Lecture 1 - Introduction to the course

Sai Amrit Patnaik

August 2020

Introduction to IR and IE

Vasudeva Varma

Personal Notes to Introduction to Information Retrieval and Information Extraction course taught in Monsoon 2020 by Dr.Vasudeva Varma, Manish Gupta and Niyati Chaya.

1 Course Structure

- Slides Available at : **Lesson 1, Slides**.
- **Introduction (4 lectures)**
- **IR Fundamentals (9 lectures) :** Models, Scoring functions, Index design, Crawling, IR Evaluation

- **NLP/Text Mining for IR (4 lectures)**
- **Machine Learning & IR (9 lectures)**
- **Information Extraction (3 lectures)** : IE Fundamentals, Named Entity Recognition
- **Information Access and IR Applications (9 lectures)** : Summarization, Social Computing

2 Evaluation

- **Quiz/In Class Activities : 10%**
- **Assignments : 15%**
- **Project : 60% (20% Mini Project + 40% Major Project)**
- **Term Paper : 15% (Can be given at any time of the semester)**

3 Textbooks and Reference Books

- Introduction Information Retrieval – Chris Manning et al (the Stanford IR Book)
- Search Engines IR in Practice – Bruce Craft et. al

4 Minor Project details

- Individual Project (4 Weeks)
- Two deliverables
- You can use compression techniques
- Explore several ranking functions (tf, tf-idf, normalized tf, normalized idf etc)
- Create a secondary index if required.

4.1 Objective

Design and develop a scalable and efficient search engine using the Wikipedia data.

4.2 Features

- Dump of Wikipedia as document repository
- Results obtained in less than a sec (even for long queries)
- Supports field queries (ex: a particular word in title section or any other section)
- Index size should be less than ${\frac{1}{4}}^{th}$ of the data size.
- You have to build your own indexing mechanism.

4.3 Evaluation Criteria

Evaluation will be done on **4 major criteria**:

- **Search time**
- **Search efficiency**
- **Indexing time**
- **Index Size**

4.4 DEADLINES

- First evaluation (29^{th} **August**) :
  - Indexing time and Indexing efficiency will be evaluated.
  - Dummy queries will be provided before 26^{th} August
- Second (Final) evaluation (7^{th} **September**) - All four parameters will be considered.

5 Major Project details

- Group of 4 people (10 Weeks)
- 5 touch points, 3 deliverables
- report progress every 2 weeks.
- Three deliverables :
  - Scope document by **26th September** - Defining scope of what is to be worked on.
  - End-to-end system by **25th October** – Most Viable Product deliverable
  - Complete system by **14th November** - Demo/presentation Video, Code, Report