Text Technologies for Data Science

INFR11145

Introduction

Instructor:
Youssef Al Hariri

20/09/2023
Lecture Objectives

• Know about the course:
  • Topic
  • Objectives
  • Requirements
  • Format
  • Logistics

• Note:
  • No much technical content today
  • Don’t assume next lectures would be the same!
Text Technologies for Data Science

= documents, words, terms, …
≠ images, videos, music (with no text)

Information Retrieval
Text Classification
Text Analytics

Search Engines
Technologies
What is Information Retrieval (IR)?

IR is NOT just Web search
What is IR?

Speech - QA
What is IR?

Social search

Information Filtering

Recommendation

Youssef Al Hariri, TTDS 2023/2024
What is IR?

Library (book) search

1950’s
What is IR?

Legal search
What is IR?

Cross-Language search
What is IR?

Content-based music search
What is IR?

*Source: Matt Lease (IR Course at U Texas)*

---

**Query suggestion / correction**

![Image of Google search suggestions for Edinburgh](image_url)
What is IR?

- Snippet selection / summarisation
- Query suggestion
- Categorisation (search verticals)

*Source: Matt Lease (IR Course at U Texas)
What is IR?

*Source: Matt Lease (IR Course at U Texas)*

Advertising
What is IR? Find?

IR ≠ Find

- Sequential
- Exact match
What is IR?

- IR is finding material of an unstructured nature that satisfies an information need from within large collections.

- Find → Task
- Unstructured → Nature
- Information need → Target
- Satisfies → Evaluation
Text classification

Second man held
Text classification

James, me (2)

Hiking trip on Saturday - Yay - so glad you can join. We should leave from I
3:14 pm

Hannah Cho

Thank you - Keri - so good that you and Steve were able to come over. Thank you :
3:05 pm

Jay Birdsong

Upcoming school conference dates. Hello everyone. A few people have
12:26 pm
Text classification

(12) United States Patent
Magdy et al.

(54) PERSONALIZED EVENT NOTIFICATION USING REAL-TIME VIDEO ANALYSIS

(75) Inventors: Walid Magdy, Giza (EG); Motaz El-Saban, Giza (EG)

(73) Assignee: Microsoft Corporation, Redmond, WA (US)

(*) Notice: Subject to any disclaimer. The term of this

(10) Patent No.: US 8,881,191 B2
(45) Date of Patent: Nov. 4, 2014

(51) Int. Cl.
H04H 60/65 (2008.01)
H04H 60/48 (2008.01)
G06F 17/30 (2006.01)

(52) U.S. Cl.
CPC ............... H04H 60/48 (2013.01); H04H 60/65 (2013.01); G06F 17/30787 (2013.01); G06F 17/30831 (2013.01)
USPC ............... 725/32; 725/43; 725/52; 382/181; 348/460
What is text classification?

- **Text classification** is the process of **classifying** documents into **predefined categories** based on their **content**.

  - Input: Text (document, article, sentence)
  - Task: Classify into one/multiple categories
  - Categories:
    - Binary: relevant/irrelevant, spam .. etc.
    - Few: sports/politics/comedy/technology
    - Hierarchical: patents
In this course, we will learn to

• How to build a search engine
  • which search results to rank at the top
  • how to do it fast and on a massive scale

• How to evaluate a search algorithm
  • is system A really better than system B

• How to work with text
  • two tweets talk about the same topic?
  • handle misspellings, morphology, synonyms

• How to classify text
  • into categories (sports, news, comedy, …)
  • features to use
  • evaluate classification quality

• Apply text analytics
  • Find what makes a set of document different from others
How this course is different from others?

• ANLP, FNLP
  • Some text processing
  • Text laws
  • No NLP (word/phrase level vs document level)

• ML practical
  • Text classification
  • No ML (using off-the-shelf ML tool)

• It does not overlap with others on:
  • Search engines
  • IR methods/models
  • IR evaluation
  • Text analysis
  • Processing large amount of textual data
Some terms you will learn about

- Inverted index
- Vector space model
- Retrieval models: TFIDF, BM25, LM
- Page rank
- Learning to rank (L2R)
- MAP, MRR, nDCG
- Mutual information, information gain, Chi-square
- binary/multiclass classification, ranking, regression
This Course is Highly Practical

• 70% of the mark is on practical work
• You will implement **50+%** of what you learn
• By W5, you should have developed a basic working Search Engine from scratch
• Practical Lab every week
• Two coursework, mostly coding
• A course group project to develop a full system
Pre-requests (1/3)

• Maths requirements:
  • Linear algebra: vectors/matrices (addition, multiplication, inverse, projections ... etc).
  • Calculus: Functions of several variables. Partial differentiation. Multivariate maxima and minima.
  • Special functions: Log, Exp, Ln.

\[
BM25(D, Q) = \sum_{i=1}^{n} \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \cdot \left[ \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)} + \delta \right]
\]
Pre-requests (2/3)

- Programming requirements:
  - **Python**
  - Knowledge in regular expressions
  - Shell commands (cat, sort, grep, uniq, sed, ...)
  - Data structures and software engineering for course project.

- We **DO NOT** teach coding skills in this course! We assume you can code!

```bash
\b[\w.-%+-]+@[\w.-]+\.[a-zA-Z]{2,6}\b
```

Parse: username@domain.TLD (top level domain)
Pre-requests (3/3)

• Team-work requirement:
  • Final course project would be in groups of 5-6 students.
  • Working in a team for the project is a requirement.
  • No exceptions will be allowed!
Skills to be gained !!!

• Working with large text collections
• Few shell commands
• Some Python programming
• Software engineering skills
• Build text classifier in few mins
• TEAM WORK
  • Project management
  • Time management
  • Task assignment + system integration
Course Structure

• 20 Lectures:
  • 2 lectures → Introduction (today)
  • 14 lectures → IR (50% practical lectures)
  • 4 lectures → Text Analytics/Classification

• 8-10 Labs:
  • Practice what you learn

• No Tutorials
• Some self-reading
• Lots of system implementation
• Few online videos
Course Instructors

Youssef Al Hariri
Lecturer
(15 lectures)

Bjorn Ross
Lecturer
(4 lectures)

+ 1 guest lecture
Lecture Format

• 2 Lectures at a time

• Questions are allowed any time. Feel free to interrupt

• 5-10 mins break after L1
  • Feel free to go out and come back
  • Discuss 1st lecture with friends
  • Questions on L1 are allowed before starting L2

• Some lectures are interactive. Please participate

• Some lectures will include demos (running code)
Labs

• How it works:
  • Relevant lab will be announced with each lecture on Wednesday
  • You should implement lab directly after lecture
  • Any issues → ask on Piazza (tag question by lab number)
  • Produced output → Share on Piazza (publicly)
  • Demonstrators → answer questions + validate your output
  • TA → answer questions about the course
  • DO NOT ask a question before checking if it was asked before
  • Tuesdays → Optional in-person labs for those still require support

• Optional in-person labs:
  • Location: AT 6.06 (TBC)
  • Times: Tuesday, 10:00, 11:00, 13:00 (TBC)

• Demonstrators:
  TBA
Lab Zero (Lab 0)

- Please check Lab 0 before next week lectures
- Lab 0 is designed for one purpose: Help you decide to take TTDS or not
- Lab content:
  - Read a text file word by word, lower-case letters, print
  - Count the number of occurrence of few words
- If Lab 0 challenging →
  → Probably, TTDS would be very challenging to you
  → You will need much extra effort to implement labs and CW
  → Think wisely before you decide to take the course
Assessments

• Coursework 1: **10%**
  The same as labs 1-3 → Build your first search engine

• Coursework 2: **20%**
  IR Evaluation, Text classification/analytics

• Group project: **40%**
  A full running search engine supported by text technologies

• Final Exam: **30%**
Group Project

• The largest weight: 40% of the total mark

• Teamwork \rightarrow \text{Group 5-6 (you select your own group)}

• Design a full end-to-end search engine that searches a large collection of documents with many functionalities.

• Mark = Mark_{\text{project}} \times \text{weight}_{\text{individual}}
  - Mark_{\text{project}} \rightarrow \text{the same for all team members}
    - How complete/effective/fast/nice is your search engine?
  - \text{weight}_{\text{individual}} \rightarrow \text{weight for individual contribution.}
    - ranges from 0 to 1. It should be 1.0 by default but can be different for each member according to their contribution.

• Project prize \rightarrow \text{a prize will be awarded to best project}
Example: BetterReads

BetterReads

Filters
Change the filters to refine your search.

<table>
<thead>
<tr>
<th>Filter</th>
<th>Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating</td>
<td>0 - 5</td>
</tr>
<tr>
<td>Year of Release</td>
<td>1950 - 2021</td>
</tr>
<tr>
<td>Weighting</td>
<td></td>
</tr>
</tbody>
</table>

BetterReads
Version 1.0.5

1. Cowsex
   2017, Lesley Jones
   Completely laugh out loud funny!!! This book is by far my favorite Lesley Jones book ever!

2. The Idiot Girls' Action-Adventure Club
   2002, Laurie Notaro
   I love Laurie Notaro!! I read this book about 6 years ago and LOVED it! I need to make sure I read her other books, too! Thely laugh out loud funny!!

3. Anyone But You
   1995, Jennifer Crusie
   This is a wonderful story! I laughed out loud so many times! I also kept catching myself just smiling while reading it!
Example: BetterReads

- 11.5M Book reviews from Good reads
- Average query time: 1 secs
- New reviews are crawled and indexed automatically every day
- Ranking: Relevance + Sentiment
- Engine hosted on Google cloud compute

• Note: we will provide credit to Google cloud to host your engine
Timeline

- 2 Semesters (or one?)

Semester 1:
- W5: CW 1 & 2

Semester 2:
- W9: Group Project
- W11: Labs

Exam
Logistics

• Lectures:
  • Two lectures on Wednesdays, 15.00-17.00
  • Recording will be available
  • Handouts to be posted on the day of the lecture

• Course webpage:
  • Link: https://opencourse.inf.ed.ac.uk/ttds/
  • Handouts, Labs, CW details

• Learn:
  • Lecture recordings
  • Deadlines

• Note: all course materials are made public, including recordings. Feel free to share with anyone interested.
Piazza

- All communication will be there
- Questions about lectures/labs/CW are there
- Feel free to answer each other questions
- Lab support will be mainly there
- Please share your lab answers there
- Tag each question/post by its relevant topic (lab, CW … etc)

- Join NOW: link
FAQ

• How the project would be managed? What if one member does not work?
• I am not that solid in programming, should I take this course?
• Can I audit the course?
• Anything else?
Next Lecture

• Definitions of IR main concepts (more introduction)
Credits

• These slides are originally created by previous lecturers Walid Magdy and Bjorn Ross.