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
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?

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

IR ≠ Find
• Sequential
• Exact match

What is IR? Find?

Find
Retrieved
Previous Next

After scoring all terms, the top expansion terms, are added to the top-drift, the topic vector is reset to the vector) before each expansion, as:

\[
\tilde{q} = q + \beta \delta
\]

where \( \delta \) is the normalized vector of the \( k \) expansion terms, and \( \beta \) is a parameter used to restrict the influence of expansion terms on the new topic vector.

2.3 Periodic E-mail Digest Scenario

In this scenario, the RTS system is required to compile a daily list of a maximum of \( m \) tweets per interest profile and send it as an email digest to the user. For that, we adopted a similar but even simpler approach than the approach for push notification scenario. At the end of each day of the evaluation period, the system issues the title of the interest profile against the local tweet index that is incrementally updated over time. We experimented with three versions:

https://dev.twitter.com/rest/public/search
**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

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)

US088191B2
US 8,881,191 B2
Nov. 4, 2014

(12) United States Patent
Magdy et al.

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

(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)

(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 17/30/787 (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

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
- SVMs: 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).
  - Special functions: Log, Exp, Ln.
**Pre-requests (2/3)**

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

```regex
\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

Walid Magdy
Reader
(14 lectures)

Steve Wilson
Research Associate
(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)
• 2 tutorial lectures about using tools
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, 10am, 12am, 6pm
- Demonstrators: Marina Posti, Maysara Hammouda and Zheng Zhao

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 $\rightarrow$ 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:
  - 50% on project $\rightarrow$ the same for all team members
    - How complete/effective/fast/nice is your search engine?
  - 50% on individual contribution $\rightarrow$ different for each member
    - How useful/much is your contribution? (Mark can be negative!)
- Project prize $\rightarrow$ a prize will be awarded to best project

---

**Example**

- A search engine that retrieves quotes of movies and TV shows.
- Collection size: 77 million movie quotes
- Search options
  - Phrase search of quotes
  - Movie info search
  - Advanced search: movie title, actor/actress, years, keywords
- Query suggestion
- Classifying results by genre
- Demo: [http://167.71.139.222/](http://167.71.139.222/)
**Timeline**

- 2 Semesters (or one?)

**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

- Learn:
  - Lecture recordings
  - Deadlines
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)