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

Social search

Information Filtering

Recommendation

Walid Magdy, TTDS 2018/2019
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?

Query suggestion / correction

Snippet selection / summarisation

Categorisation (search verticals)

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**

[Image of a BBC News page with a headline: Second man held]
Text classification

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: Subject to any disclaimer. The term of this

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

US Pat. 8,881,191 B2

Date of Patent: Nov. 4, 2014

THE UNIVERSITY of EDINBURGH
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
- How to evaluate a search algorithm
- How to work with text
- How to classify text
- Apply text analytics
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
Skills to be gained !!!

- Working with large text collections
- Few shell commands
- Some Perl/Python programming
- IR tools: Lemur / Indri / 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
- Some lectures are interactive. Please participate
- Some lectures will include demos (running code)
- 1 or 2 optional tutorial lectures about using tools

Assessments

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

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

- 2 Semesters (or one?)

Pre-requests (1/3)

- Maths requirements:
  - Linear algebra: vectors/matrices (addition, multiplication, inverse, projections ... etc).
  - Special functions: Log, Exp, Ln.

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

```
\b\w\.+[@]?\w\.-+\.
```

Parse: username@domain.TLD (top level domain)

Pre-requests (3/3)

• Team-work requirement:
  • Final course project would be in groups of 4-6 students. Working in a team for the project is a requirement.
Logistics (1/2)

- Course webpage: [http://www.inf.ed.ac.uk/teaching/courses/tts/](http://www.inf.ed.ac.uk/teaching/courses/tts/)
- Lectures:
  - 2 Lectures on the same day (10 mins break in-between)
  - Wednesdays, 15.00-17.30
  - 2.13, Geography (Old Infirmary), Central (might change)
- Practical labs:
  - Option 1: Thursdays, AT 6.06, 16.00-17.00
  - Option 2: Thursdays, AT 6.06, 17.00-18.00
- Demonstrators:
  - Lushi Chen
  - Abeer Aldayel

Logistics (2/2)

- 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/jksc9tw0xiz15l](https://piazza.com/class/jksc9tw0xiz15l)