

Курсова работа по

“Извличане на информация в Интернет”

1. Да се създаде програмна система на C++, която да реализира опростена система за търсене в web със следната функционалност:
 - a. Получава съдържанието на уеб-страницата в изходен вид от HTTP сървър.
 - b. Извлича текстовото съдържание (мета-данни: заглавие, ключови думи, описание; съдържание на html-документа: игнорирайки текста в таговете за изображения) в балансирано двоично дърво.
 - c. Извлича от страницата всички линкове към други страници (.htm или .html) и създава списък с линковете.
 - d. Обхожда този списък, обработвайки аналогично всяка една от страниците, като дълбочината на вложеност да не бъде повече от 10.
 - e. Извежда в подходящ графичен вид граф на връзките между страниците.
 - f. Генерира резултат от търсене по въведена от потребителя определена дума – извежда линковете на страниците, в които е открита (по реда на откриване, без рейтинг).
2. Да се поставят в сорса достатъчно коментари, поясняващи предназначението на отделните функции и атрибути.
3. Разработката се предава в папка, със заглавна страница, разпечатана **черно-бяла** обяснителната записка, съдържаща кратко описание на отделните класове, техните променливи и функции, разпечатка на хедърните и сорсови файлове, кратко описание на проведеното тестване- подадените входни файлове и резултатите от тестовете. В найлонов джоб на последна страница се предава CD със записаните файлове на курсовата работа.
4. Защитата на курсовата работа е персонална, с тестване с реален сайт (<http://crawllertest.cs.tu-varna.bg>).