CORSO DI
WEB MINING E RETRIEVAL
- INTRODUZIONE AL WM -

Corso di Laurea in Informatica, Ing. Internet, Ing. Informatica, Ing. Gestionale

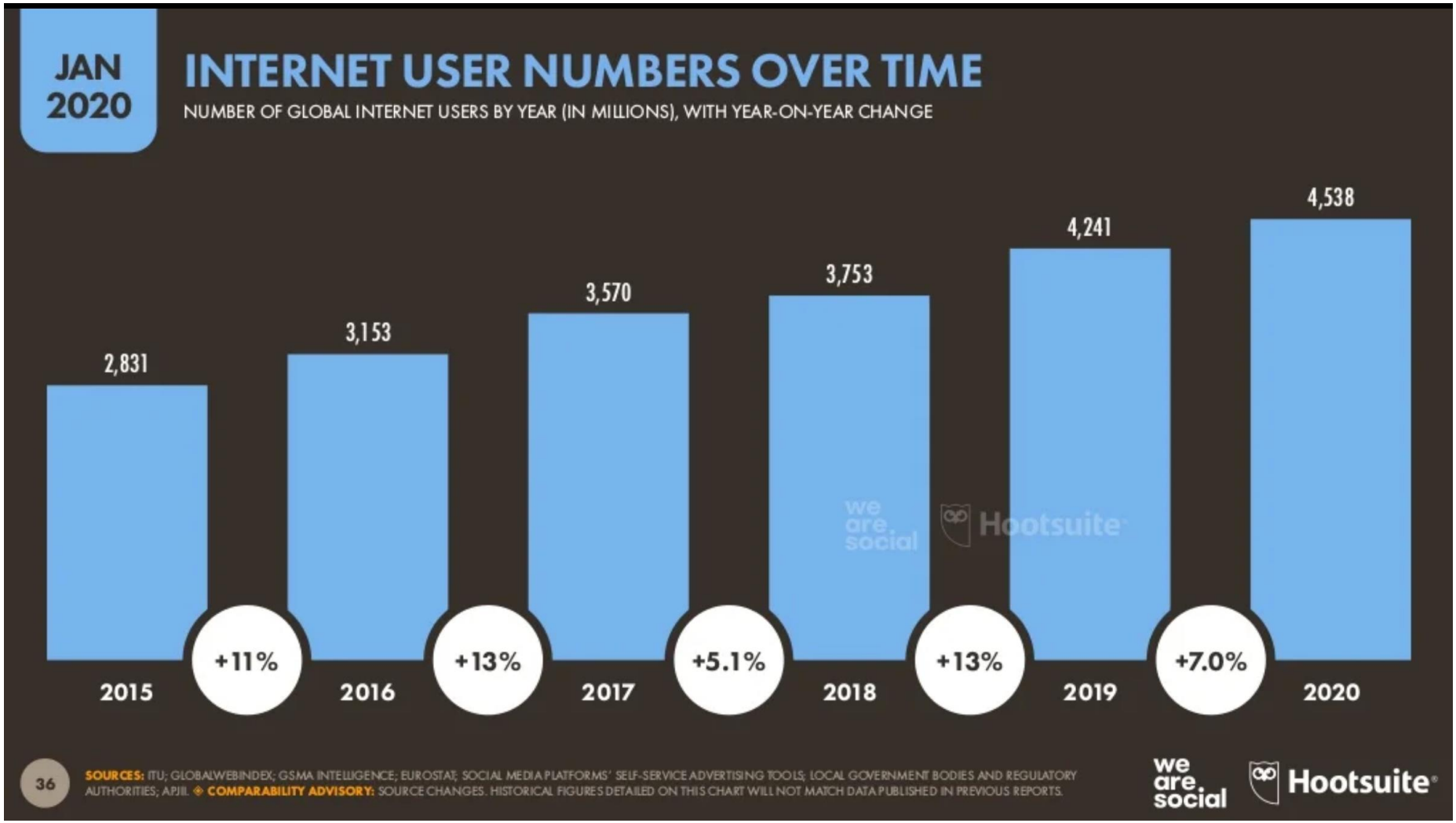
(a.a. 2021-2022)

Roberto Basili

Overview

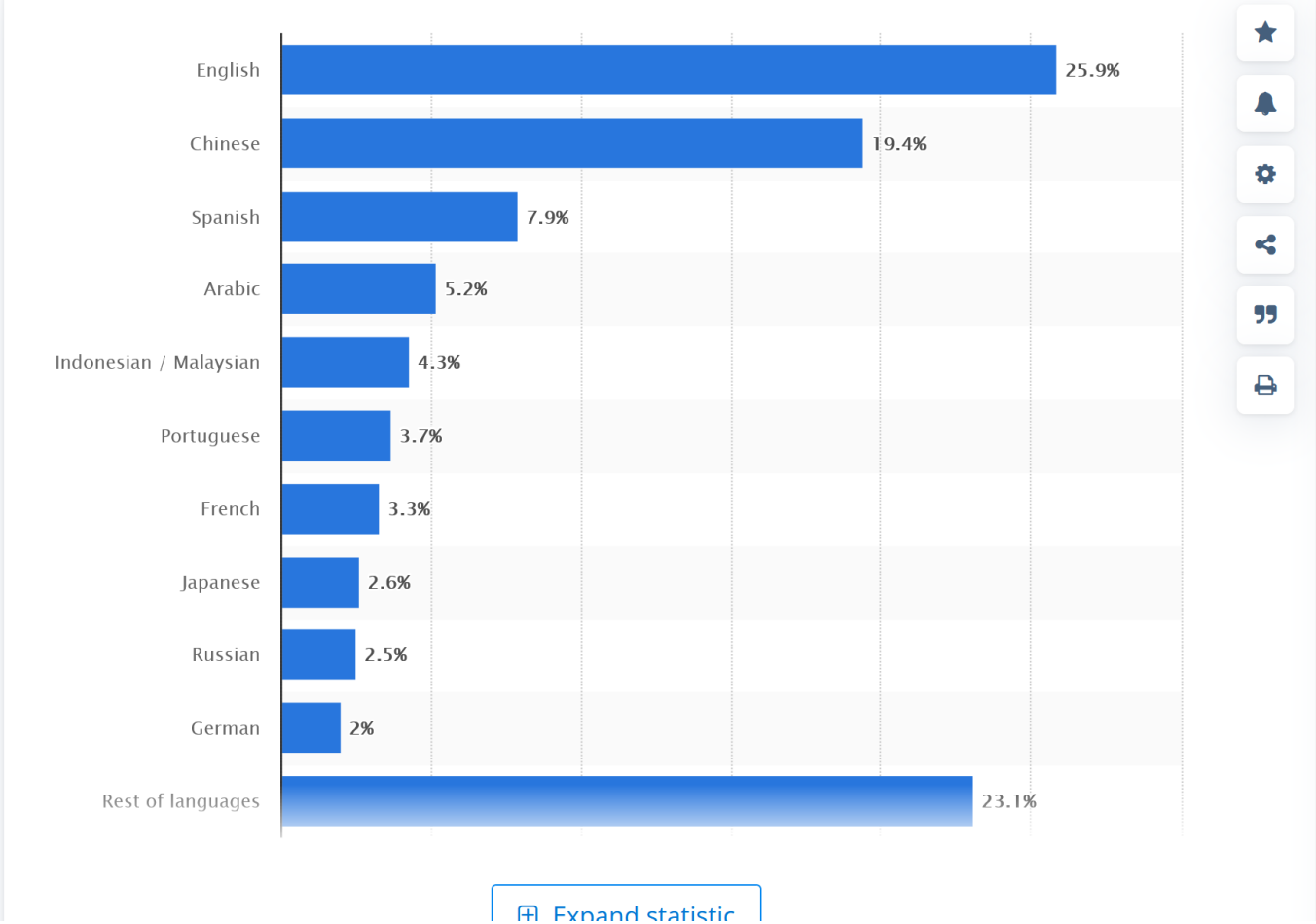
- Web Mining & Retrieval: Motivations & perspectives
  - Web, User-generated contents, Social Media
  - The role of *learning*
  - What is Machine Learning?
  - Data-driven algorithms: sources of complexity
- Main Applications
  - Intelligent Web Search
  - User Profiling for Marketing or Brand reputation management
  - Web Recommending
  - Spoken Dialogue Interaction in Robotics or in Web/mobile Interfaces

Internet statistics (Jan 2021)

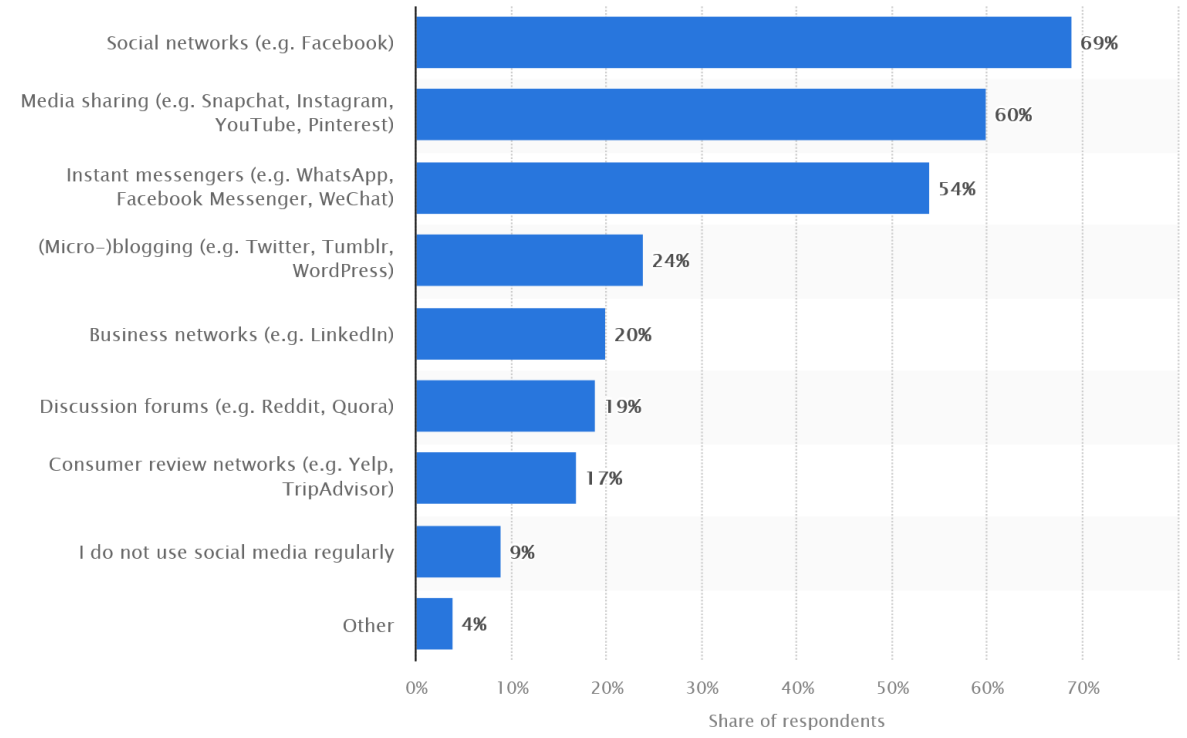


Internet Statistics (Jan 2020)

Most common languages used on the internet as of January 2020, by share of internet users



What kinds of social media do you use regularly?



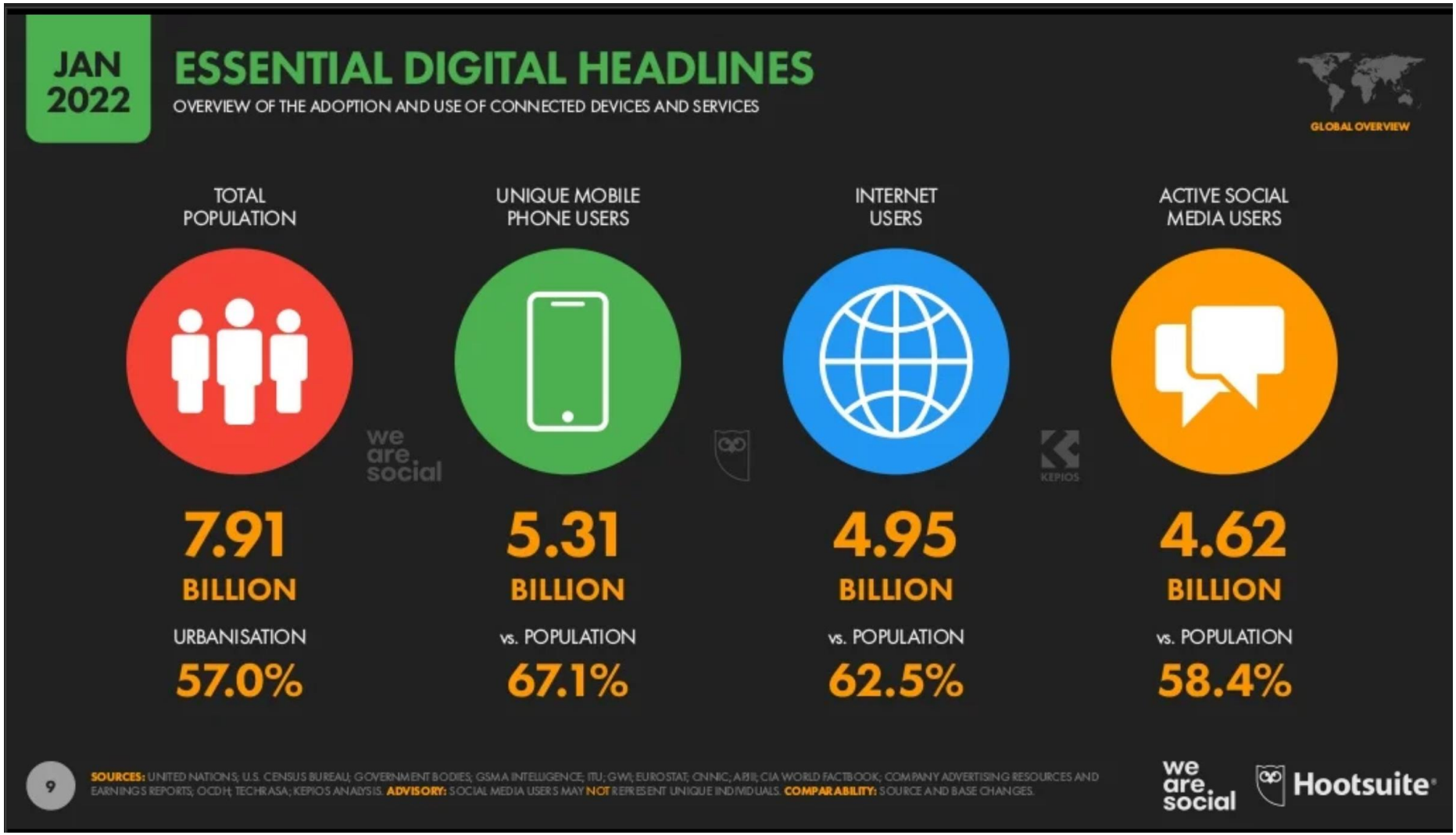
Details:

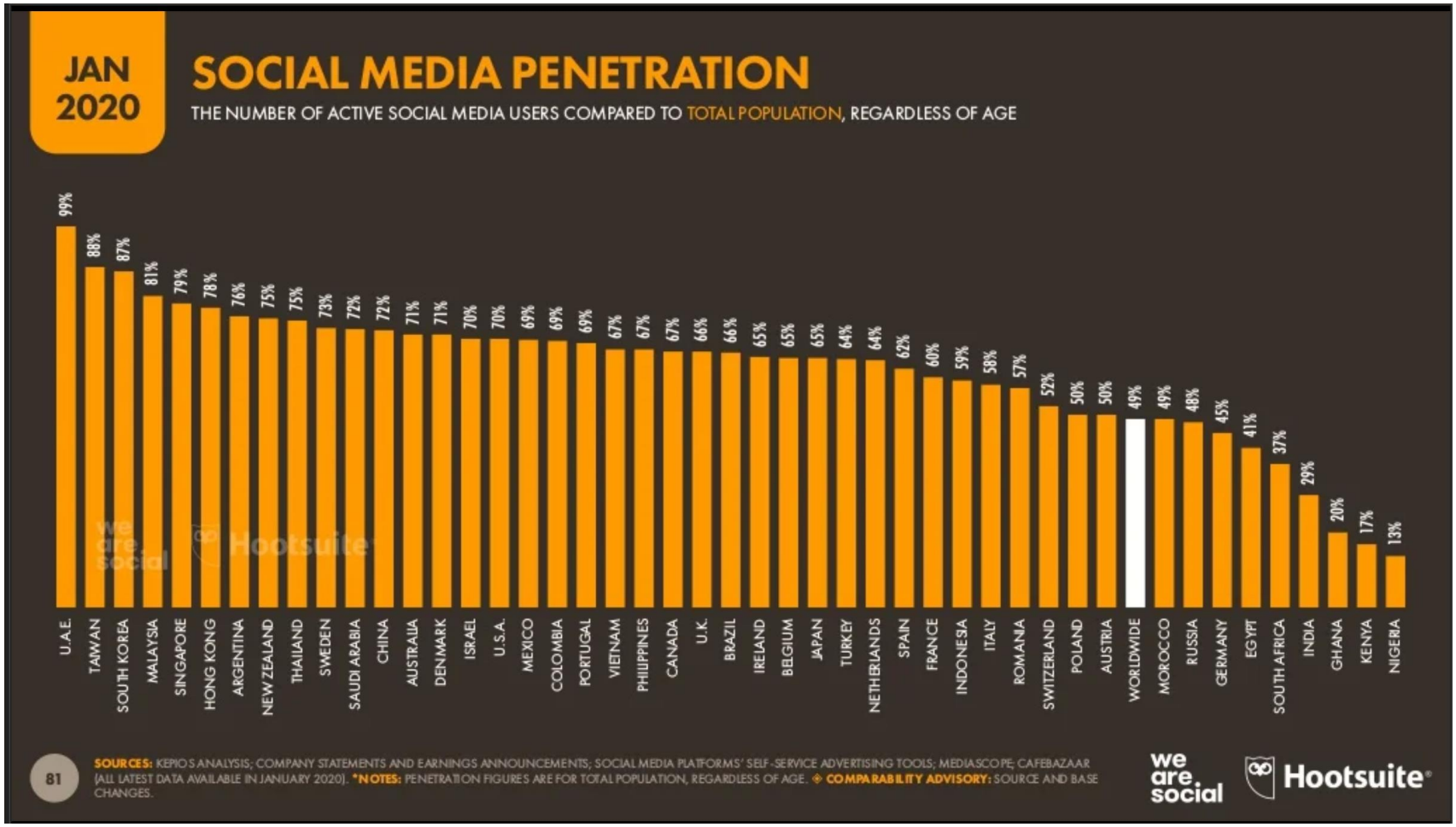
Show source

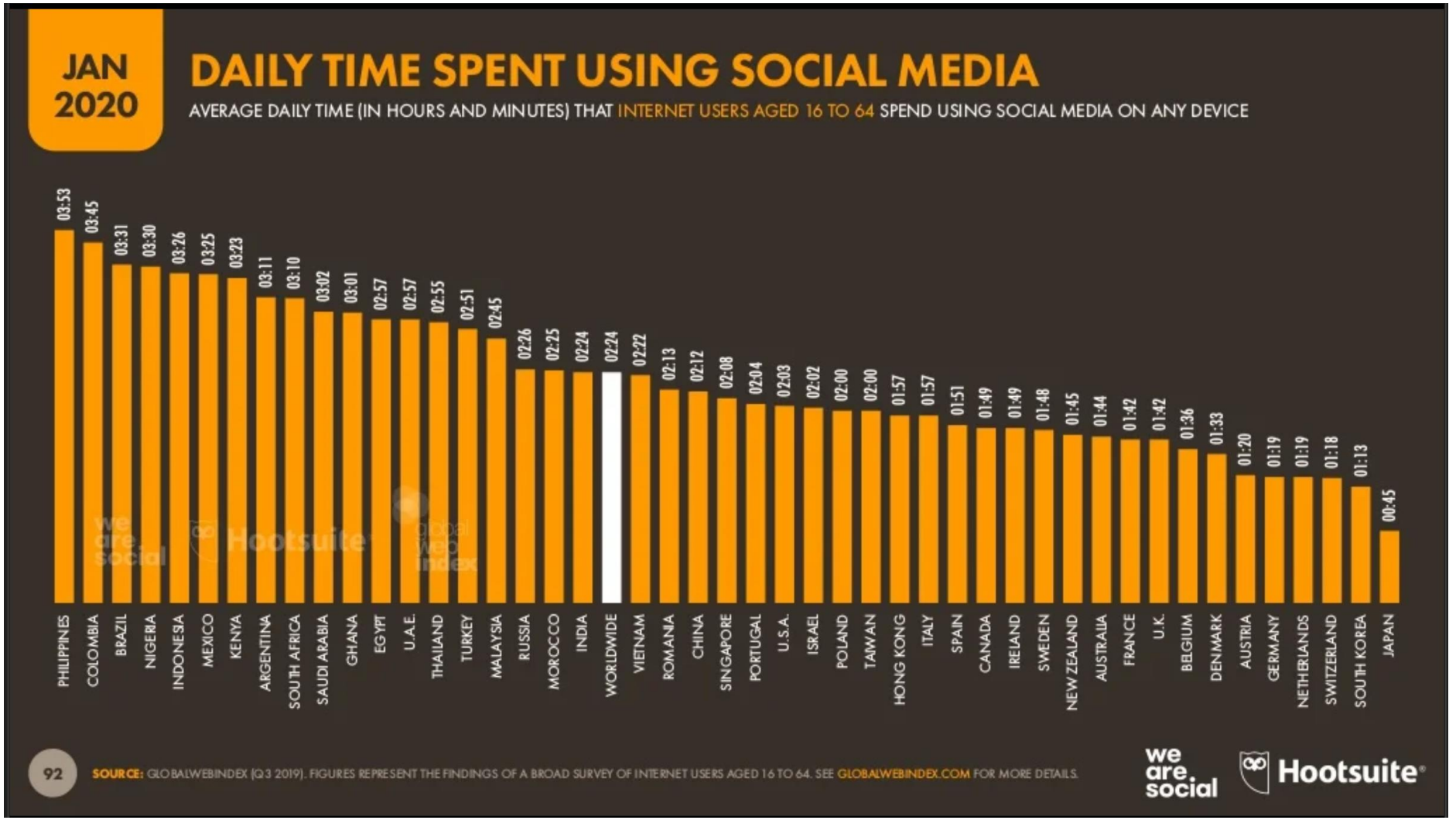
Do you know

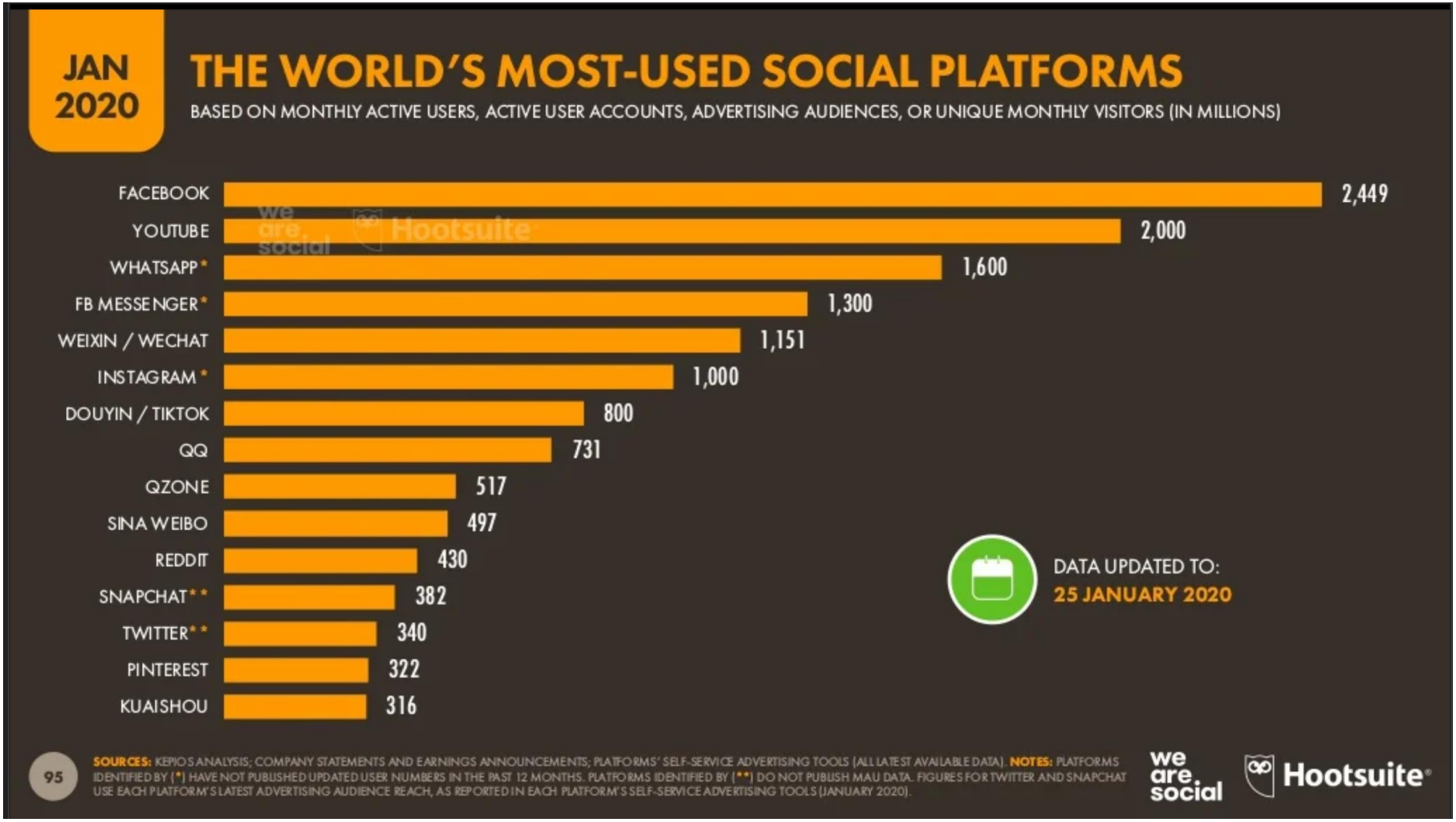
Do you know

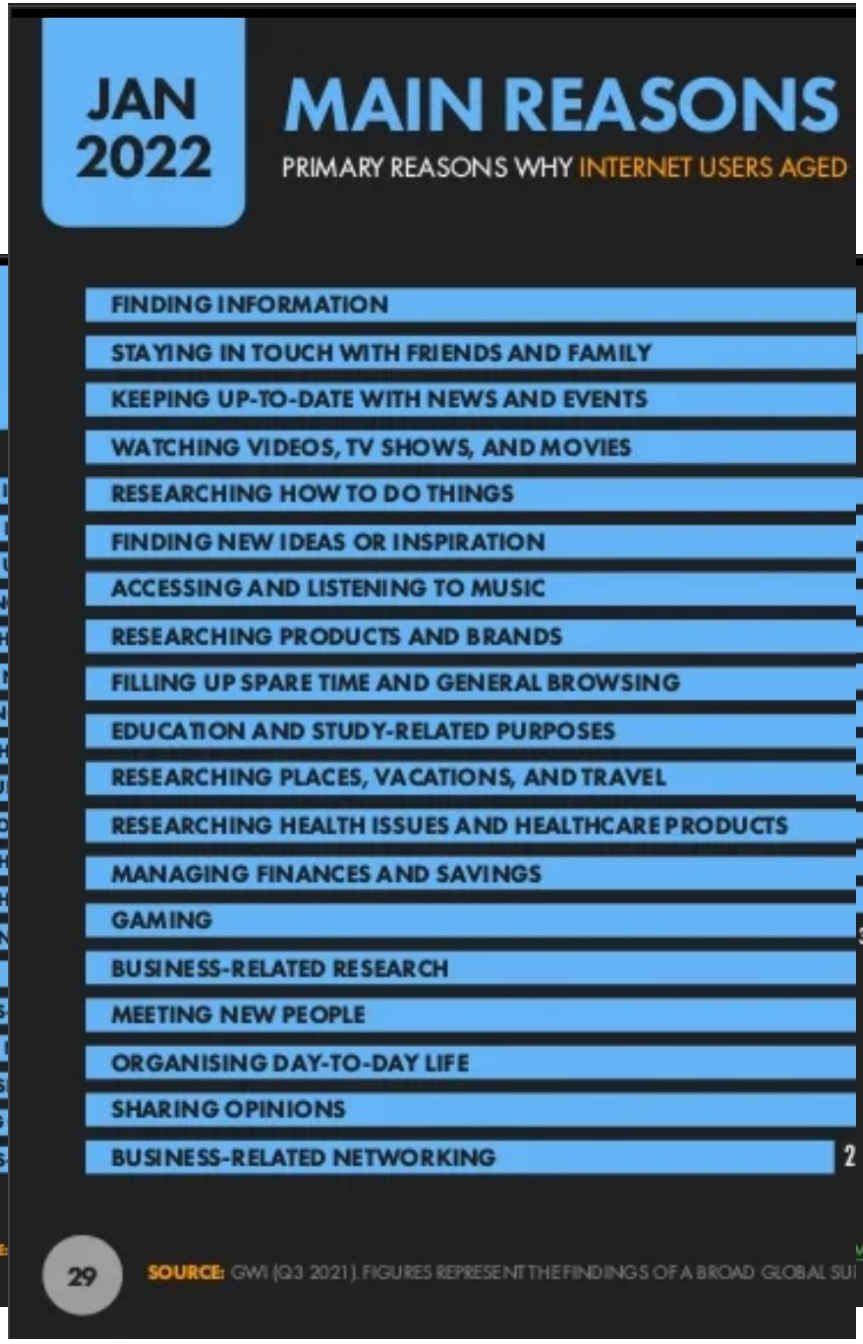
Contains more information than a person was likely to come across in a lifetime in the 18th century...

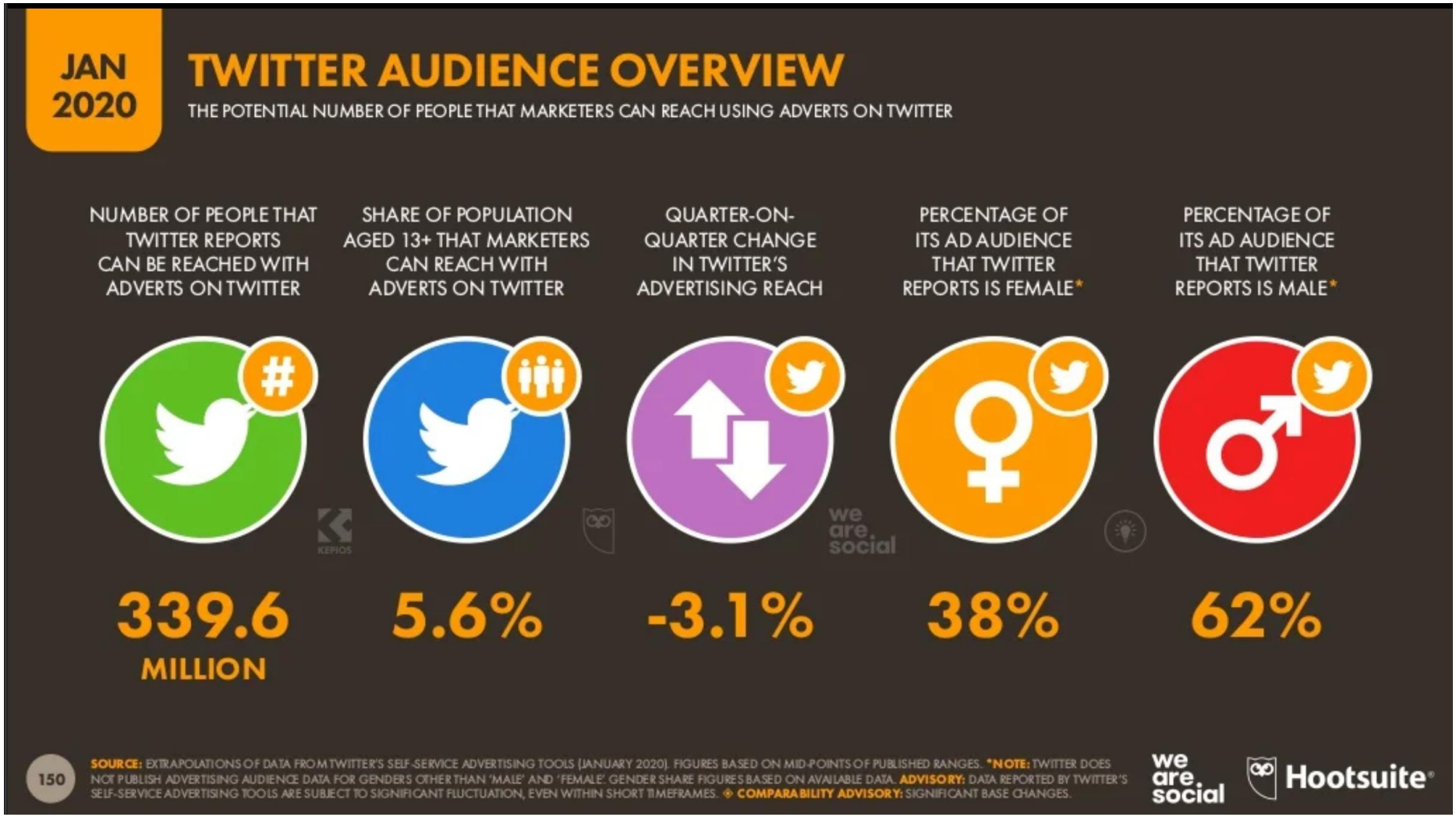

JAN 2022
ESSENTIAL DIGITAL HEADLINES
OVERVIEW OF THE ADOPTION AND USE OF CONNECTED DEVICES AND SERVICES
GLOBAL OVERVIEW
TOTAL POPULATION
UNIQUE MOBILE PHONE USERS
INTERNET USERS
ACTIVE SOCIAL MEDIA USERS
we are social
KEPIOS
7.91 BILLION
5.31 BILLION
4.95 BILLION
4.62 BILLION
URBANISATION
57.0%
vs. POPULATION
67.1%
vs. POPULATION
62.5%
vs. POPULATION
58.4%
9
SOURCES: UNITED NATIONS; U.S. CENSUS BUREAU; GOVERNMENT BODIES; GSMA INTELLIGENCE; ITU; GWI; EUROSTAT; CNNIC; ABII; CIA WORLD FACTBOOK; COMPANY ADVERTISING RESOURCES AND EARNINGS REPORTS; OCDH; TECHRASA; KEPIOS ANALYSIS. ADVISORY: SOCIAL MEDIA USERS MAY NOT REPRESENT UNIQUE INDIVIDUALS. COMPARABILITY: SOURCE AND BASE CHANGES.
we are social
Hootsuite

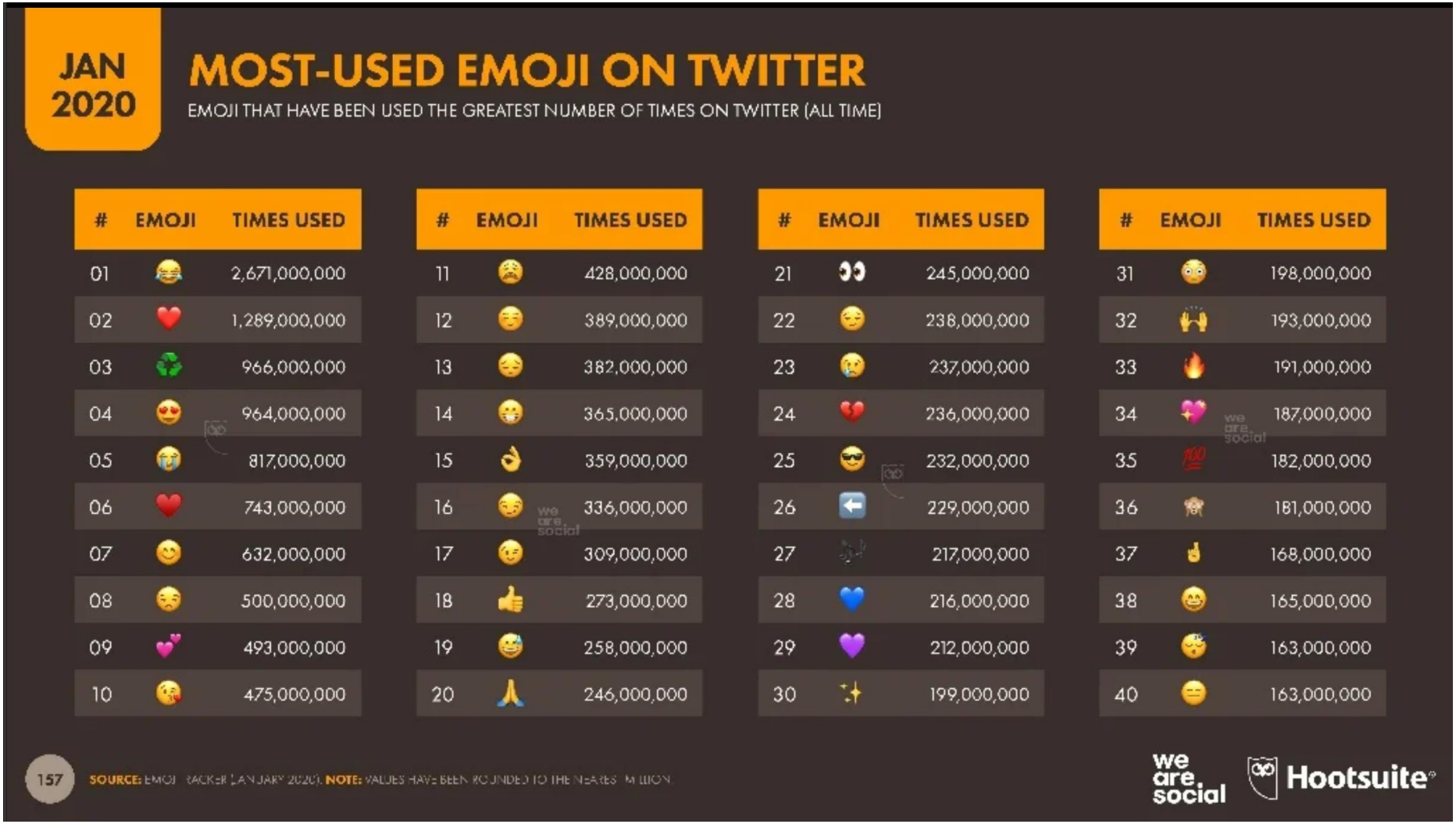

JAN 2020
SOCIAL MEDIA PENETRATION
THE NUMBER OF ACTIVE SOCIAL MEDIA USERS COMPARED TO TOTAL POPULATION, REGARDLESS OF AGE
U.A.E. 99%
TAIWAN 88%
SOUTH KOREA 87%
MALAYSIA 81%
SINGAPORE 79%
HONG KONG 78%
ARGENTINA 76%
NEW ZEALAND 75%
THAILAND 75%
SWEDEN 73%
SAUDI ARABIA 72%
CHINA 72%
AUSTRALIA 71%
DENMARK 71%
ISRAEL 70%
U.S.A. 70%
MEXICO 69%
COLOMBIA 69%
PORTUGAL 69%
VIETNAM 67%
PHILIPPINES 67%
CANADA 67%
U.K. 66%
BRAZIL 66%
IRELAND 65%
BELGIUM 65%
JAPAN 65%
TURKEY 64%
NETHERLANDS 64%
SPAIN 62%
FRANCE 60%
INDONESIA 59%
ITALY 58%
ROMANIA 57%
SWITZERLAND 52%
POLAND 50%
AUSTRIA 50%
WORLDWIDE 49%
MOROCCO 49%
RUSSIA 48%
GERMANY 45%
EGYPT 41%
SOUTH AFRICA 37%
INDIA 29%
GHANA 20%
KENYA 17%
NIGERIA 13%
81
SOURCES: KEPIOS ANALYSIS; COMPANY STATEMENTS AND EARNINGS ANNOUNCEMENTS; SOCIAL MEDIA PLATFORMS' SELF-SERVICE ADVERTISING TOOLS; MEDIASCOPE; CAFEBAZAAR (ALL LATEST DATA AVAILABLE IN JANUARY 2020). *NOTES: PENETRATION FIGURES ARE FOR TOTAL POPULATION, REGARDLESS OF AGE. ◈ COMPARABILITY ADVISORY: SOURCE AND BASE CHANGES.
we are social
Hootsuite

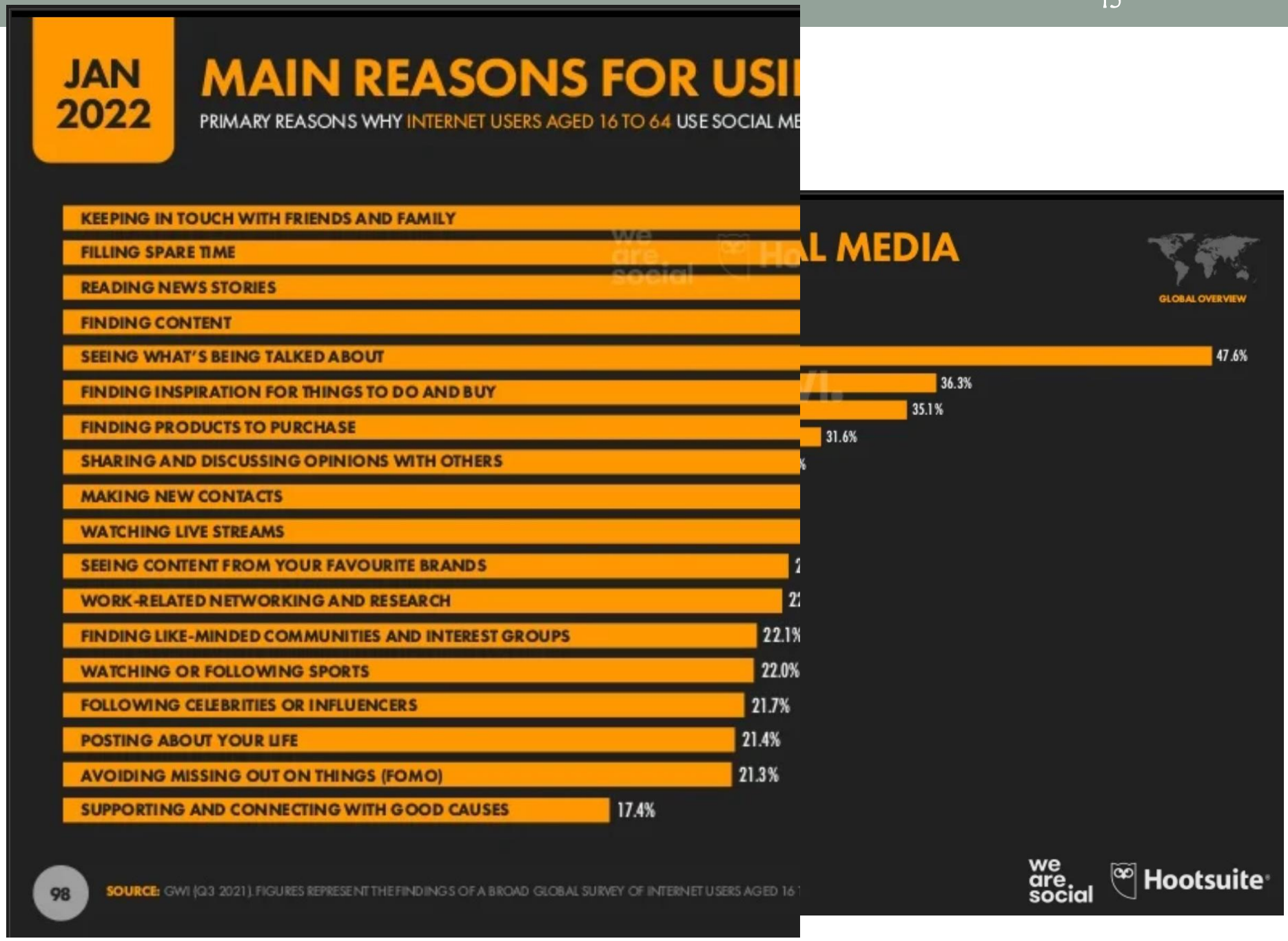

JAN 2020
DAILY TIME SPENT USING SOCIAL MEDIA
AVERAGE DAILY TIME (IN HOURS AND MINUTES) THAT INTERNET USERS AGED 16 TO 64 SPEND USING SOCIAL MEDIA ON ANY DEVICE
PHILIPPINES 03:53
COLOMBIA 03:45
BRAZIL 03:31
NIGERIA 03:30
INDONESIA 03:26
MEXICO 03:25
KENYA 03:23
ARGENTINA 03:11
SOUTH AFRICA 03:10
SAUDI ARABIA 03:02
GHANA 03:01
EGYPT 02:57
U.A.E. 02:57
THAILAND 02:55
TURKEY 02:51
MALAYSIA 02:45
RUSSIA 02:26
MOROCCO 02:25
INDIA 02:24
WORLDWIDE 02:24
VIETNAM 02:22
ROMANIA 02:13
CHINA 02:12
SINGAPORE 02:08
PORTUGAL 02:04
U.S.A. 02:03
ISRAEL 02:02
POLAND 02:00
TAIWAN 02:00
HONG KONG 01:57
ITALY 01:57
SPAIN 01:51
CANADA 01:49
IRELAND 01:49
SWEDEN 01:48
NEW ZEALAND 01:45
AUSTRALIA 01:44
FRANCE 01:42
U.K. 01:42
BELGIUM 01:36
DENMARK 01:33
AUSTRIA 01:20
GERMANY 01:19
NETHERLANDS 01:19
SWITZERLAND 01:18
SOUTH KOREA 01:13
JAPAN 00:45
92
SOURCE: GLOBALWEBINDEX (Q3 2019). FIGURES REPRESENT THE FINDINGS OF A BROAD SURVEY OF INTERNET USERS AGED 16 TO 64. SEE GLOBALWEBINDEX.COM FOR MORE DETAILS.
we are social
Hootsuite


JAN 2020
THE WORLD'S MOST-USED SOCIAL PLATFORMS
BASED ON MONTHLY ACTIVE USERS, ACTIVE USER ACCOUNTS, ADVERTISING AUDIENCES, OR UNIQUE MONTHLY VISITORS (IN MILLIONS)
FACEBOOK 2,449
YOUTUBE 2,000
WHATSAPP* 1,600
FB MESSENGER* 1,300
WEIXIN / WECHAT 1,151
INSTAGRAM* 1,000
DOUYIN / TIKTOK 800
QQ 731
QZONE 517
SINA WEIBO 497
REDDIT 430
SNAPCHAT** 382
TWITTER** 340
PINTEREST 322
KUAISHOU 316
DATA UPDATED TO:
25 JANUARY 2020
95
SOURCES: KEPIOS ANALYSIS; COMPANY STATEMENTS AND EARNINGS ANNOUNCEMENTS; PLATFORMS' SELF-SERVICE ADVERTISING TOOLS (ALL LATEST AVAILABLE DATA). NOTES: PLATFORMS IDENTIFIED BY (*) HAVE NOT PUBLISHED UPDATED USER NUMBERS IN THE PAST 12 MONTHS. PLATFORMS IDENTIFIED BY (**) DO NOT PUBLISH MAU DATA. FIGURES FOR TWITTER AND SNAPCHAT USE EACH PLATFORM'S LATEST ADVERTISING AUDIENCE REACH, AS REPORTED IN EACH PLATFORM'S SELF-SERVICE ADVERTISING TOOLS (JANUARY 2020).
we are social
Hootsuite


JAN 2022
MAIN REASONS
PRIMARY REASONS WHY INTERNET USERS AGED
FINDING INFORMATION
STAYING IN TOUCH WITH FRIENDS AND FAMILY
KEEPING UP-TO-DATE WITH NEWS AND EVENTS
WATCHING VIDEOS, TV SHOWS, AND MOVIES
RESEARCHING HOW TO DO THINGS
FINDING NEW IDEAS OR INSPIRATION
ACCESSING AND LISTENING TO MUSIC
RESEARCHING PRODUCTS AND BRANDS
FILLING UP SPARE TIME AND GENERAL BROWSING
EDUCATION AND STUDY-RELATED PURPOSES
RESEARCHING PLACES, VACATIONS, AND TRAVEL
RESEARCHING HEALTH ISSUES AND HEALTHCARE PRODUCTS
MANAGING FINANCES AND SAVINGS
GAMING
BUSINESS-RELATED RESEARCH
MEETING NEW PEOPLE
ORGANISING DAY-TO-DAY LIFE
SHARING OPINIONS
BUSINESS-RELATED NETWORKING
29
SOURCE: GWI (Q3 2021). FIGURES REPRESENT THE FINDINGS OF A BROAD GLOBAL SU


JAN 2020
TWITTER AUDIENCE OVERVIEW
THE POTENTIAL NUMBER OF PEOPLE THAT MARKETERS CAN REACH USING ADVERTS ON TWITTER
NUMBER OF PEOPLE THAT TWITTER REPORTS CAN BE REACHED WITH ADVERTS ON TWITTER
SHARE OF POPULATION AGED 13+ THAT MARKETERS CAN REACH WITH ADVERTS ON TWITTER
QUARTER-ON-QUARTER CHANGE IN TWITTER'S ADVERTISING REACH
PERCENTAGE OF ITS AD AUDIENCE THAT TWITTER REPORTS IS FEMALE*
PERCENTAGE OF ITS AD AUDIENCE THAT TWITTER REPORTS IS MALE*
KEPIOS
we are social
339.6 MILLION
5.6%
-3.1%
38%
62%
150
SOURCE: EXTRAPOLATIONS OF DATA FROM TWITTER'S SELF-SERVICE ADVERTISING TOOLS (JANUARY 2020). FIGURES BASED ON MID-POINTS OF PUBLISHED RANGES. *NOTE: TWITTER DOES NOT PUBLISH ADVERTISING AUDIENCE DATA FOR GENDERS OTHER THAN 'MALE' AND 'FEMALE'. GENDER SHARE FIGURES BASED ON AVAILABLE DATA. ADVISORY: DATA REPORTED BY TWITTER'S SELF-SERVICE ADVERTISING TOOLS ARE SUBJECT TO SIGNIFICANT FLUCTUATION, EVEN WITHIN SHORT TIMEFRAMES. ◈ COMPARABILITY ADVISORY: SIGNIFICANT BASE CHANGES.
we are social
Hootsuite


JAN 2020
MOST-USED EMOJI ON TWITTER
EMOJI THAT HAVE BEEN USED THE GREATEST NUMBER OF TIMES ON TWITTER (ALL TIME)
EMOJI TIMES USED
01 2,671,000,000
02 1,289,000,000
03 966,000,000
04 964,000,000
05 817,000,000
06 743,000,000
07 632,000,000
08 500,000,000
09 493,000,000
10 475,000,000
11 428,000,000
12 389,000,000
13 382,000,000
14 365,000,000
15 359,000,000
16 336,000,000
17 309,000,000
18 273,000,000
19 258,000,000
20 246,000,000
21 245,000,000
22 238,000,000
23 237,000,000
24 236,000,000
25 232,000,000
26 229,000,000
27 217,000,000
28 216,000,000
29 212,000,000
30 199,000,000
31 198,000,000
32 193,000,000
33 191,000,000
34 187,000,000
35 182,000,000
36 181,000,000
37 168,000,000
38 165,000,000
39 163,000,000
40 163,000,000
157
SOURCE: EMOJI TRACKER JANUARY 2020. NOTE: VALUES HAVE BEEN ROUNDED TO THE NEAREST MILLION
we are social
Hootsuite


JAN 2022
MAIN REASONS FOR USI
PRIMARY REASONS WHY INTERNET USERS AGED 16 TO 64 USE SOCIAL ME
KEEPING IN TOUCH WITH FRIENDS AND FAMILY
FILLING SPARE TIME
READING NEWS STORIES
FINDING CONTENT
SEEING WHAT'S BEING TALKED ABOUT
FINDING INSPIRATION FOR THINGS TO DO AND BUY
FINDING PRODUCTS TO PURCHASE
SHARING AND DISCUSSING OPINIONS WITH OTHERS
MAKING NEW CONTACTS
WATCHING LIVE STREAMS
SEEING CONTENT FROM YOUR FAVOURITE BRANDS
WORK-RELATED NETWORKING AND RESEARCH
FINDING LIKE-MINDED COMMUNITIES AND INTEREST GROUPS
22.1%
WATCHING OR FOLLOWING SPORTS
22.0%
FOLLOWING CELEBRITIES OR INFLUENCERS
21.7%
POSTING ABOUT YOUR LIFE
21.4%
AVOIDING MISSING OUT ON THINGS (FOMO)
21.3%
SUPPORTING AND CONNECTING WITH GOOD CAUSES
17.4%
L MEDIA
GLOBAL OVERVIEW
47.6%
36.3%
35.1%
31.6%
98
SOURCE: GWI (Q3 2021). FIGURES REPRESENT THE FINDINGS OF A BROAD GLOBAL SURVEY OF INTERNET USERS AGED 16
we are social
Hootsuite

WE ARE SOCIAL'S PERSPECTIVE: SOCIAL IN 2020

SHIFTS IN HOW PEOPLE BEHAVE AND INTERACT ON SOCIAL

BAD INFLUENCE

Being a creator has lost its lo-fi sheen; many lifestyle influencers lead unrelatable lives, while celebrity 'creators' like Will Smith are blowing up on platforms like YouTube and TikTok. As a result, there's a growing backlash against influencer culture and the metrics that drive it.

In 2020, brands will look beyond likes, followers and reach to generate genuine engagement

ADDED VALUE

The internet has long been a wild west where intellectual property is barely there. But in a maturing digital frontier, creators have grown dedicated audiences who not only see value in their content, but recognise their style anywhere. As a result, communities are rallying to protect creators.

In 2020, brands will take greater steps to ensure they're being respectful of digital communities

RUNNING COMMENTARY

Audiences are increasingly willing to invest time and attention in content and narratives they deem to have a higher value. This isn't about a shift back to traditional media. It's about longer, more complex content designed to be consumed in-platform and on smaller screens.

In 2020, brands will tell more complex stories across multiple touchpoints on social

we are social Hootsuite

Dealing with *real* Social media data

WM&R: Motivations

- *What does Web Mining mean*?
- *Why Information Retrieval is involved*?
- *Why Machine Learning*?
- *Which are the contributions of IR/ML to technologies that support and exploit Web Data, Information and Knowledge*?
- *Which are the technological perspectives in the medium-long term*?

What is Web Mining?

- *Web Mining* refers to a body of technologies currently needed for the *exploitation of publicly available information from the Web and the I*oT
  - Contents: data but also … people, locations, events, concepts, …
  - Relations:
    - Links within structured networks
    - Thematic, interpersonal and semantic associations
    - Analogies
  - On-Line Structured and semi-structured resources (e.g. Wikipedia)
  - Textual, Multimedia and Multilingual Contents
  - Trends e time-related information (community on-line behaviours)
  - Opinions, Preferences, Expectations

Why IR?

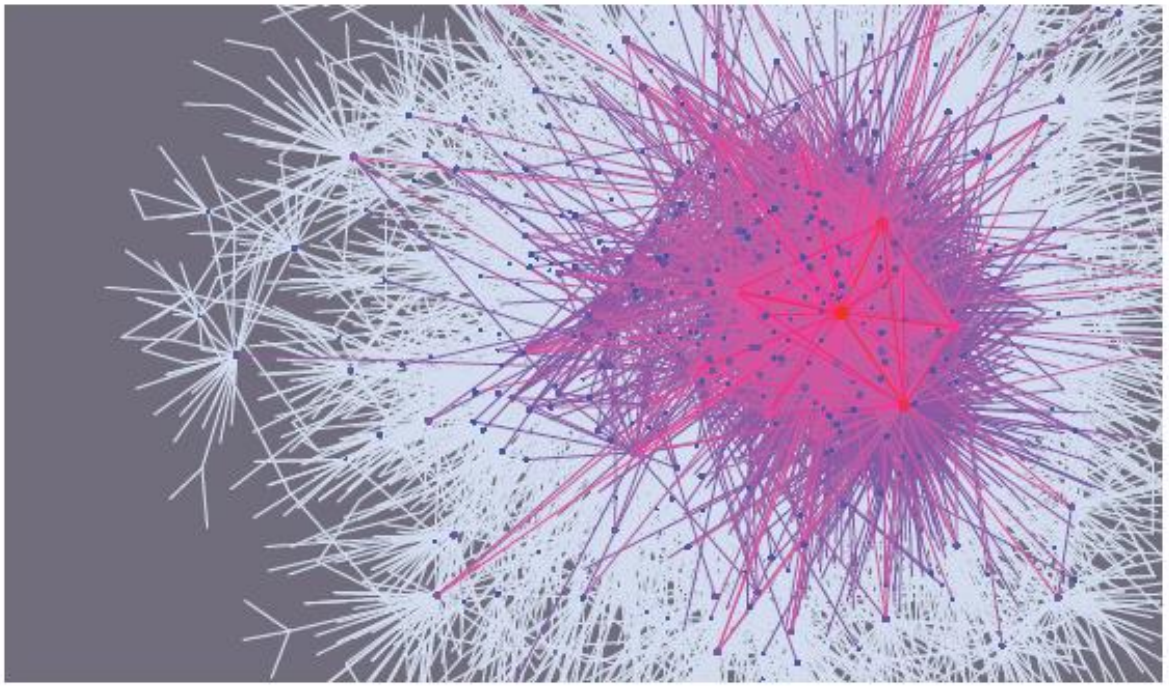
- The volumes involved in Web Mining pose the crucial problem of *locating information* beforehand
- Automatic information access is possible only if we solve the two major challenges
  - What is relevant
  - Where the relevant information is located
- Searching information corresponds to computing an uncertain function that models the *mapping* between information needs and the targeted data

Machine Learning vs IR?

- Web mining involve heterogeneous information that is characterized search as strongly uncertainy process
- The available information is characterized by:
  - Incompleteness:
    - Short queries as an incomplete description of the information need
  - Variability: Wealth of data vs. heterogeneity of formats and access modes
    - Contents are dispersed in various forms across data sources
  - Vague Requirements
    - Information is often implicit (i.e. partially and qualitatively expressed) in the operational contexts
  - Subjectivity
    - Relevance depends on the user and not just on the contents
  - Timeliness
  - Authority

Machine Learning vs. IR

- Uncertainty os so pervasive that exhaustive solutions (i.e. global *optima*) are not available or even not existing
- "*Finding diamonds in the rough*" (Fan Chung, UCSD)



Machine Learning vs. IR

- Le tecniche di ML propongono una ampia serie di algoritmi, strategie e tecniche per la produzione di soluzioni *sub-ottime* ma efficaci
- Nel processo di *learning* i dati suggeriscono la ipotesi risolutiva per la funzione di *mapping*
- Tale ipotesi è attesa migliorare la prestazione complessiva del sistema di base
  - Accuratezza
  - Efficienza computazionale

Machine Learning

- (Langley, 2000): l'Apprendimento Automatico si occupa dei meccanismi attraverso i quali un agente intelligente migliora nel tempo le sue prestazioni *P* nell'effettuare un compito *C*.
- La prova del successo dell'apprendimento è quindi nella capacità di misurare l'incremento ΔP delle prestazioni sulla base delle esperienze *E* che l'agente è in grado di raccogliere durante il suo ciclo di vita.
- La natura dell'apprendimento è quindi tutta nella caratterizzazione delle nozioni qui primitive di *compito*, *prestazione* ed *esperienza*.

Esperienza ed Apprendimento

- L'esperienza, per esempio, nel gioco degli scacchi può essere interpretata in diversi modi:
  - i dati sulle vittorie (e sconfitte) pregresse per valutare la bontà (o la inadeguatezza) di strategie e mosse eseguite rispetto all'avversario.
  - valutazione fornita sulle mosse da un docente esterno (oracolo, guida).
  - Adeguatezza dei comportamenti derivata dalla auto-osservazione, cioè dalla capacità di analizzare partite dell'agente contro se stesso secondo un modello esplicito del processo (partita) e della sua evoluzione (comportamento, vantaggi, …).

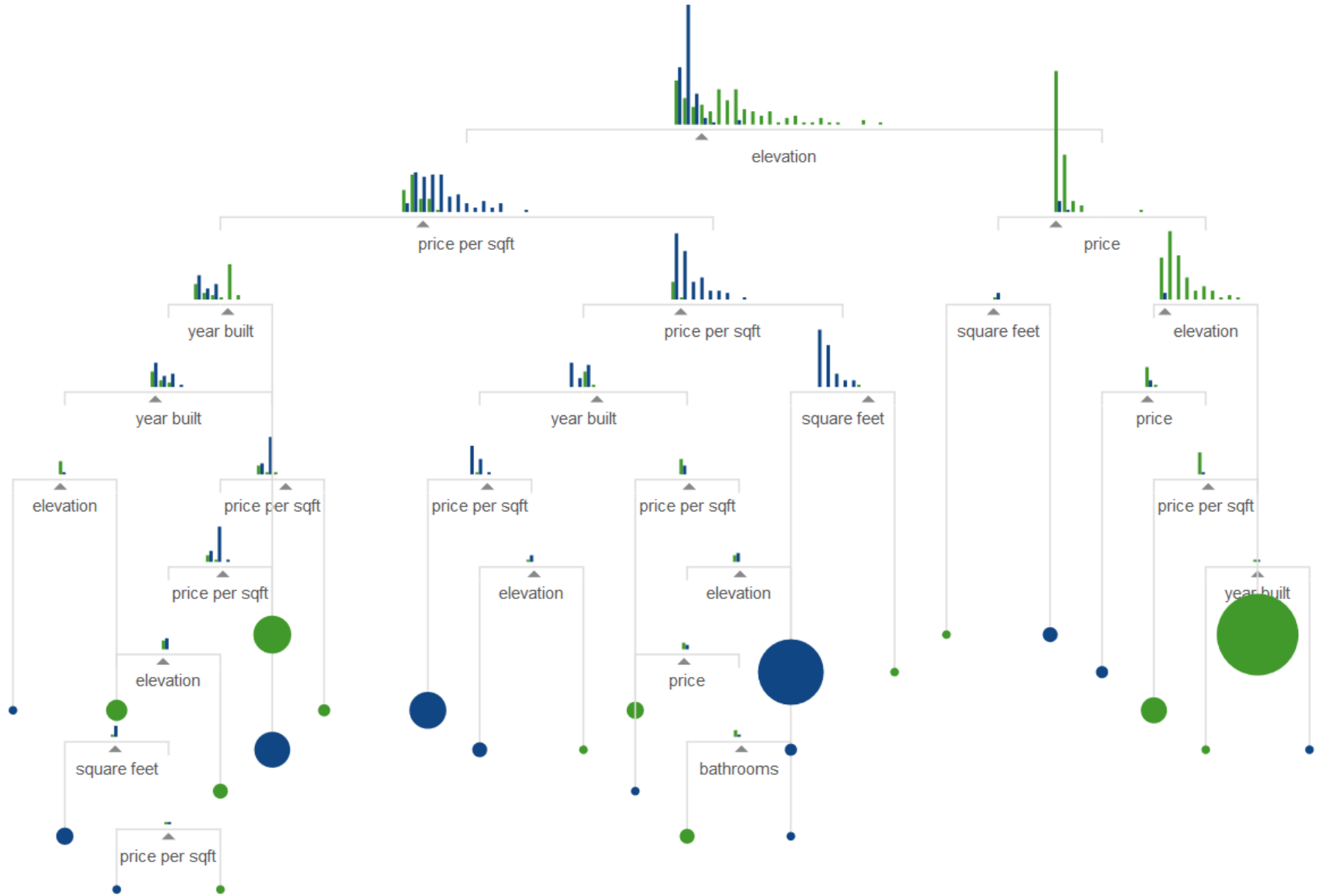
ML: una introduzione visuale

- See URL: http://www.r2d3.us/visual-intro-to-machine-learning-part-1/?imm_mid=0d76b4&cmp=em-data-na-na-newsltr_20150826

Apprendimento e Classi di Algoritmi

- Acquisizione di:
  - Funzioni logiche booleane, (ad es., alberi di decisione)
  - Induzione: determinazione ricorsiva delle CNES che caratterizzano i diversi sottogruppi .
- Approcci probabilistici:
  - Funzione target di Probabilità, (ad es., classificatore Bayesiano)
  - Induzione: Stima delle probabilità (in quanto parametri).
- Approcci geometrici
  - Funzioni di separazione in spazi vettoriali (lineari e non)
    - KNN
    - Funzioni Lineari, percettroni, Neural Networks, Support Vector Machines,…
    - Embeddings, analisi spettrale (trasformazioni di spazio)
  - Induzione: parametrizzare la funzione appartenente ad una certa classe (ad es. polinomi di grado *n*)

Es. apprendimento alberi di decisione



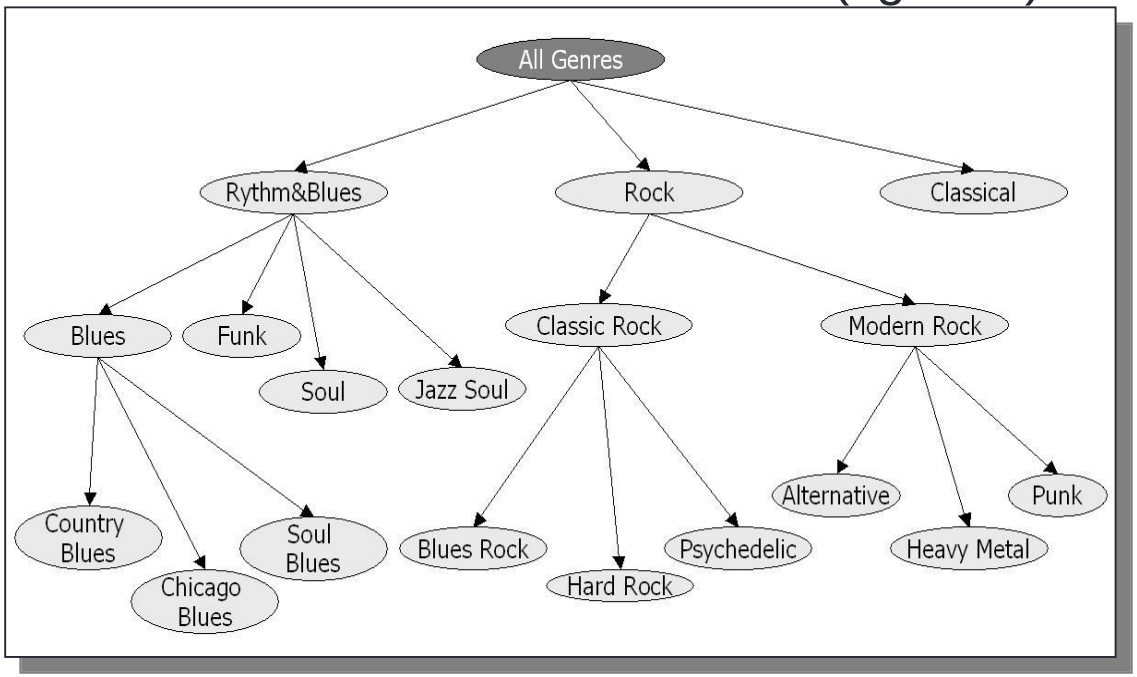
Apprendimento senza supervisione

- In assenza di un oracolo o di conoscenze sul task esistono ancora molti modi di migliorare le proprie prestazioni, ad es.
  - Migliorando il proprio modello del mondo (acquisizione/*discovery* della conoscenza)
  - Migliorando le proprie prestazioni computazionali (ottimizzazione)

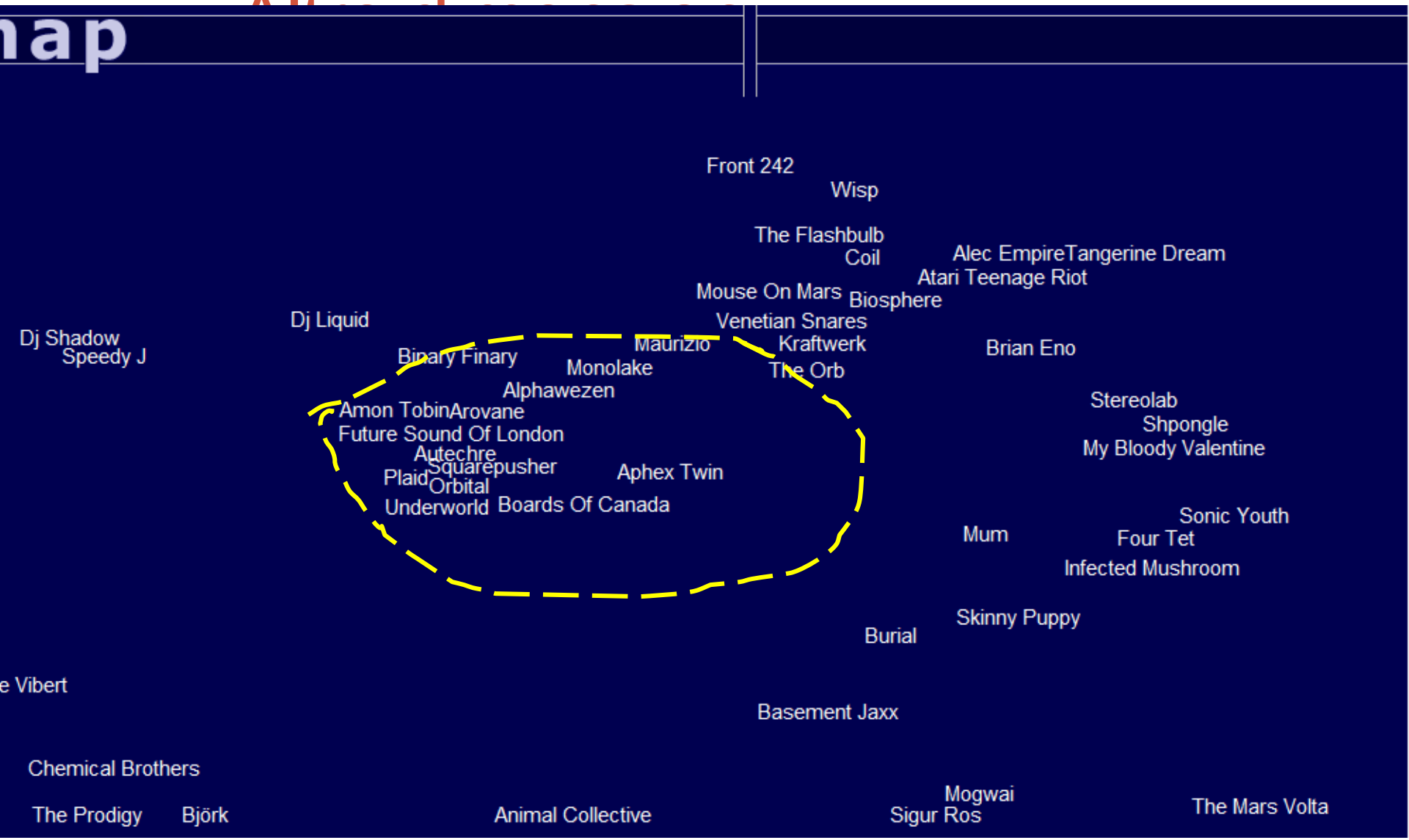
Apprendimento senza supervisione

Al termine del processo di acquisizione il sistema dispone di un sistema di classi e relazioni indotti che migliora la sua interazione futura con l'ambiente operativo (ad es. l'utente)

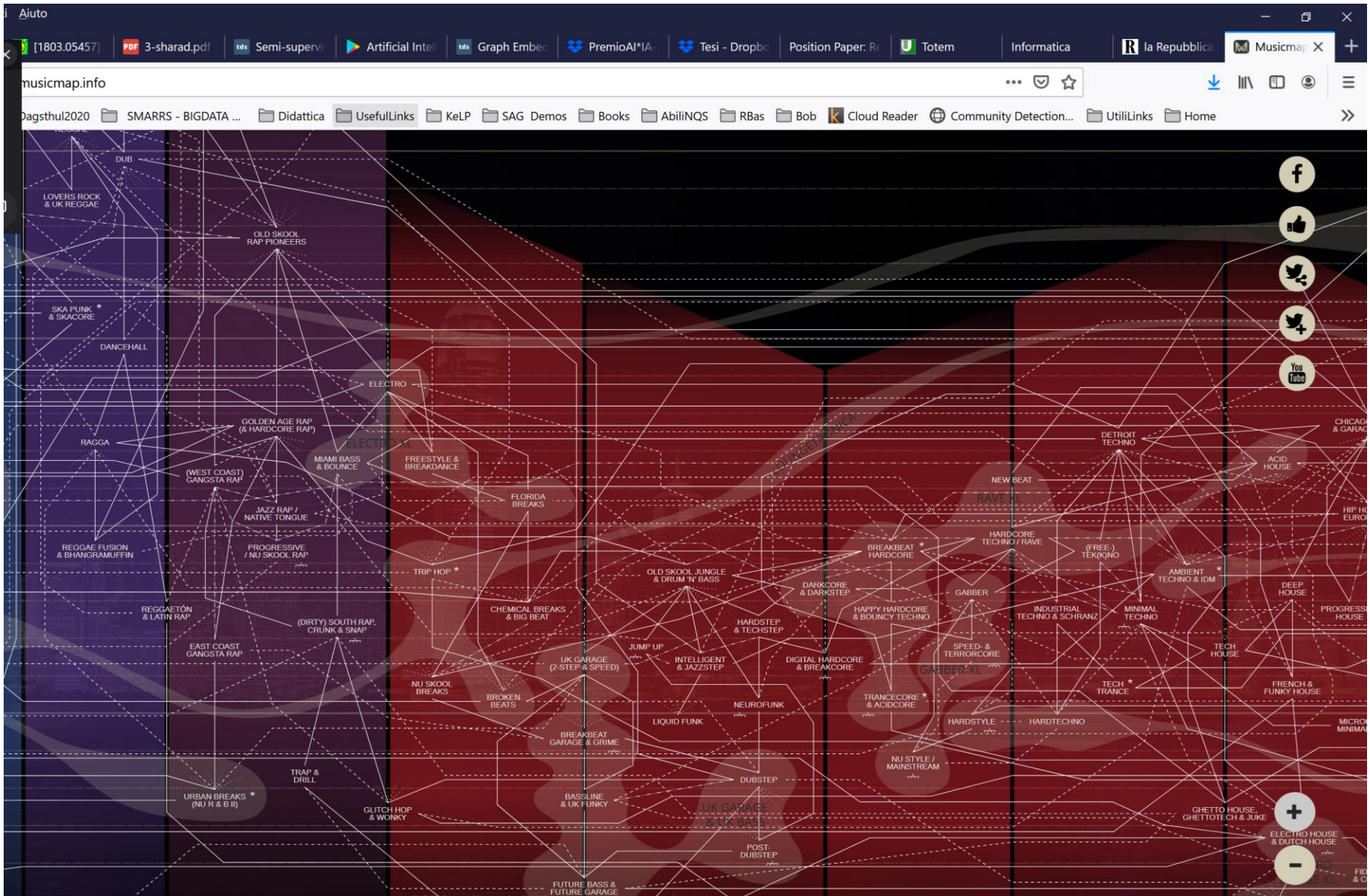
- Ese
  - u
    a
    a
    g
  - Il miglioramento avviene quindi almeno rispetto agli algoritmi di ricerca: la organizzazione gerarchica consente di esaminare solo i membri dell'insieme in alcune classi (i generi).



nap


Front 242
Wisp
The Flashbulb
Coil
Alec Empire
Tangerine Dream
Atari Teenage Riot
Mouse On Mars
Biosphere
Dj Liquid
Venetian Snares
Dj Shadow
Speedy J
Maurizio
Kraftwerk
Brian Eno
Binary Finary
Monolake
The Orb
Alphawezen
Stereolab
Amon Tobin
Arovane
Shpongle
Future Sound Of London
My Bloody Valentine
Autechre
Squarepusher
Plaid
Orbital
Aphex Twin
Underworld
Boards Of Canada
Sonic Youth
Mum
Four Tet
Infected Mushroom
Skinny Puppy
Burial
e Vibert
Basement Jaxx
Chemical Brothers
Mogwai
The Prodigy
Björk
Animal Collective
Sigur Ros
The Mars Volta

Music maps: 2020



Web IR

- Processing Web data: content detection, link detection, …
- Web Crawling
- Web Search: indici, link analysis
- Ranking: weighting contents, links and formats, authority, timeliness
- Meta-search
- Link Analysis

Information, Web and Natural Language

Web contents, characterized by rich multimedia information, are mostly opaque from a semantic standpoint

Information, Web and language

Hu meets KMT honorary chairman in Hawaii - People's Daily Online - Mozilla Firefox

File Modifica Visualizza Cronologia Segnalibri Yahoo! Strumenti Aiuto

Hu meets KMT honorary chairman

Indietro Avanti Down

Chinese President Hu Jintao (R) shakes hands with Honorary Chairman of the Chinese Kuomintang (KMT) Lien Chan, in Honolulu, Hawaii, the U.S., Nov. 11, 2011. (Xinhua/Huang Jingwen)

HONOLULU, United States, Nov. 11 (Xinhua) -- Hu Jintao, general secretary of the Central

Latest News: • Indonesia to host European Higher Education Fair

Beijing Sunny 15 / 1 City Forecast

Home >> China Politics

Hu meets KMT honorary chairman in Hawaii

(Xinhua)

11:10, November 12, 2011

Chinese President Hu Jintao (R) shakes hands with Honorary Chairman of the Chinese Kuomintang (KMT) Lien Chan, in Honolulu, Hawaii, the U.S., Nov. 11, 2011.

Selections for you

Miao ethnic group celebrates Miao's New Year in SW China

World's first Angry Birds exclusive shop opens in Helsinki

Who is Hu Jintao?

Most Popular

1. Hu reaffirms support to Hong Kong's sta...
2. Hu meets KMT honorary chairman in Hawaii
3. China in APEC: a mutually beneficial en...
4. Night life in Shanghai
5. China's 2011 foreign trade to grow 20 p...
6. Beijing house prices stumble 5.1 pct as...
7. Lama students start school in Tibet Col...
8. Police in central China crack phoney ca...

+Tu Web Immagini Video Mappe News Gmail Altro -

Google

Hu Jintao

Ricerca

Circa 725.000 risultati (0,09 secondi)

Tutto
Immagini
Mappe
Video
Notizie
Shopping
Più conte

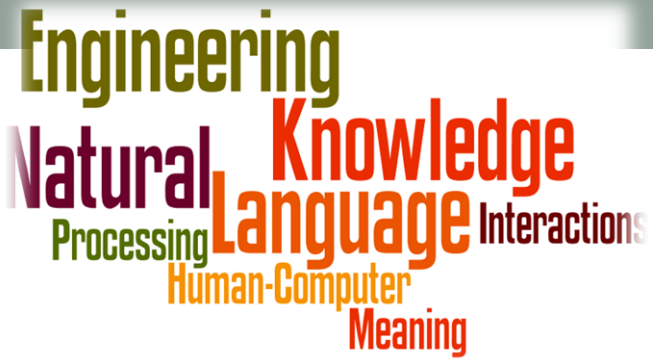
Tutti i ri
Per argomento

Qualsiasi dimensione
Grandi
Medie
Icone
Maggiori di...
Dimensioni esatte...

Qualsiasi colore
A colori
Bianco e nero

Qualsiasi tipo
Volti
Foto
Clip art
Disegni

Visual. standard
Mostra dimensioni

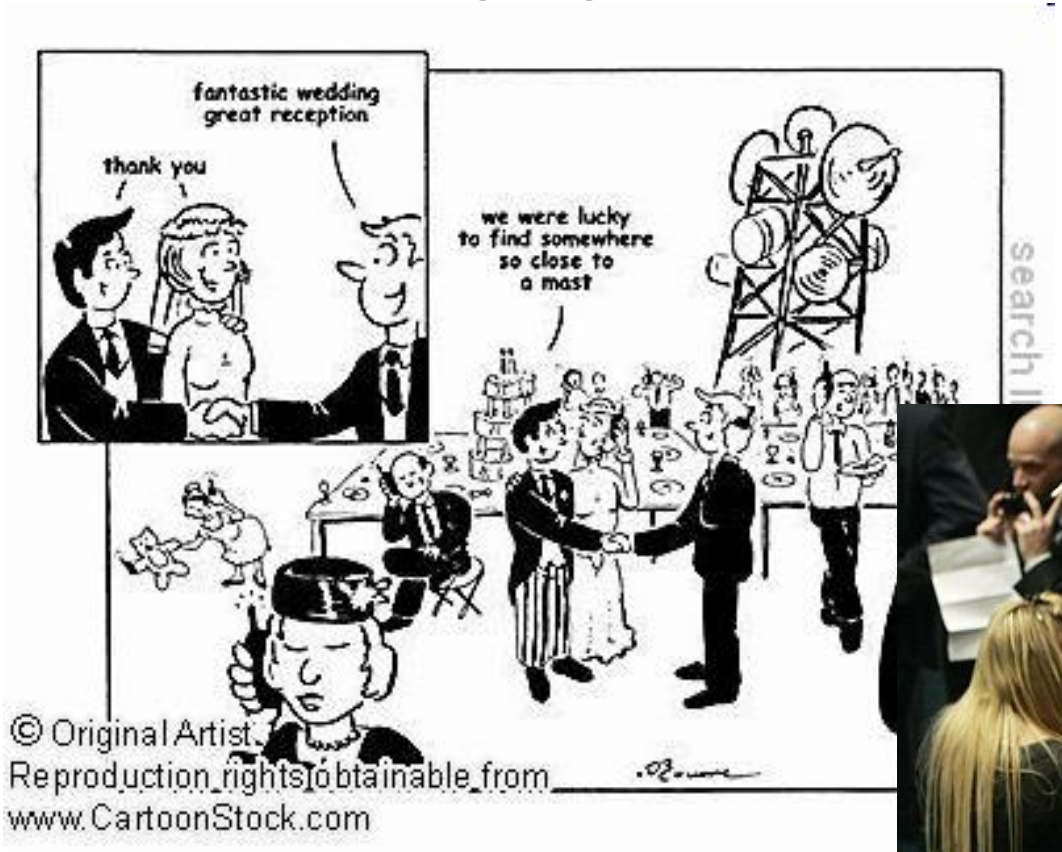


Content Semantics and Natural Language

- Human languages are the main carrier of the information involved in processes such as *retrieval*, *publication* and *exchange* of knowledge as it is associated to the open Web contents
- Words and NL syntactic structures express concepts, activities, events, abstractions and conceptual relations we usually share through data
- "*Language is parasitic to knowledge representation languages but the viceversa is not true*" (Wilks, 2001)

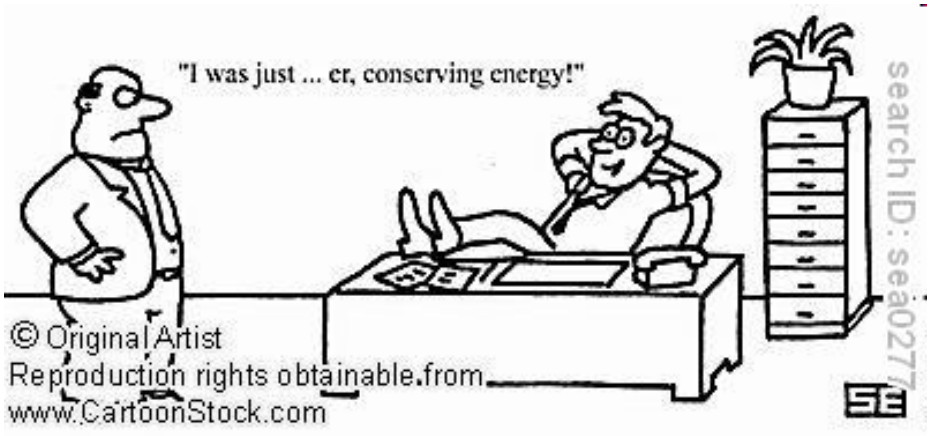
NL and Knowledge

- Natural Languages are even too successfull in modern



Benefits of a data-driven approach

- Very effective learning algorithms available (e.g. Support Vector Machines)
- The ML technology is *portable* while imperative coding is *task (*i.e. *scenario) specific*
- Very accurate solutions can be obtained
- Gathering training data *much less expensive* than rule coding
- In dynamically evolving scenarios, incremental refinement of the system only consists in re-training



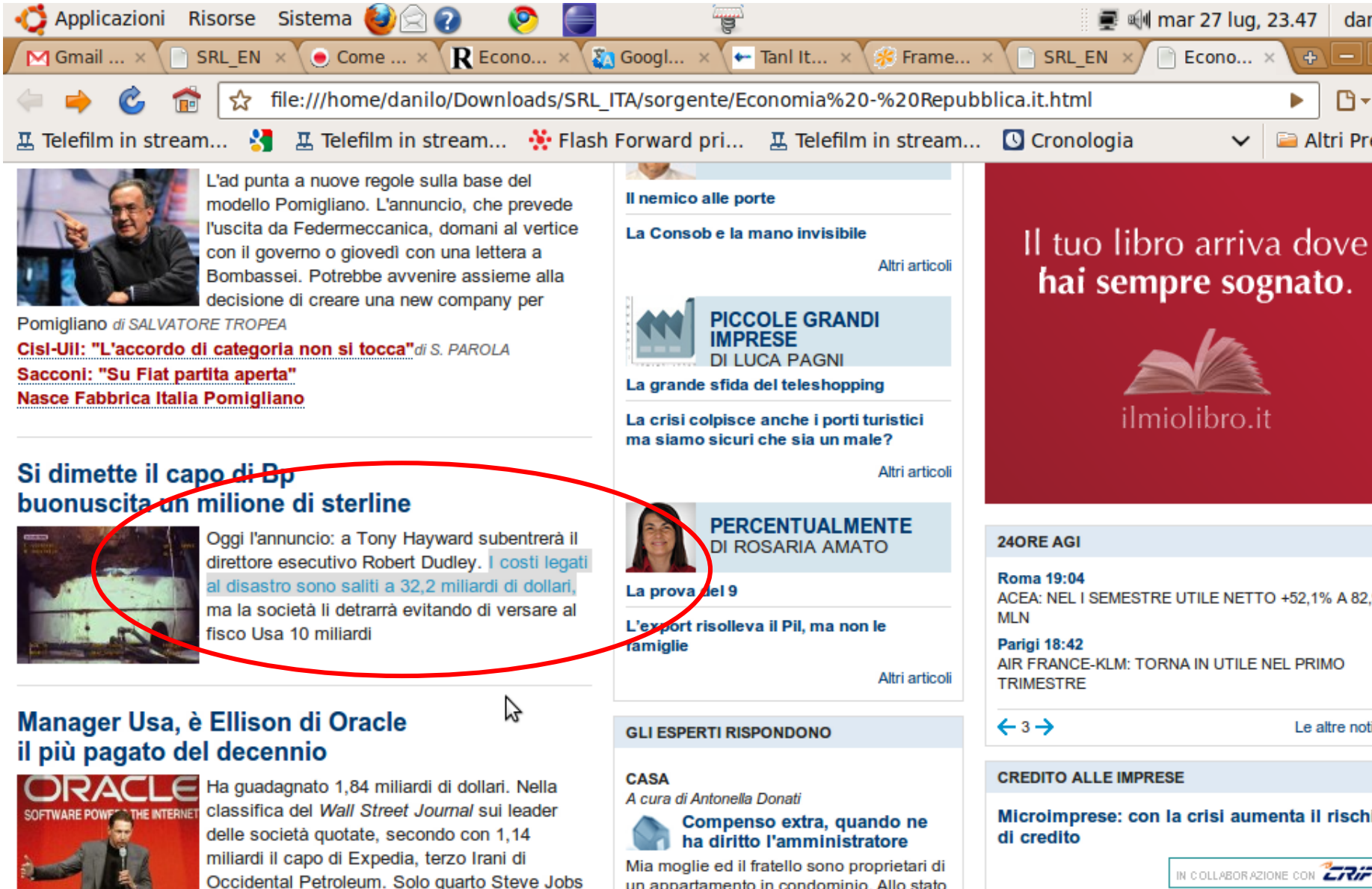
Data Mining: perspectives and benefits

- Technical advantages
  - Self adaptivity to changing operational conditions (i.e. domains)
  - Better SW management and incremental maintenance
  - More flexibility for special-purpose versioning:
    - No need for re-engineering or independent software developments
    - Just new domain-specific examples are needed

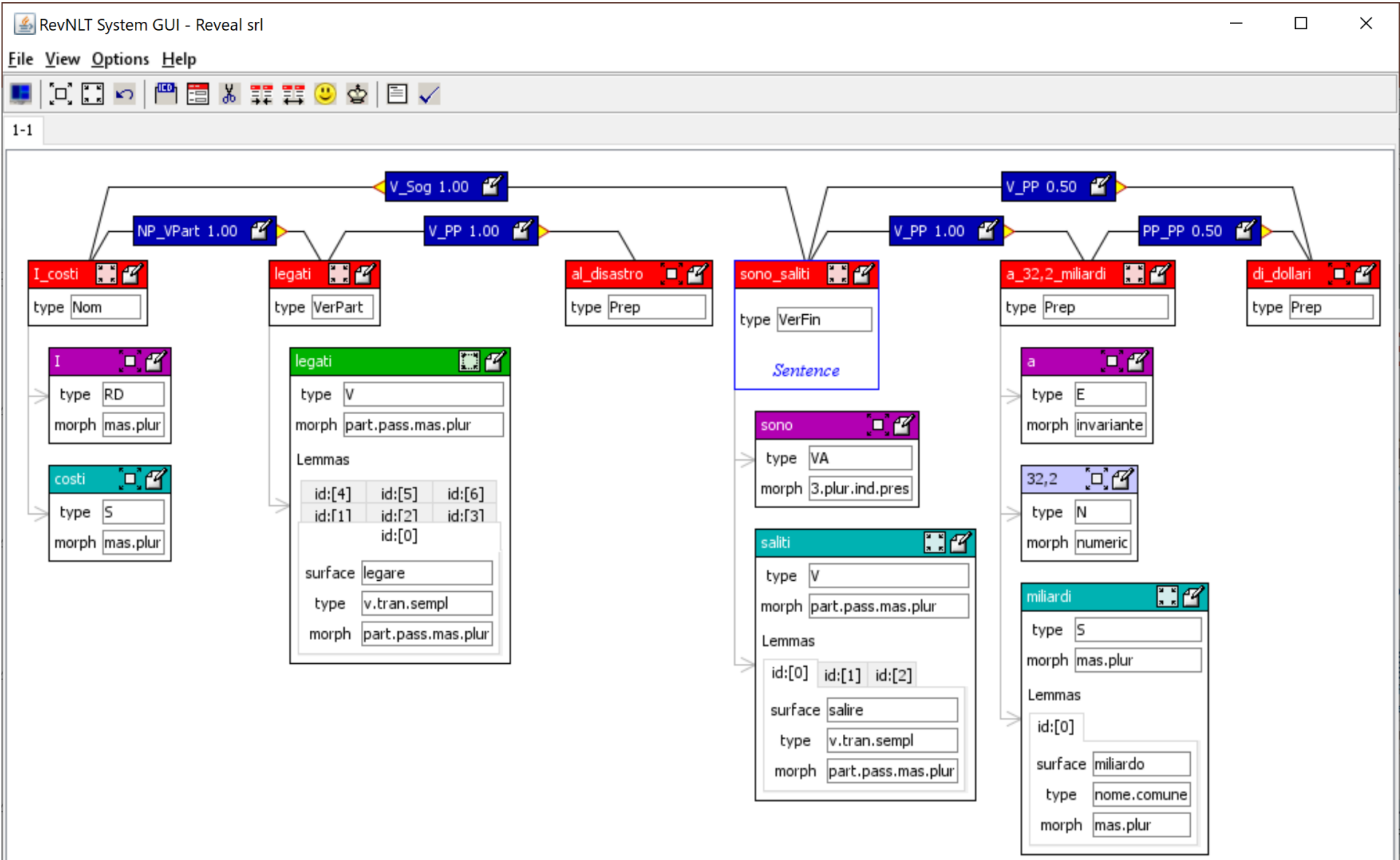
- Cost benefits
  - The data-driven approach has been shown to reduce the development costs up to 80-90% in several NLP tasks

- Market benefits
  - Reduced time-to-market
  - Competitive advantages: the lack of similar products makes the system targeted strongly competitive solutions

Semantics and News



Ouput del parser NLP



Laboratori del Corso

- Nell'ambito dei Laboratori agli studenti saranno resi disponibili:
  - Piattaforme di Machine Learning: Weka, KELP, Pytorch, SciKit
  - Motori di Ricerca: Lucene, Solr
  - Strumenti di AI per l'elaborazione dei testi:
    - Recursive Neural Networks per l'apprendimento di lessici vettoriali
    - Transformers per il semantic parsing e la textual inference
    - Parser grammaticali delle lingue (ita, eng)
    - Named Entity recognition and Wikification tools

Un esempio: Kelp: Java-based kernel framework

GitHub Search GitHub Explore Features Enterprise Pricing Sign up Sign in

KeLP

KeLP (Kernel-based Learning Platform) is a Java machine learning platform developed within the SAG group and the QCRI.

University of Roma, Tor Vergata http://sag.art.uniroma2.it/demo-software/kelp

Repositories People 3

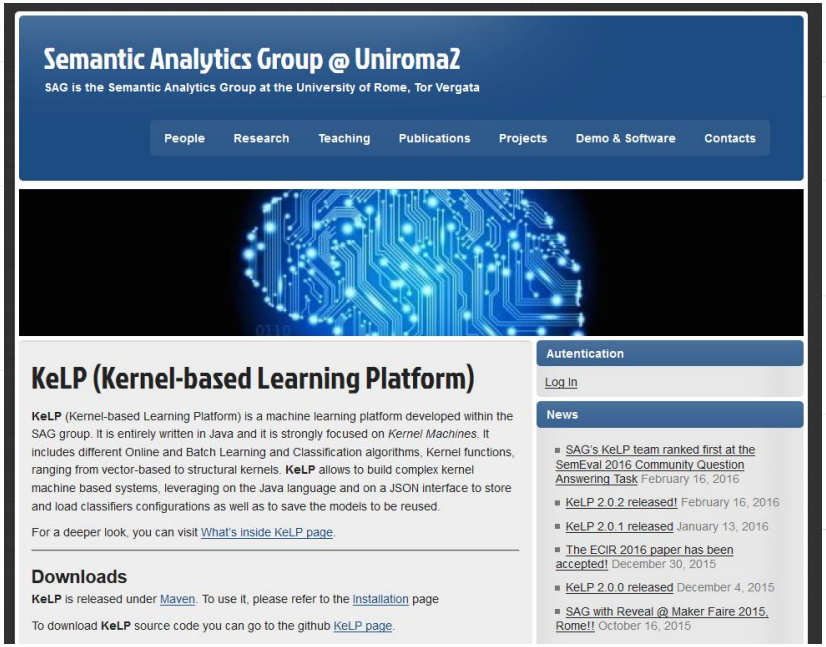
Filters Find a repository...

kelp-additional-algorithms

Updated 9 days ago

kelp-full

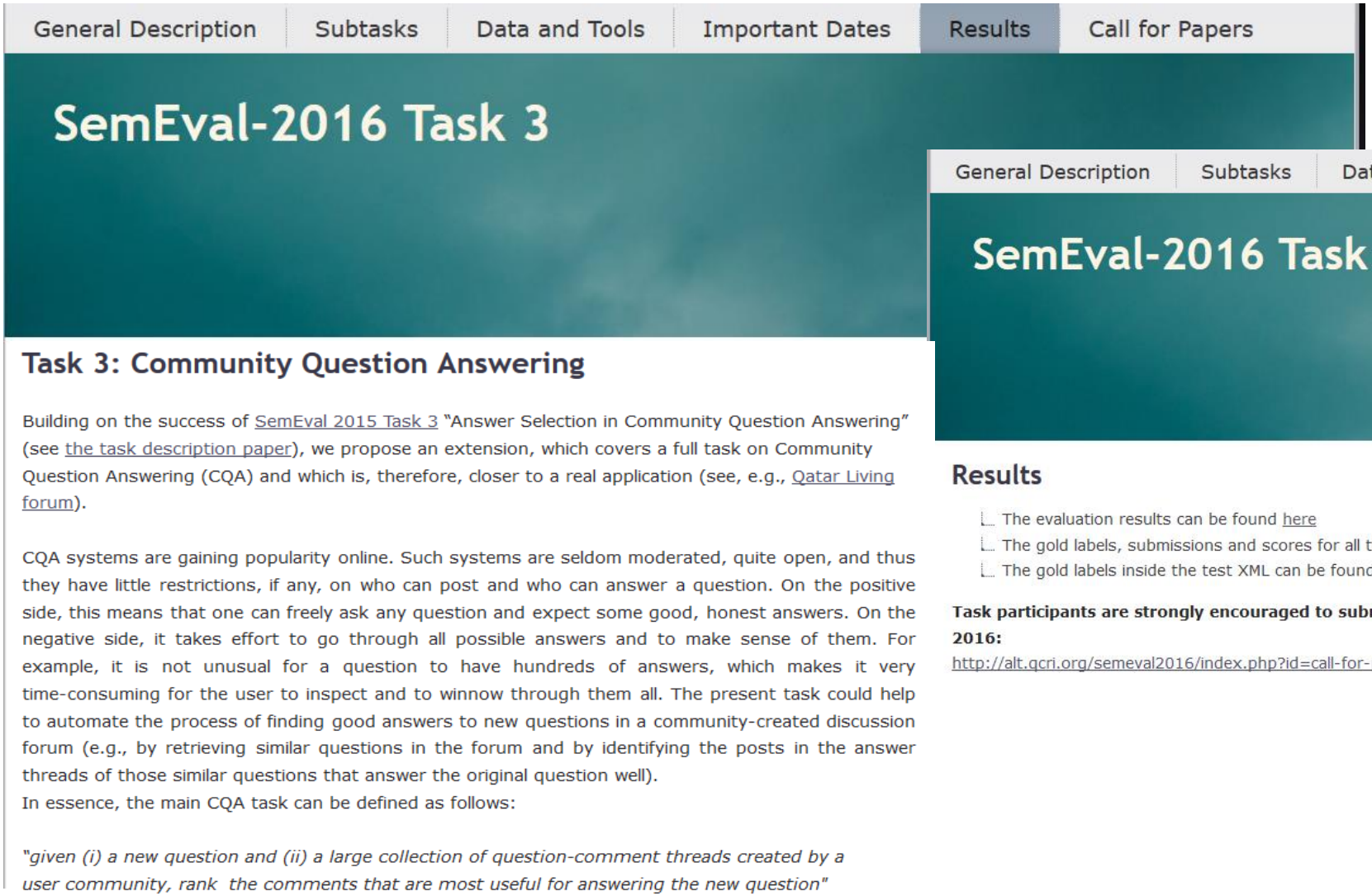
Updated 12 days ago



https://github.com/SAG-KeLP

http://sag.art.uniroma2.it/demo-software/kelp/

KELP applications: cQA



General Description | Subtasks | Data and Tools | Important Dates | Results | Call for Papers

SemEval-2016 Task 3

Task 3: Community Question Answering

Building on the success of SemEval 2015 Task 3 "Answer Selection in Community Question Answering" (see the task description paper), we propose an extension, which covers a full task on Community Question Answering (CQA) and which is, therefore, closer to a real application (see, e.g., Qatar Living forum).

CQA systems are gaining popularity online. Such systems are seldom moderated, quite open, and thus they have little restrictions, if any, on who can post and who can answer a question. On the positive side, this means that one can freely ask any question and expect some good, honest answers. On the negative side, it takes effort to go through all possible answers and to make sense of them. For example, it is not unusual for a question to have hundreds of answers, which makes it very time-consuming for the user to inspect and to winnow through them all. The present task could help to automate the process of finding good answers to new questions in a community-created discussion forum (e.g., by retrieving similar questions in the forum and by identifying the posts in the answer threads of those similar questions that answer the original question well).
In essence, the main CQA task can be defined as follows:

"given (i) a new question and (ii) a large collection of question-comment threads created by a user community, rank the comments that are most useful for answering the new question"

General Description | Subtasks | Data and Tools

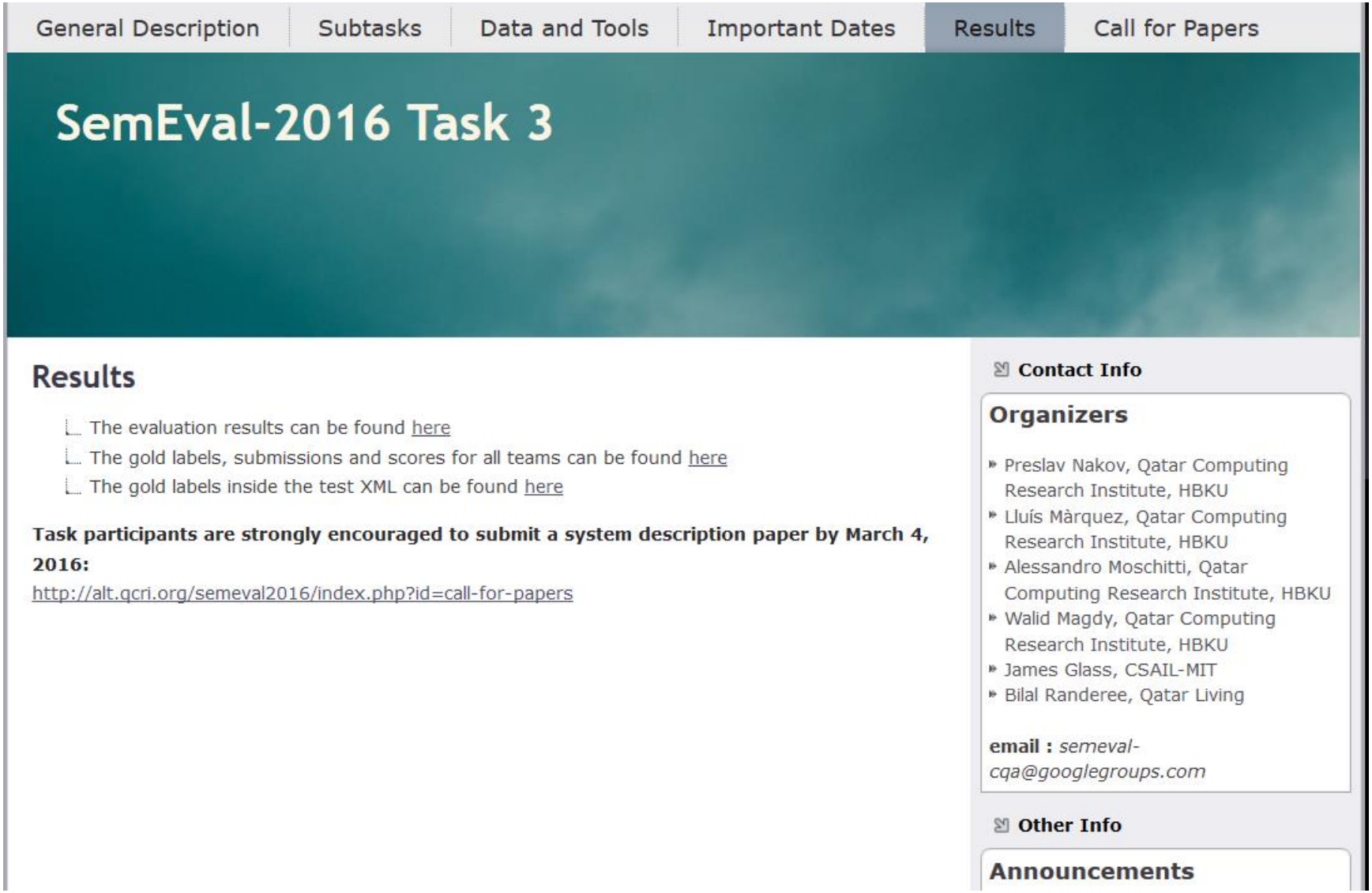
SemEval-2016 Task 3

Results

- The evaluation results can be found here
- The gold labels, submissions and scores for all teams can be foun
- The gold labels inside the test XML can be found here

Task participants are strongly encouraged to submit a system des 2016:
http://alt.qcri.org/semeval2016/index.php?id=call-for-papers

KELP applications: cQA



KELP applications: cQA

General Description | Subtasks

SemEval-2016 Tas

Team ID	Team Affiliation
ConvKN	Qatar Computing Research Institute,
ECNU	East China Normal University, Chin
ICL00	Institute of Computational Linguistic
ICRC-HIT	Intelligence Computing Research Ce
ITNLP-AiKF	Intelligence Technology and Natural
Kelp	University of Roma, Tor Vergata, Ita
MTE-NN	Qatar Computing Research Institute,
overfitting	University of Waterloo, Canada
PMI-cool	Sofia University, Bulgaria
QAIIIT	IIIT Hyderabad, India
QU-IR	Qatar University, Qatar
RDI_team	RDI Egypt, Cairo University, Egypt
SemanticZ	Sofia University, Bulgaria
SLS	MIT Computer Science and Artificia
SUper_team	Sofia University, Bulgaria; Qatar Co
UH-PRHLT	Pattern Recognition and Human Lan
	Universitat Politècnica de València;
UniMelb	The University of Melbourne, Austr
UPC_USMBA	Universitat Politècnica de Catalunya

Table 5: The particip

	Submission	MAP	AvgRec	MRR	P	R	F1	Acc
1	**Kelp-primary**	**79.19**$_1$	**88.82**$_1$	**86.42**$_1$	**76.96**$_1$	**55.30**$_8$	**64.36**$_5$	**75.11**$_2$
	ConvKN-contrastive1	78.71	88.98	86.15	77.78	53.72	63.55	74.95
	SUper_team-contrastive1	77.68	88.06	84.76	75.59	55.00	63.68	74.50
2	**ConvKN-primary**	**77.66**$_2$	**88.05**$_3$	**84.93**$_4$	**75.56**$_2$	**58.84**$_6$	**66.16**$_2$	**75.54**$_1$
3	**SemanticZ-primary**	**77.58**$_3$	**88.14**$_2$	**85.21**$_2$	**74.13**$_4$	**53.05**$_{10}$	**61.84**$_8$	**73.39**$_5$
	ConvKN-contrastive2	77.29	87.77	85.03	74.74	59.67	66.36	75.41
4	**ECNU-primary**	**77.28**$_4$	**87.52**$_5$	**84.09**$_6$	**70.46**$_6$	**63.36**$_4$	**66.72**$_1$	**74.31**$_4$
	SemanticZ-contrastive1	77.16	87.73	84.08	75.29	53.20	62.35	73.88
5	**SUper_team-primary**	**77.16**$_5$	**87.98**$_4$	**84.69**$_5$	**74.43**$_3$	**56.73**$_7$	**64.39**$_4$	**74.50**$_3$
	MTE-NN-contrastive2	76.98	86.98	85.50	58.71	70.28	63.97	67.83
	SUper_team-contrastive2	76.97	87.89	84.58	74.31	56.36	64.10	74.34
	MTE-NN-contrastive1	76.86	87.03	84.36	55.84	77.35	64.86	65.93
	SLS-contrastive2	76.71	87.17	84.38	59.45	67.95	63.41	68.13
	SLS-contrastive1	76.46	87.47	83.27	60.09	69.68	64.53	68.87
6	**MTE-NN-primary**	**76.44**$_6$	**86.74**$_7$	**84.97**$_3$	**56.28**$_9$	**76.22**$_1$	**64.75**$_3$	**66.27**$_8$
7	**SLS-primary**	**76.33**$_7$	**87.30**$_6$	**82.99**$_7$	**60.36**$_8$	**67.72**$_3$	**63.83**$_6$	**68.81**$_7$
	ECNU-contrastive2	75.71	86.14	82.53	63.60	66.67	65.10	70.95
	SemanticZ-contrastive2	75.41	86.51	82.52	73.19	50.11	59.49	72.26
	ICRC-HIT-contrastive1	73.34	84.81	79.65	63.43	69.30	66.24	71.28
8	**ITNLP-AiKF-primary**	**71.52**$_8$	**82.67**$_9$	**80.26**$_8$	**73.18**$_5$	**19.71**$_{12}$	**31.06**$_{12}$	**64.43**$_9$
	ECNU-contrastive1	71.34	83.39	78.62	66.95	41.31	51.09	67.86
9	**ICRC-HIT-primary**	**70.90**$_9$	**83.36**$_8$	**77.38**$_{10}$	**62.48**$_7$	**62.53**$_5$	**62.50**$_7$	**69.51**$_6$
10	**PMI-cool-primary**	**68.79**$_{10}$	**79.94**$_{10}$	**80.00**$_9$	**47.81**$_{12}$	**70.58**$_2$	**57.00**$_9$	**56.73**$_{12}$
	UH-PRHLT-contrastive1	67.57	79.50	77.08	54.10	50.11	52.03	62.45
11	**UH-PRHLT-primary**	**67.42**$_{11}$	**79.38**$_{11}$	**76.97**$_{11}$	**55.64**$_{10}$	**46.80**$_{11}$	**50.84**$_{11}$	**63.21**$_{10}$
	UH-PRHLT-contrastive2	67.33	79.34	76.73	54.97	49.13	51.89	62.97
12	**QAIIIT-primary**	**62.24**$_{12}$	**75.41**$_{12}$	**70.58**$_{12}$	**50.28**$_{11}$	**53.50**$_9$	**51.84**$_{10}$	**59.60**$_{11}$
	QAIIIT-contrastive2	61.93	75.22	69.95	49.48	49.96	49.72	58.93
	QAIIIT-contrastive1	61.80	75.12	69.76	49.85	50.94	50.39	59.24
	Baseline 1 (IR)	**59.53**	**72.60**	**67.83**	—	—	—	—
	Baseline 2 (random)	52.80	66.52	58.71	40.56	74.57	52.55	45.26
	Baseline 3 (all 'true')	—	—	—	40.64	100.00	**57.80**	40.64
	Baseline 4 (all 'false')	—	—	—	—	—	—	**59.36**

Table 1: **Subtask A, English (Question-Comment Similarity):** results for all submissions. The first column shows the rank of the primary runs with respect to the official MAP score. The second column contains the team's name and its submission type (primary vs. contrastive). The following columns show the results for the primary, and then for other, unofficial evaluation measures. The subindices show the rank of the primary runs with respect to the evaluation measure in the respective column.

… a further example

- EvalIta 2018:
  - IronIta: [Irony Detection in Italian Tweets (IronITA)]
  - Task B: irony type

5.2 Task B: Different types of irony

34 The best performing UNITOR team is also the only team that participated to Task B with an unconstrained run.

Table 5: Results Task B. Unconstrained runs are marked by grey background

team name	id	F1-score			
		not-iro	iro	sarc	macro
UNITOR	2	0.668	0.447	0.446	0.520
UNITOR	1	0.662	0.432	**0.459**	**0.518**
ItaliaNLP	1	0.707	0.432	0.409	0.516
ItaliaNLP	2	0.693	0.423	0.392	0.503
Aspie96	1	0.668	**0.438**	0.289	0.465
baseline-random		0.503	0.266	0.242	0.337
venses-itgetarun	1	0.431	0.260	0.018	0.236
baseline-mfc		0.668	0.000	0.000	0.223
venses-itgetarun	2	0.413	0.183	0.000	0.199

35 All participating systems show an improvement over the baselines, with the exception of the only unsupervised system (venses-itgetarun, see details in Section 6).

References

- Mitchell, Tom. M. 1997. *Machine Learning*. New York: McGraw-Hill.

- Kernel machines, neural networks and graphical models, P. Frasconi, A. Sperduti, A. Starita, Rivista AI*IA Numero speciale per i “50 anni di IA”, 2007.

- Very good video lectures by Andrew Ng (Stanford) **http://academicearth.org/courses/machine-learning**