Zadania domowe – Przetwarzanie tekstów 2

Gotowe zadanie domowe należy przesłać na adres junczys@amu.edu.pl . Należy dołączyć wszystkie pliki potrzebne do poprawnego wyświetlenia wraz z krótkim **komentarzem** do każdego zadania.

Proszę umieścić w mailu Imię, Nazwisko, Specjalizację i Rok. Termin złożenia zadania domowego to 16.04.2008.

Zadanie 1

Napisać program, który usuwa z dowolnego pliku HTML wszystkie znaczniki HTML, w tym treść komentarzy i skryptów. Na wyjściu może się pojawić tylko sam tekst strony. Program nie ma usuwać tekstów pojawiających się w menu lub podobnych miejscach. Zachowujemy całość tekstu widocznego w przeglądarce www, czyli też opisy do linków, stopki strony itp.

Zastosować program do pliku gazeta1.html

Wskazówka: Ewentualnie warto wczytać cały plik do jednej zmiennej skalarnej.

Punkty: 2

Zadanie 2

Dokończyć program z zajęć. Tzn. program ma przetworzyć plik gazeta1.html zawierający artykuł z Gazety Wyborczej w taki sposób, aby został sam tekst artykułu oraz istotne informacje jak tytuł, autor, data, itp. Nie mogą się pojawić elementy menu, teksty do linków, czy opisy do zdjęć. Oczywiście nie mogą się pojawić elementy HTML, czy Javascript lub inne fragmenty kodu.

Wyjście programu powinno być podobne do pliku przykładowego gazeta1.txt. Opracować program tylko na podstawie pliku gazeta1.html nie sugerując się plikiem gazeta2.html

Uwaga: Warto uwzględnić elementy
 w szczególny sposób.

Punkty: 3

Sprawdzić, czy program będzie działał dla pliku gazeta2.html. Jeśli nie, to proszę zmodyfikować program by obsługiwał oba pliki.

Punkty: +2

Zadanie 3

Przetworzyć plik eur-lex.html do postaci eur-lex.txt. Zwrócić szczególną uwagę na zachowanie kolejności odpowiedników.

Punkty: 6