Курсов проект

Craw-ване на kaldata.com и индексиране на статиите

Изработил: Ташос Митусис
Факултетен номер: 8MI3400166
Специалност: PCMT

Подход и инструменти за реализация

Създаване на скрипт на питон, който отваря “robot.txt” и извлича новините от https://www.kaldata.com/news-sitemap.xml за извличане на текущи новини и обикаляне на “ https://www.kaldata.com/телефони/page/1” и другите подобни(it-новини, хардуер, игри, автомобили, софтуер) за намиране на исторически новини в страници. След извличане, текстовете се индексират чрез “TF-IDF” и интерфейс за показване на страници на база избран текст.

Мотивация и формулировка

Индексирането и кеширане на статии за техника и технологии. По този начин могат да се проследят тенденциите в техническият сектор и да се правят предположения за нови трендове. Също така може да се създаде абонаментен план, който да изпраща имейли, при излизане на позитивни/негативни/всички статии за дадена тема.

План за реализация/експерименти

- Написване на crawler
- Запазване на суровите страници
- Изваждане на текста от страниците и запазването ѝм
- Генериране на TF-IDF за всеки документ и запазването ѝм
- Създаване на Web интерфейс за търсене и показване на допълнителна информация