Information Retrieval Exercises

Introduction

Patrick Schäfer (patrick.schaefer@hu-berlin.de)

Marc Bux (buxmarcn@informatik.hu-berlin.de)

Idea

- we will build groups of two students each
- each group has to solve 5 exercises
  - all exercises must be solved by all groups
- to solve each exercise, you'll get 2-3 weeks
- "solving" means: a Java implementation or some configuration / extension of Lucene
- partial solutions are presented by groups in the form of 5-10 minutes talks
  - each group has to present roughly 2 times in total
  - you'll be able to pick when and what you'd like to present (first-come-first-served)
  - if you don't decide yourself, you'll be assigned a presentation slot

Tentative Schedule

- today: group formation; assignment 1
  - Crawl IMDB, build an index, and implement a set of queries
- 21/22 Nov: presentation of solutions; assignment 2
  - Implement fast Boolean search in a 100 MB corpus
- 12/13 Dec: presentation of solutions, assignment 3
  - Use Lucene for Boolean information retrieval
- 3/9 Jan: presentation of solutions, assignment 4
  - Implement synonym expansion with Lucene
- 23/24 Jan: presentation of solutions, assignment 5
  - Find significant co-occurrences
- 13/14 Feb: presentation of solutions, final ceremony
- In between: Q/A sessions

Challenge - Voluntary

- most assignments can be solved more or the less well
- “good” usually means: very fast

- the best groups in each assignment will get points
  - best: 5 points, second: 3 points; third: 2 point; fourth-fifth: 1 point
- the overall best group will get a little present at the end of the semester

Additional Information

- group formation via E-Mail
  - notify us of your group (including a self-chosen group name) until / while you hand in the first assignment
  - we don‘t accept submissions by groupless individuals
  - we provide a Google Sheet for group formation: https://goo.gl/NhuV4G
- questions?