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 Crawling. RSS feeds.
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. Колекции за оценяване. Журнали.
11. Оценка на ефективност и ефикасност.
12. Архитектура на Google търсеща машина.
13. Google File System.

Литература:

1. Х. Вълчанов, В.Алексиева. Извличане на информация в Интернет. Ръководство за лаб. упражнения. Варна, 2015.
2. Baeza R., B. Riberto. Modern Retrieval: The Concepts and Technology behind Search. 2nd ed. Addison-Wesley, 2011.
3. Bopp, R.E. & Smith, L.C. (eds), Reference and information services: An introduction, 4th ed, Libraries Unlimited, Santa Barbara, Calif., 2011.
4. Ceri S., A. Bozzon. Web Information Retrieval. Springer-Verlag. 2013.
5. Croft B. Search Engines: Information Retrieval in Practice. Pearson, 2010.
6. Hedden, H. The Accidental Taxonomist, Information Today, Medford, NJ, 2010.
7. Morville, P. & Callender, J., Search patterns, O'Reilly, Sebastopol CA, 2010.
8. Search Engine Land. http://searchengineland.com/.

Лектор: доц.д-р инж. Христо Вълчанов

Формат на изпита:

писмен - 90 минути, последван от устно препитване

Изпитният вариант се състои от отворени въпроси (с кратък отговор) и задачи. За всеки от тях са посочени максималния брой точки, които носят в крайната оценка.

Оценка:

- Точките от текущ контрол (до 40т.) се събират с точките, получени от изпита (до 60т.)