Information Retrieval and Extraction

11.8.2020

Course Logistics

- 3 major parts:
  - 4 sessions: introduction from multiple aspects
  - 9 " : IR fundamentals:
    - models
    - scoring functions
    - crawling, evaluation, etc.

    NLP required specifically for IR
    ML for IR
  - : IE fundamentals
    - majorly: NER
  - : Information Access / Applications of IR
    - moving specifics: social media, sentiments,
    - computational advertising
    - sentiment analysis.

- Tutorials: Thu / Tue evening

- Recommended Books
  - Stanford IR book
  - Search Engines in Practice

- Grading:

Quizzes / In-class:		10%
Assignments (best 3/4)		15%
Project		60%
├ mini	20	
└ major	40	
Term Paper (Final Exam, can be done anytime)		15%

_____ × _____ × _____

- Project Details
  - MINI
    - Individual
    - 4 weeks, starting today
    - Deliverables:
      1.
      2.

Lang:
Py / C++ / Java

DEADLINES
1. 24th August: offline
2. 7th Sept: online + offline

Design and develop a scalable and efficient search engine on wikipedia.

REQUIREMENTS
- Query (1s) → result
- Support "Field Queries"[1]
- Total index size: < 1/4th the size of the doc repository
- build your own indexing scheme

EVAL
- online:
  - search time
  - search efficiency
- offline:
  - Indexing time
  - Indexing size

26th: Dummy Queries

1. Field Query: Searching within subset.

— MAJOR

- Team of 4 constrained choice
- 10 weeks.

Advanced topics
Scope well defined (by us)

- 5 touchpoints
  Report to mentor every 2 weeks
- 3 Evaluations
  - first deliverable: (26th Sept) — scope doc
  - MVP deliverable: (25th Oct) — full system, v1
  - Complete system: (14th Nov) — Demo, presentation report code, etc.

Basic Overview

- Searching is important
- Computational advertising