
- Cp870 (IBM Multilingual Latin-2),
- ISO-8859-2 (ISO 8859-2, Latin Alphabet No. 2),
- MacCentralEurope (Macintosh Latin-2).

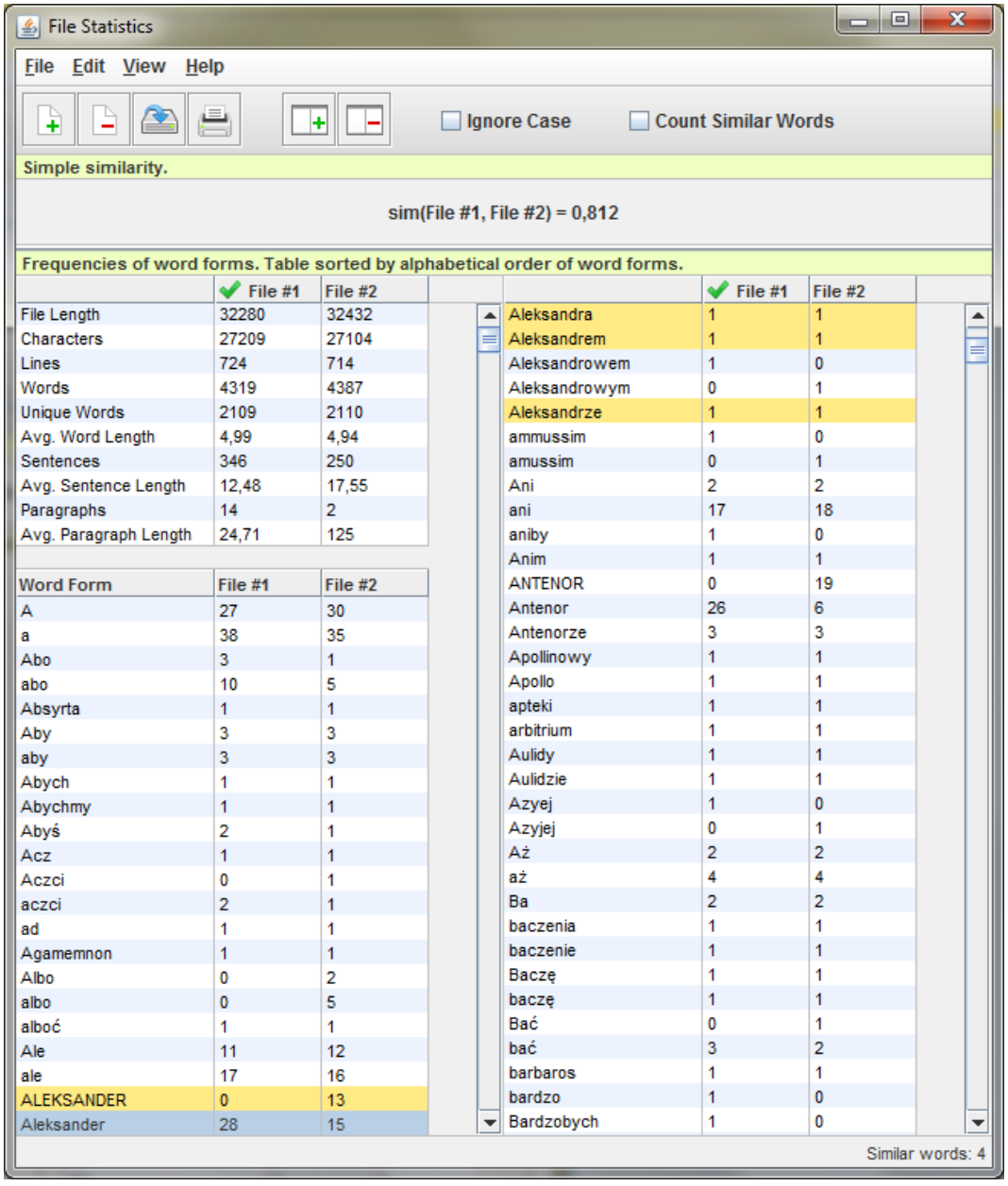
Domyślnym kodowaniem plików tekstowych jest format UTF8.

Wybierając plik do analizy możemy ustalić metodę dzielenia tekstu na paragrafy. Może być to znak końca linii lub pusta linia tekstu. W przypadku plików TXT najodpowiedniejsza wydaje się ta druga metoda. W przypadku plików utworzonych przez takie programy jak Microsoft Word, czy Open Office, bardziej naturalną jest metoda pierwsza. Domyślnym separatorem paragrafów jest pusta linia.

Pliki o nierozpoznanym przez program rozszerzeniu traktowane są jak pliki tekstowe TXT.

W przypadku polecenia **Open from URL** należy podać adres URL pliku, który chcemy pobrać z internetu oraz nazwę użytkownika i hasło, jeśli podany adres internetowy jest zabezpieczony hasłem.

Po wczytaniu pliku, w głównym oknie programu wyświetla się tabela zawierająca różne statystyki dotyczące tego pliku oraz, dla każdej formy wyrazowej lub bigramu, wybraną wartość. Wartością tą może być frekwencja (absolutna lub procentowa) występowania, lub współczynnik IPF albo PMI, w zależności od tego, czy tabela zawiera formy wyrazowe, czy bigramy. Tabelę tę można podzielić na kilka

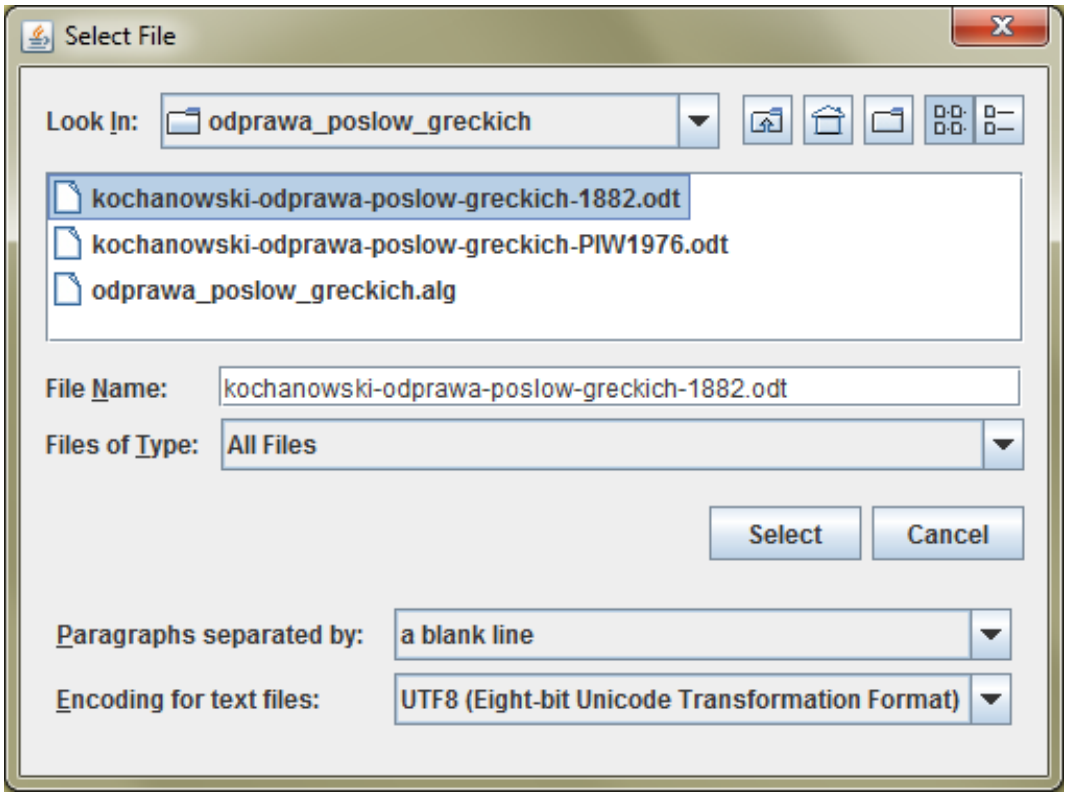


	File #1	File #2
File Length	32280	32432
Characters	27209	27104
Lines	724	714
Words	4319	4387
Unique Words	2109	2110
Avg. Word Length	4,99	4,94
Sentences	346	250
Avg. Sentence Length	12,48	17,55
Paragraphs	14	2
Avg. Paragraph Length	24,71	125

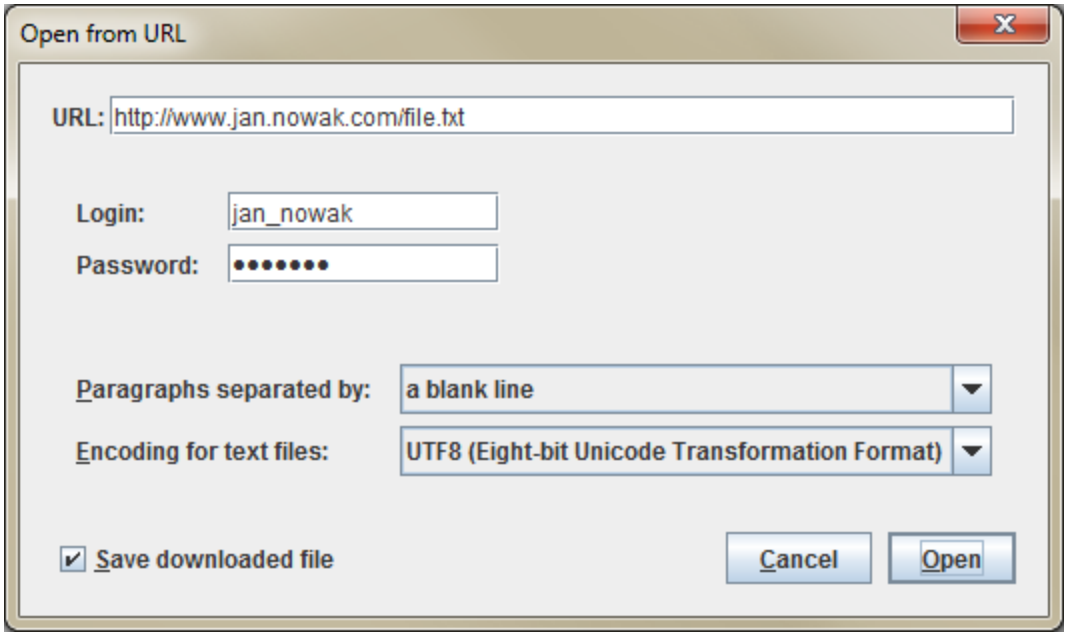
Word Form	File #1	File #2
A	27	30
a	38	35
Abo	3	1
abo	10	5
Absyrta	1	1
Aby	3	3
aby	3	3
Abych	1	1
Abychmy	1	1
Abyś	2	1
Acz	1	1
Aczci	0	1
aczci	2	1
ad	1	1
Agamemnon	1	1
Albo	0	2
albo	0	5
alboć	1	1
Ale	11	12
ale	17	16
ALEKSANDER	0	13
Aleksander	28	15

	File #1	File #2
Aleksandra	1	1
Aleksandrem	1	1
Aleksandrowem	1	0
Aleksandrowym	0	1
Aleksandrze	1	1
ammussim	1	0
amussim	0	1
Ani	2	2
ani	17	18
aniby	1	0
Anim	1	1
ANTENOR	0	19
Antenor	26	6
Antenorze	3	3
Apollinowy	1	1
Apollo	1	1
apteki	1	1
arbitrium	1	1
Aulidy	1	1
Aulidzie	1	1
Azyej	1	0
Azyjej	0	1
Aż	2	2
aż	4	4
Ba	2	2
baczenia	1	1
baczenie	1	1
Baczę	1	1
baczę	1	1
Bać	0	1
bać	3	2
barbaros	1	1
bardzo	1	0
Bardzobych	1	0

Rysunek 1: Główne okno programu FileStat



Rysunek 2: Wybór pliku do analizy



Rysunek 3: Wybór pliku do analizy ze strony internetowej

fragmentów umieszczając je obok siebie. Na rysunku 1 tabela została podzielona na dwie części. Do tego celu służy polecenie **Add Right View** w menu **View** lub przycisk na pasku narzędzi. Taki podział likwidowany jest poleceniem **Remove Right View** lub przyciskiem .

Tabela poszerza się o dane kolejno wczytywanych plików. Plikom tym nadaje się identyfikatory: *File1*, *File2*, itd. Gdy chcemy uzyskać informację o nazwie pliku i jego lokalizacji, należy ustawić wskaźnik myszy na nagłówku tabeli z identyfikatorem tego pliku. Wówczas jego nazwa i lokalizacja wyświetli się na dole ekranu. Klikając myszą w identyfikator pliku w nagłówku tabeli dokonujemy jego selekcji. Wybrany plik zaznaczony jest znakiem ✔. Dla wybranych w ten sposób plików dostępne są dodatkowe polecenia. Wykonanie polecenia **Close** z menu **File** (przycisk) skutkuje usunięciem wybranego pliku z tabeli wyników. Zawartość wybranego pliku można podejrzeć przy pomocy polecenia **View Selected File** znajdującego się w menu **View**.

Rodzaj wyświetlanej tabeli możemy zmienić przy pomocy poleceń **Word Forms** i **Bigrams** znajdujących się w menu **View**.

Tabela wartości liczbowych może zawierać frekwencje absolutne, tzn. liczby wystąpień danych form wyrazowych (bigramów) w poszczególnych plikach, frekwencje względne w stosunku do ogólnej liczby występujących form wyrazowych (bigramów) wyrażone w procentach, a także współczynniki IPF lub PMI. Wy-

boru typu wyświetlanych wartości można dokonać przy pomocy poleceń z menu **Calculate** z poziomu menu **View**.

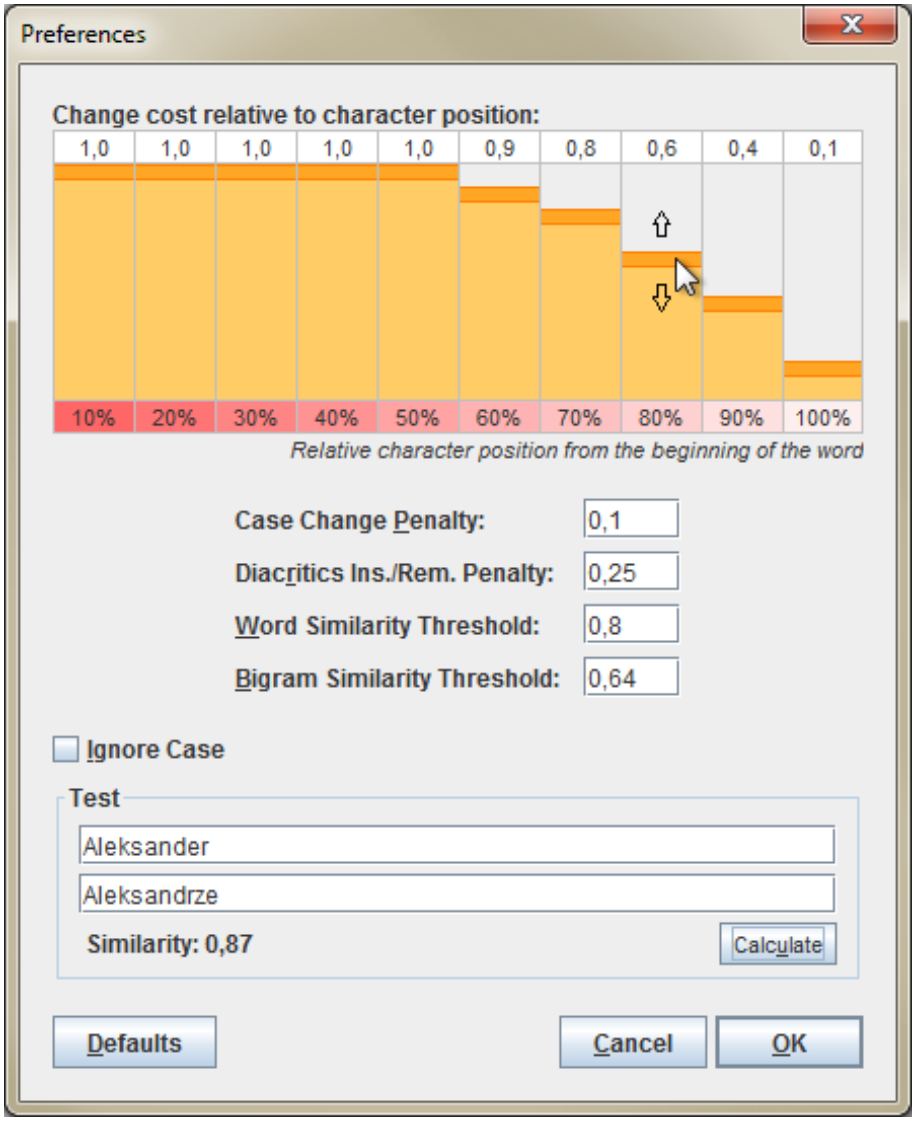
Tabelę wartści można posortować na trzy sposoby: alfabetycznie, względem obliczonych wartości we wszystkich plikach oraz względem obliczonych wartości w wybranym pliku. Polecenia sortujące tabelę znajdują się w menu **Sort** dostępnym z poziomu menu **View**.

Jeśli wczytane zostały przynajmniej dwa pliki, to oprócz wspomnianej już tabeli, wyświetlana jest także informacja o wzajemnym stopniu podobieństwa plików. Wyboru odpowiedniej miary podobieństwa dokumentów można dokonać z menu **Similarity Measure** dostępnego z poziomu menu **View**. W tym samym menu znajduje się polecenie, dzięki któremu możemy uwzględnić lub pominąć podobieństwo wyrazów w procesie porównywania plików. To samo polecenie jest również dostępne na pasku narzędzi.

Parametry wykorzystywane przez algorytm badania podobieństwa wyrazów możemy zmodyfikować przy pomocy dialogu przedstawionego na rysunku 4. Dialog ten wywołujemy poleceniem **Preferences...** w menu **Edit**[1]. Przy pomocy tego dialogu mażemy zmodyfikować koszt zmiany wielkości liter, koszt wstawienia lub usunięcia znaku diakrytycznego, wartość progową, od której dwa wyrazy uznaje się za podobne oraz współczynniki kosztu podstawowego działania edycyjnego w zależności od pozycji w ciągu znaków, na której działanie to zostało wykonane. Możemy także poinstruować program, aby nie brał pod uwagę wielkości liter. To samo polecenie dostępne jest na pasku narzędzi.

Wyniki programu możemy zapamiętać na dysku lub wydrukować, czemu służą odpowiednio polecenia **Save...** () i **Print** ().

[1]Pod systemem Mac OS X polecenie to jest umieszczone standardowo w głównym menu aplikacji.



Rysunek 4: Parametry podobieństwa wyrazów