

Конспект

по дисциплината „Извличане на информация в Интернет“ за специалност „Софтуерни и Интернет технологии“

1. Базови принципи на извличане на информацията (ИИ). Структурирани и неструктурирани документи. Подходи за ИИ.
2. Архитектура на търсеща машина. Основни компоненти. Функциониране.
3. Извличане на web страници. Web crawler.
4. Обработване на текст. Tokenizing. Определяне на stop words. Stemming.
5. Парсване на документ. Структура на парсер.
6. HTTP клиент под Windows.
7. Индексиране. Изграждане на индекси. Инвертни индекси.
8. Анализ на връзки. Рейтинговане. Google PageRank.
9. Boolean модел на извличане на информация.
10. Оценка на търсещи машини
11. Архитектура на Google търсеща машина.
12. Google File System.
13. Облачни изчисления.

Литература:

1. Х. Вълчанов, В.Алексиева. Извличане на информация в Интернет. Ръководство за лаб. упражнения. Варна, 2015.
2. Yi Chang, Hongbo Deng. Query Understanding for Search Engines. Springer, 2020, ISBN 978-3-030-58333-0
3. Nicola Ferro, Carol Peters. Information Retrieval Evaluation in a Changing World. Springer, 2019, ISBN 978-3-030-22947-4
4. Tetsuya Sakai, Douglas W. Oard, Noriko Kando. Evaluating Information Retrieval and Access Tasks. Springer, 2021, ISBN 978-981-15-5553-4
5. Introduction to Information Retrieval. <https://nlp.stanford.edu/IR-book/information-retrieval-book.html>
6. Ceri S., A. Bozzon. Web Information Retrieval. Springer-Verlag. 2013.
7. Croft B. Search Engines: Information Retrieval in Practice. <https://ciir.cs.umass.edu/downloads/SEIRiP.pdf>
8. Using search engines. <https://edu.gcfglobal.org/en/internetbasics/using-search-engines/1/>

Лектор: доц,д-р инж. Христо Вълчанов

Формат на изпита: текуща оценка

Две контролни под формата на тест (писмен или в системата за е-бучение на ТУ).

Оценка:

- Точките от упражнения (до 60т.) се събират с точките, получени от двете контролни или поправителен изпит (до 40т.)