Text Technologies for Data Science
INFR11145

Introduction

Instructor:
Walid Magdy

Lecture Objectives

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

What is IR?

Information Filtering

Recommendation

Social search
What is IR?

Library (book) search
1950's

What is IR?

Legal search
What is IR?

Cross-Language search

Content-based music search
What is IR? Find?

IR ≠ Find
- Sequential
- Exact match

*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 $\rightarrow$ Task
  - Unstructured $\rightarrow$ Nature
  - Information need $\rightarrow$ Target
  - Satisfies $\rightarrow$ Evaluation

---

**Text classification**

- BBC
- Sign in
- Home, News, Sport, Weather, iPlayer

**NEWS**

**Second man held**

- Search Twitter
- Today, News, Sports, Entertainment, Fun

---

Walid Magdy, TTDS 2019/2020
Text classification

1. Personalized Event Notification Using Real-Time Video Analysis
2. Assignee: Microsoft Corporation, Redmond, WA (US)
3. Notice: Subject to any disclaimer. The term of this

United States Patent
Magdy et al.

Patent No.: US 8,881,191 B2
Date of Patent: Nov. 4, 2014

Int. Cl. IH04H 60/65 (2008.01) IH04H 60/48 (2008.01) G06F 17/30 (2006.01)
U.S. Cl. IH04H 60/48 (2013.01); IH04H 60/65 (2013.01); G06F 17/30787 (2013.01); G06F 17/30831 (2013.01)
CPC ............... 725/32; 725/43; 725/52; 382/18; 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
  - Detect important and distinctive words
This course overlaps a bit with

- 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

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
- SVMs: binary/multiclass classification, ranking, regression
This Course is Highly Practical

• You will implement 50+% of what you learn
• By W5, you should have developed a fully working Search Engine from scratch
• Practical Lab every week
• Two coursework, mostly coding
• A course group project to develop a full system

Skills to be gained !!!

• Working with large text collections
• Few shell commands
• Some Perl/Python programming
• Software engineering skills
• IR tools: Solr
• Build text classifier in few mins
• TEAM WORK
Course Structure

• 18 Lectures:
  • 2 lectures -> Introduction (today)
  • 10 lectures -> IR (50% practical lectures)
  • 2 lectures -> Applications
  • 2 lectures -> Text Classification

• 8-10 Labs:
  • Practice what you learn

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

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)
• 2-4 tutorial lectures about using tools
Assessments

- Coursework 1: 10%
- Coursework 2: 10%
- Group project: 30%
- Final Exam: 50%

- \[ Mark_{\text{project}} = 0.5 \ Mark_{\text{team}} + 0.5 \ Mark_{\text{individual}} \]

Timeline

- 2 Semesters (or one?)

  - Semester 1:
    - W5: Lectures
    - W9: Labs
    - CW 1 & 2
  - Semester 2:
    - W4: Group Project
    - Exam
Pre-requests (1/3)

- Maths requirements:
  - Linear algebra: vectors/matrices (addition, multiplication, inverse, projections ... etc).
  - 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 and/or Perl, and good knowledge in regular expressions
  - Shell commands (cat, sort, grep, uniq, sed, ...)
  - Additional programming language could be useful for course project.

.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.

Logistics (1/2)

- Lectures:
  - 2 Lectures on the same day (10 mins break in-between)
  - Wednesdays, 15.00-17.30
  - Location: F.21, 7 George Square

- Practical labs → Location: AT 6.06
  - Option 1: Monday, 9.00-10.00
  - Option 2: Monday, 11.00-12.00
  - Option 3: Tuesday, 11.00-12.00
  - You need to signup for one of those labs

- Demonstrators:
  - Abeer Aldayel, Silviu Oprea, Zheng Zhao
Logistics (2/2)

- Course webpage: [http://www.inf.ed.ac.uk/teaching/courses/tts/](http://www.inf.ed.ac.uk/teaching/courses/tts/)
- Material: TBA with each lecture
- Assignments/Project:
  - W5, W9, and S2 W4
  - Details to be announced
- Textbooks:
  - “Introduction to Information Retrieval”. Manning et al.
  - “Search Engines: Information Retrieval in Practice”
- Discussion forum: [https://piazza.com/class/k0hc6o94ljp7mc](https://piazza.com/class/k0hc6o94ljp7mc)
- Sign up for Labs on [Piazza](https://piazza.com)

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)