



THE UNIVERSITY  
of EDINBURGH

# Text Technologies for Data Science

INFR11145

## Introduction

Instructor:  
**Walid Magdy**

19-Sep-2017

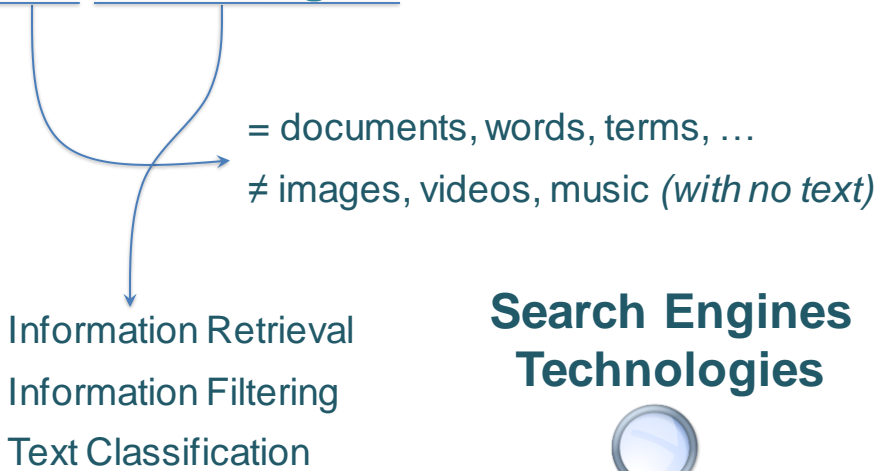
## Lecture Objectives

- Know about the course:
  - Topic
  - Objectives
  - Format
  - Requirements
  - Logistics
- Have a decision on the course
  - Stay
  - Run away



THE UNIVERSITY  
of EDINBURGH

## Text Technologies for Data Science



## Search Engines Technologies



Walid Magdy, TTDS 2017/2018



THE UNIVERSITY  
of EDINBURGH

## What is Information Retrieval (IR)?

IR is **NOT** just

Google

Google Search

I'm Feeling Lucky

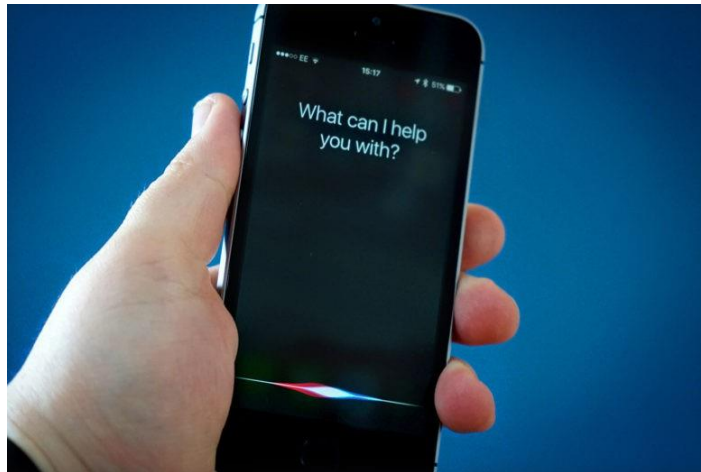
Web search

Walid Magdy, TTDS 2017/2018



THE UNIVERSITY  
of EDINBURGH

## What is IR?



### Speech - QA

Walid Magdy, TTDS 2017/2018



## What is IR?

The screenshot shows a Twitter search results page for the hashtag #Harvey. The page includes a search bar at the top, navigation tabs (Home, Moments, Notifications, Messages), and a search filter section. The 'Who to follow' section is circled in red, with an arrow pointing to it from the word 'Recommendation'. The tweet by HSSawakening is also circled in red, with an arrow pointing to it from the words 'Information Filtering'. The tweet text is: 'Close the loopholes, stop tax evasion and use corporations fair tax share to care for our citizens after #Harvey and #hurricaneirma2017'. Below the tweet is a photo of people in a flooded area.

### Social search

Walid Magdy, TTDS 2017/2018



## What is IR?



### Library (book) search 1950's

Walid Magdy, TTDS 2017/2018



## What is IR?

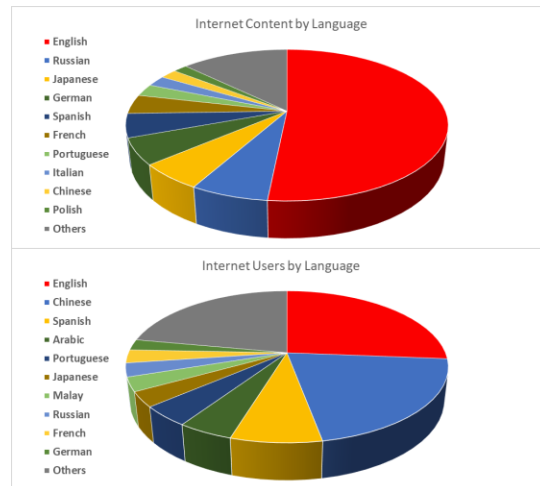


### Legal search

Walid Magdy, TTDS 2017/2018



## What is IR?



## Cross-Language search

Walid Magdy, TTDS 2017/2018



## What is IR?



## Content-based music search

Walid Magdy, TTDS 2017/2018





## 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

Walid Magdy, TTDS 2017/2018



## Text classification

The screenshot shows the BBC News website interface. At the top, there's a navigation bar with the BBC logo, a 'Sign in' button, and links to Home, News, Sport, Weather, and iPlayer. Below this is a large red banner with the word 'NEWS' in white. Under the banner, there's a horizontal menu with categories: Home, UK, World, Business, Politics, Tech, Science, Health, Education, and Entertainment & Arts. The 'UK' category is highlighted with a yellow box. Below the menu, there's a sub-menu for regions: England, N. Ireland, Scotland, Alba, Wales, and Cymru. The main content area features a headline 'Second man held' with a corresponding image. At the bottom, there's a navigation bar with icons for Home, Moments, Notifications, Messages, and a Twitter logo, along with a 'Search Twitter' button. Below this bar, there's another set of tabs: Today, News, Sports, Entertainment, and Fun. The 'Today' tab is highlighted with a red box.

Walid Magdy, TTDS 2017/2018



## Text classification


1-21 of 21

Primary Social 1 new Google+ Promotions 2 new Google Offers, Zagat Updates 1 new Google Play

James, me (2) **Hiking** Hiking trip on Saturday - Yay - so glad you can join. We should leave from l 3:14 pm

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


James, me (2) **School** Upcoming school conference dates. Hello everyone. A few people have 4:06 pm



Walid Magdy, TTDS 2017/2018

THE UNIVERSITY of EDINBURGH

## Text classification



US00881191B2

(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)

(\*) Notice: Subject to any disclaimer, the term of this

(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

Walid Magdy, TTDS 2017/2018

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

- 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 relevant/non-relevant (filtering)
  - into categories (sports, news, comedy, ...)
  - features to use
  - classification methods to use

## 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

Walid Magdy, TTDS 2017/2018



## 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

Walid Magdy, TTDS 2017/2018



## Skills to be gained

- Working with large text collections
- Few shell commands
- Some Perl programming
- IR tools: Lemur / Indri / Solr
- Crawling: Web / Tweets
- TEAM WORK

Walid Magdy, TTDS 2017/2018



## Course Structure

- 18 Lectures:
  - 2 lectures → Introduction
  - 10 lectures → IR
  - 2 lectures → Applications
  - 2 lectures → Text Classification
- 8-10 Labs:
  - Practice what you learn
- No Tutorials
- Much self-reading + system implementation
- Few online videos

Walid Magdy, TTDS 2017/2018



## Assessments

- Assignment 1: 10%
- Assignment 2: 10%
- Group project: 20%
- Final Exam: 60%
- $Mark_{project} = 0.5 Mark_{team} + 0.5 Mark_{individual}$
- Final exam: 2<sup>nd</sup> semester

## Good to know

- Course has not been taught for two years
- New instructor
- Updated content
- Updated credit (previously 10, now 20) → 200 hours
- Additional credit is directed mainly to practical work
- Previous versions average mark: 52%

## Pre-requests (1/3)

- Maths requirements:
  - Linear algebra: vectors/matrices (addition, multiplication, inverse, projections ... etc).
  - Probability theory: Discrete and continuous univariate random variables. Bayes rule. Expectation, variance. Univariate Gaussian distribution.
  - Calculus: Functions of several variables. Partial differentiation. Multivariate maxima and minima.
  - Special functions: Log, Exp, Ln.

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

Walid Magdy, TTDS 2017/2018



THE UNIVERSITY  
of EDINBURGH

## Pre-requests (2/3)

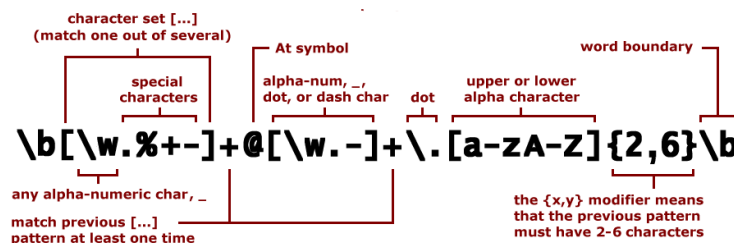


Perl



python

- 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)**

Walid Magdy, TTDS 2017/2018



THE UNIVERSITY  
of EDINBURGH

## 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.



Walid Magdy, TTDS 2017/2018



## Logistics (1/2)

- Course webpage: <http://www.inf.ed.ac.uk/teaching/courses/tts/>
- Lectures:
  - 2 Lectures on the same day (30mins break in-between)
  - Tuesdays, 14.00-16.30
  - Lecture Theatre A, David Hume Tower
- Practical labs:
  - Thursdays, 12.10-13.00 & 16.10-17.00
  - Room 6.06, Appleton Tower
- Demonstrator: Clara Vania

Walid Magdy, TTDS 2017/2018



## Logistics (2/2)

- Material: TBA with each lecture
- Assignments/Project: TBA
- Textbooks:
  - “Introduction to Information Retrieval”. Manning et al.
  - “Search Engines: Information Retrieval in Practice”
- Discussion forum: <https://piazza.com/class/j766gisdu46m>

## Questions

- Next lecture:  
Definitions of IR main concepts  
(more introduction)

**Break  
&  
continue in 30 mins**

