DOKUMENTACJA WSTĘPNA – WYKRYWANIE PARAGRAFÓW

WEDT 17L

Mariusz Pajączkowski
Mikołaj Płachta

1. Wymagania funkcjonalne.

Celem projektu jest opracowanie metody wykrywania paragrafów oraz zaimplementowanie jej w formie narzędzia. Narzędzie będzie badało różne formy tekstu wejściowego:

- sformatowany tekst w języku naturalnym(na co dzień spotykany w książkach)
- tekst sformatowany dla standardów internetowych(html oraz xml)

Narzędzie zostanie przygotowane w formie aplikacji webowej napisanej w języku Java uruchamianej za pomocą serwera aplikacyjnego. Po wykonaniu narzędzia zostaną przeprowadzone manualne testy sprawdzające poprawność działania metody.

Narzędzie będzie przystosowane tylko do procesowania dokumentów w języku angielskim.

2. Definicja paragrafu.

Paragraf jest to fragment tekstu o ilości przynajmniej dwóch zdań zawierający logiczną oraz odnoszącą się do siebie całość myśli autora. Najpopularniejszą formą wykorzystania paragrafu w sformatowanym języku naturalnym jest akapit. Poza sytuacją klasyczną paragrafem może być fragment wiersza zawarty w jednej zwrotce. Do paragrafów nie będziemy zaliczać takich fragmentów tekstu jak tytuły czy przypisy.

W przypadku języków internetowych podstawową formą paragrafu będzie akapit oznaczany znacznikiem <p> w odpowiednim połączeniu ze znakiem pustej linii
.

3. Algorytm

Działanie narzędzia rozpocznie się od wczytania pliku pdf z tekstem naturalnym bądź pliku z tekstem w standardzie internetowym(html lub xml). Następnie zostanie on przeprocesowany, czego wynikiem będą wyznaczone paragrafy zgodnie z definicją z powyższego punktu. Wynik zostanie przedstawiony w formie wygenerowanego pliku tekstowego z paragrafami oznaczonymi za pomocą tagów „<paragraf>”.