Text Technologies for Data Science
INFR11145

Introduction

Instructor:
Walid Magdy

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

Google

Web search
What is IR?

Speech - QA

Social search

Information Filtering

Recommendation

Walid Magdy, TTDS 2021/2022
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?

IR ≠ Find
- Sequential
- Exact match

IR ≠ Find
- Query suggestion / correction
- Snippet selection / summarisation
- Categorisation (search verticals)

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

Walid Magdy, TTDS 2021/2022

United States Patent
Magdy et al.

Personalized event notification using real-time video analysis

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

Assignee: Microsoft Corporation, Redmond, WA (US)

Notice: Subiect to anv disclaimer. the term of this

Int. Cl.
II04H 60/65 (2008.01)
II04H 60/48 (2008.01)
G06F 17/30 (2006.01)

U.S. Cl.
II04H 60/48 (2013.01); II04H 60/65 (2013.01); G06F 17/307787 (2013.01); G06F 17/30831 (2013.01)

USPC 725/32; 725/43; 725/52; 382/181; 348/460

US8881191B2

Walid Magdy, TTDS 2021/2022
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.

$BM25(D, Q) = \sum_{i=1}^{n} \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \cdot \left[ \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{avgDl}\right)} + \delta \right]$
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.

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

character set {...} (match one out of several)

special characters

At symbol

alpha-num, dot, or dash char

dot

upper or lower alpha character

any alpha-numeric char,

match previous {...} pattern at least one time

the {x,y} modifier means that the previous pattern must have 2-6 characters

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

Walid Magdy
Reader
(14 lectures)

Bjorn Ross
Lecturer
(5 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
  • Mind teaser math problem (for fun)
• Some lectures are interactive. Please participate
• Some lectures will include demos (running code)
Labs

• Online!

• How it will work?
  • 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
  • DO NOT ask a question before checking if it was asked before
  • Tuesdays → Optional Teams meetings for those still require support

• Live lab times: Tuesday, 9am, 11am, 6pm

• Demonstrators:
  Zheng Zhao, Ibrahim Abu Farha, and Youssef Al Hariri

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

• If you Lab 0 challenging then:
  → 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 → Group 5-6 (you select your group)
- Design a full end-to-end search engine that searches a large collection of documents with many functionalities.

Mark = Mark_{project} \times weight_{individual}

- Mark_{project} → the same for all team members
  - How complete/effective/fast/nice is your search engine?
- weight_{individual} → 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 → a prize will be awarded to best project
Example: BetterReads

- 11.5M Book reviews from Good reads
- Average query time: 1.3 secs
- New reviews are crawled and indexed automatically every month
- 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
- Lectures
- Labs
- CW 1 & 2

**Semester 2**
- W9
- W11
- Exam
- Group Project

---

**Logistics (1/2)**

- Lectures:
  - Live on 2 Wednesdays, 12.00-14.30 (some exception might occur)
  - Recording will be available
  - Handouts to be posted on the day of the lecture

- Course webpage:
  - Link: [http://www.inf.ed.ac.uk/teaching/courses/tts/](http://www.inf.ed.ac.uk/teaching/courses/tts/)
  - Handouts, Labs, CW details, link to recordings

- Learn:
  - Lecture recordings
  - Deadlines

- Note: all course materials are made public including recordings. Feel free to share with anyone interested
Logistics (2/2)

- Pizza:
  - 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
  - Join NOW: link

- Microsoft Teams:
  - Live lab support will be there
  - 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)