AI6122 Text Data Management & Analysis

Lecture 00 Introduction

About me

- Dr. Sun Aixin 孙爱欣
  - Email: axsun@ntu.edu.sg
  - Subject: **[AI6122] Your subject**

  - Office: N4-02C-102
  - Phone: 6790-5139
  - Homepage: www.ntu.edu.sg/home/axsun

  - Research Interests:
    - Information Retrieval, Text Mining, Social Computing, Digital Libraries

Course Evaluation

- Directed reading/Literature review: 10%
  - **Individual**
  - Find an interesting topic, and a few relevant papers
  - Summarize the main ideas in the papers

- Project/Assignment: 40%
  - **Group (4 – 5 students)**
  - Work on pre-defined and self-defined problems on real dataset

- Final Exam: 50%
  - Closed book exam
  - 3 hours
  - All kinds of questions

Preparation

- Pre-requisites
  - Basic understanding on English grammar,
    - e.g., verb, noun phrase, preposition
  - Basic algorithm and data structure analysis,
    - e.g., dynamic programming
  - Basic probability concepts,
    - e.g., conditional probability
  - **Decent** programming skills

My house is on top of that hill.
Possessive pronoun, noun, verb, preposition, noun, preposition, determiner, noun
Noun phrase, verb phrase, prepositional phrase

$$\boxed{P(B|A) = \frac{P(A,B)}{P(A)}}$$

Divide & Conquer	Dynamic Programming
Partitions a problem into independent smaller sub-problems	Partitions a problem into overlapping sub-problems
Doesn't store solutions of sub-problems. (Identical sub-problems may arise – results in the same computations are performed repeatedly)	Stores solutions of sub-problems: thus avoids calculations of same quantity twice.
Top down algorithms: which logically progresses from the initial instance down to the smallest sub-instances via intermediate sub-instances.	Bottom up algorithms: in which the smallest sub-problems are explicitly solved first and the results of these used to construct solutions to progressively larger sub-instances.

Preparation (Cont'd)

Machine learning?

- Machine learning knowledge can be very **helpful** for assignment and some parts of lecture
- Not everyone has the same skills
  - Assumes some ability to learn missing knowledge

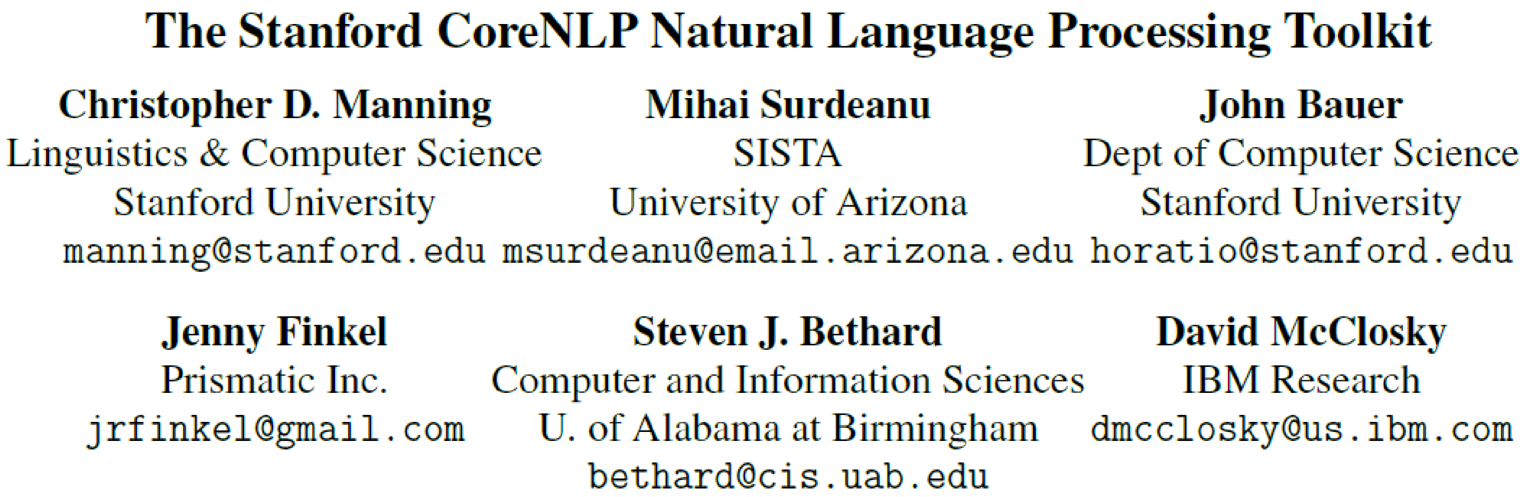
What computation?

- Some statistics!
- Some rules based on linguistic theory

What to be covered (NLP + IR)

- This course covers **fundamental** techniques to manage and process **text** data. This course does NOT cover deep learning
  - Some of the concepts will be linked to deep learning topics
- Text indexing and search
  - inverted index, query processing, ranking, and evaluation,
- Word-level, sentence-level, document-level, and collection-level processing
  - morphological analysis, part-of-speech tagging, parsing, summarization, classification and clustering, and topic modeling,
- Case studies and applications
  - social media text, sentiment analysis, and information extraction.

Why these topics



The Stanford CoreNLP Natural Language Processing Toolkit

Christopher D. Manning
Linguistics & Computer Science
Stanford University
manning@stanford.edu

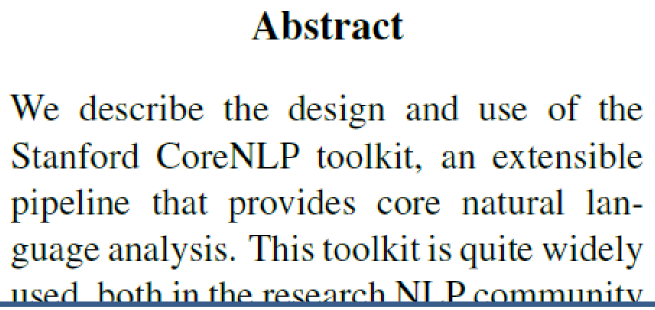
Mihai Surdeanu
SISTA
University of Arizona
msurdeanu@email.arizona.edu

John Bauer
Dept of Computer Science
Stanford University
horatio@stanford.edu

Jenny Finkel
Prismatic Inc.
jrfinkel@gmail.com

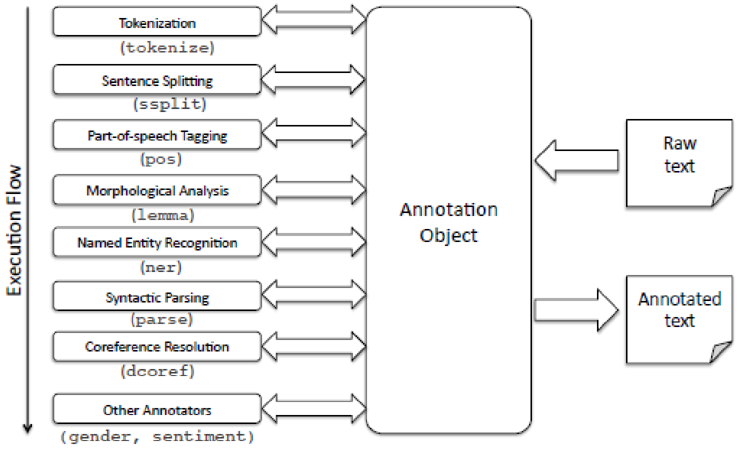
Steven J. Bethard
Computer and Information Sciences
U. of Alabama at Birmingham
bethard@cis.uab.edu

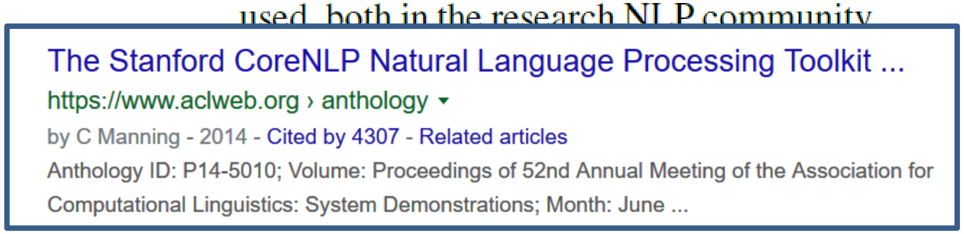
David McClosky
IBM Research
dmcclosky@us.ibm.com



Abstract

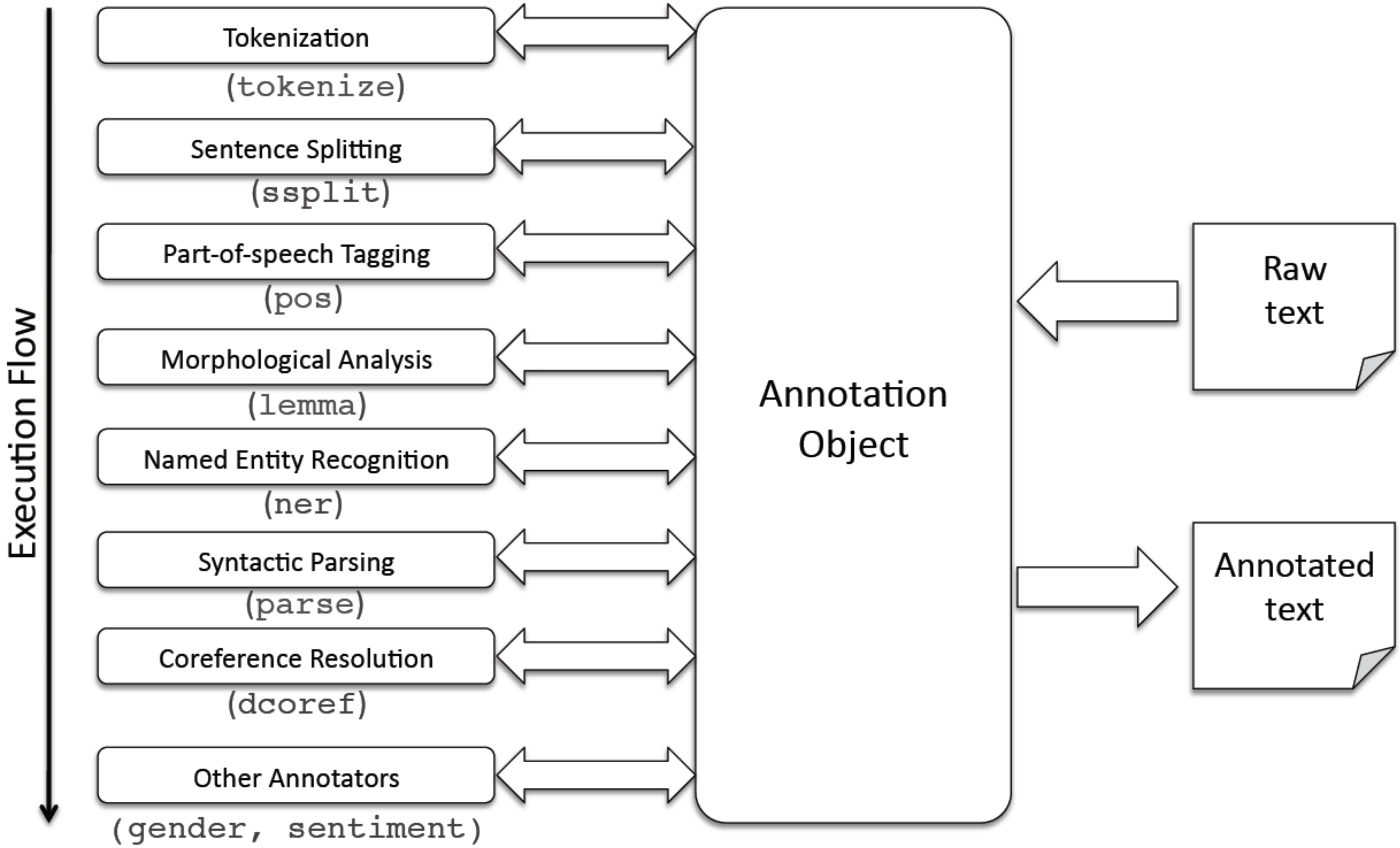
We describe the design and use of the Stanford CoreNLP toolkit, an extensible pipeline that provides core natural language analysis. This toolkit is quite widely used, both in the research NLP community



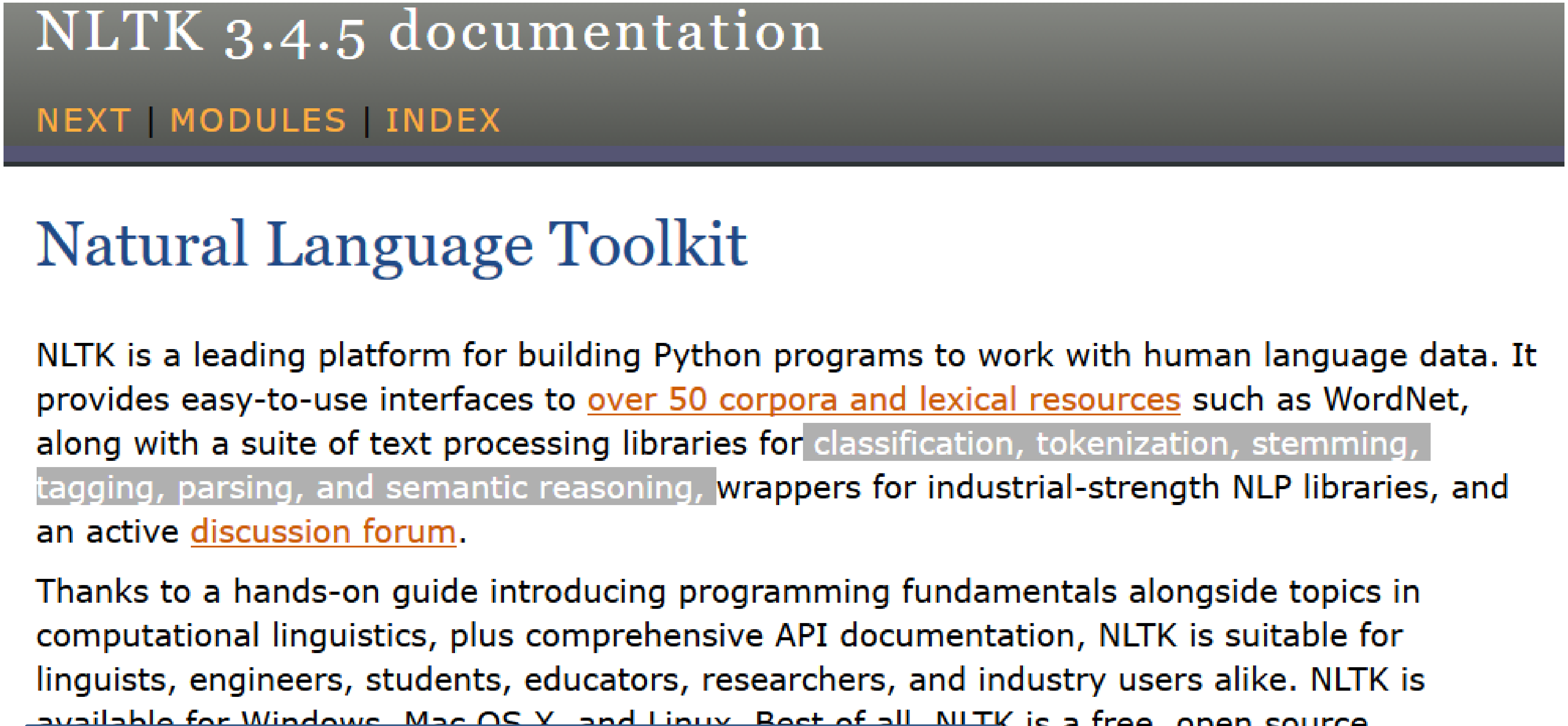


The Stanford CoreNLP Natural Language Processing Toolkit ...
https://www.aclweb.org › anthology ▾
by C Manning - 2014 - Cited by 4307 - Related articles
Anthology ID: P14-5010; Volume: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations; Month: June ...

Stanford CoreNLP



The NLTK toolkit

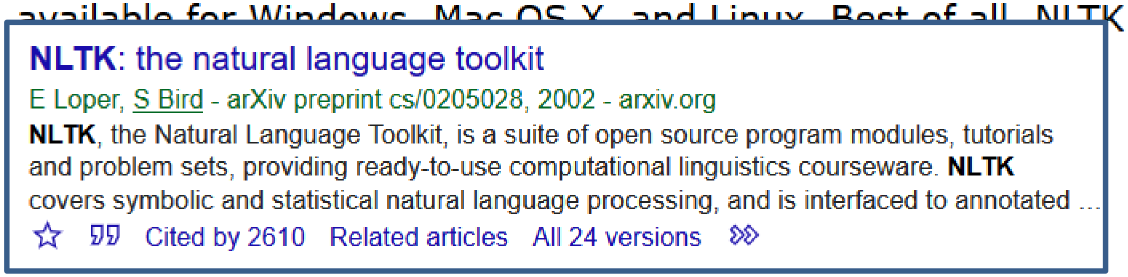


NLTK 3.4.5 documentation

Natural Language Toolkit

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, plus comprehensive API documentation, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source,

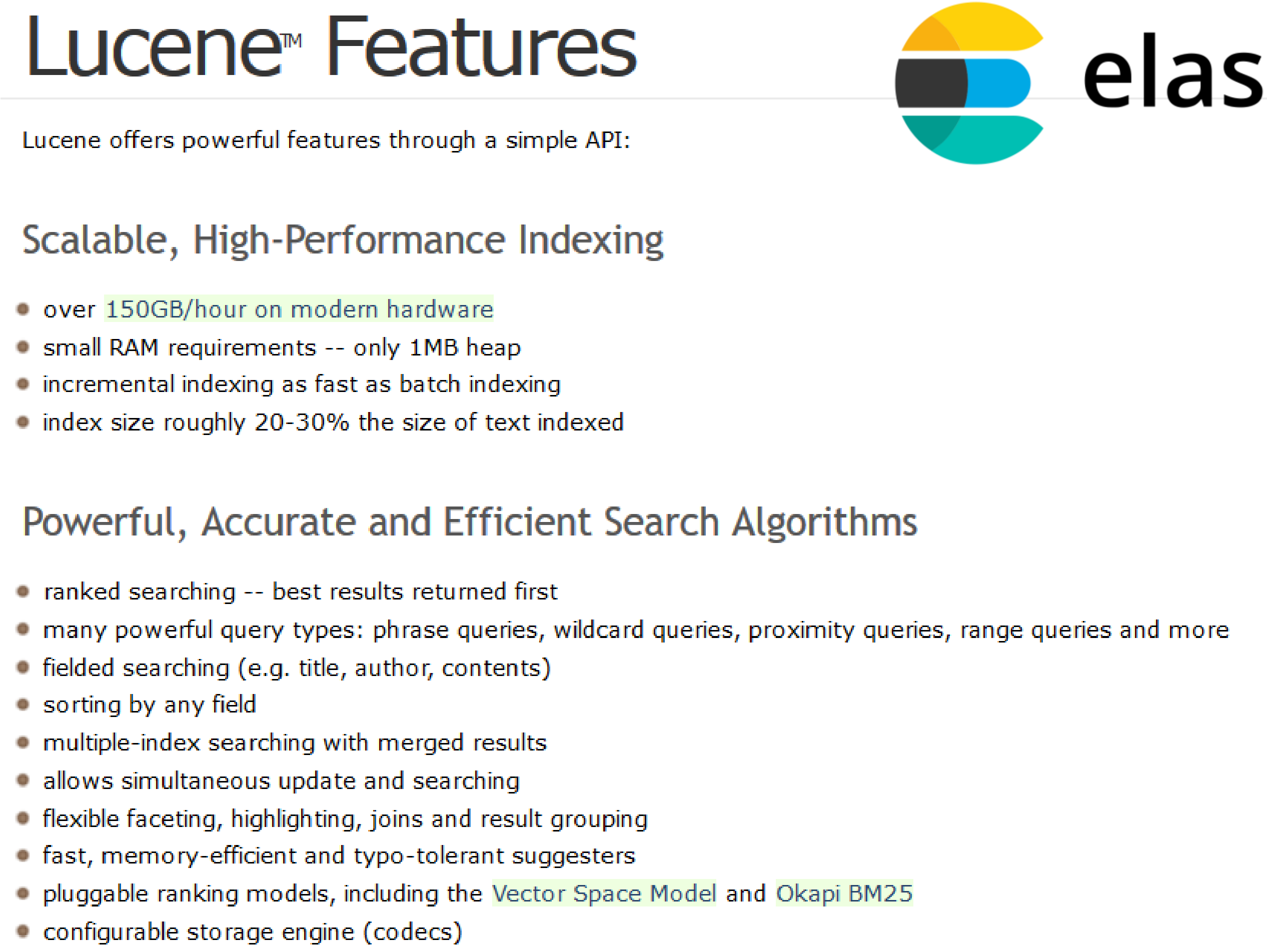


NLTK: the natural language toolkit

E Loper, S Bird - arXiv preprint cs/0205028, 2002 - arxiv.org

NLTK, the Natural Language Toolkit, is a suite of open source program modules, tutorials and problem sets, providing ready-to-use computational linguistics courseware. **NLTK** covers symbolic and statistical natural language processing, and is interfaced to annotated ...

Lucene



Lucene™ Features

Lucene offers powerful features through a simple API:

Scalable, High-Performance Indexing

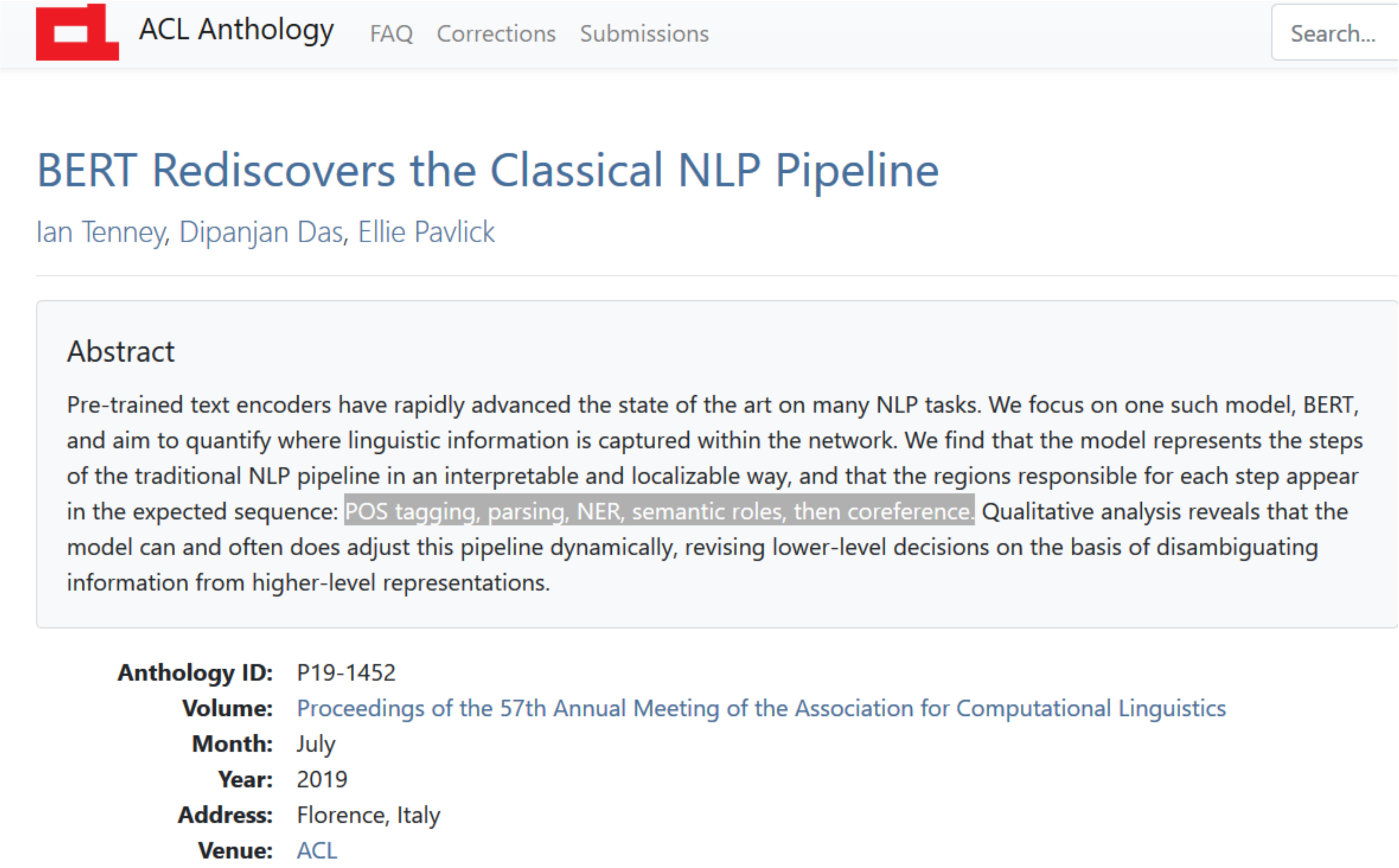
- over 150GB/hour on modern hardware
- small RAM requirements -- only 1MB heap
- incremental indexing as fast as batch indexing
- index size roughly 20-30% the size of text indexed

Powerful, Accurate and Efficient Search Algorithms

- ranked searching -- best results returned first
- many powerful query types: phrase queries, wildcard queries, proximity queries, range queries and more
- fielded searching (e.g. title, author, contents)
- sorting by any field
- multiple-index searching with merged results
- allows simultaneous update and searching
- flexible faceting, highlighting, joins and result grouping
- fast, memory-efficient and typo-tolerant suggesters
- pluggable ranking models, including the Vector Space Model and Okapi BM25
- configurable storage engine (codecs)

elasticsearch

Are these topics linked to the trending things?



BERT Rediscovers the Classical NLP Pipeline

Ian Tenney, Dipanjan Das, Ellie Pavlick

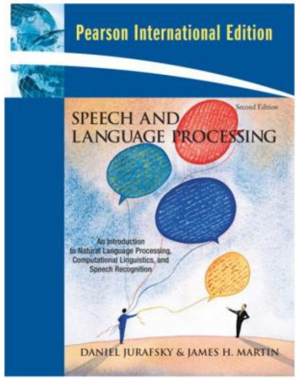
Abstract

Pre-trained text encoders have rapidly advanced the state of the art on many NLP tasks. We focus on one such model, BERT, and aim to quantify where linguistic information is captured within the network. We find that the model represents the steps of the traditional NLP pipeline in an interpretable and localizable way, and that the regions responsible for each step appear in the expected sequence: POS tagging, parsing, NER, semantic roles, then coreference. Qualitative analysis reveals that the model can and often does adjust this pipeline dynamically, revising lower-level decisions on the basis of disambiguating information from higher-level representations.

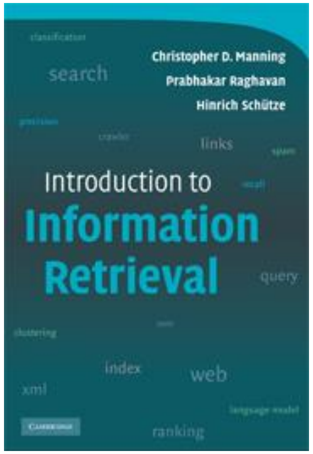
Anthology ID:	P19-1452
Volume:	Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics
Month:	July
Year:	2019
Address:	Florence, Italy
Venue:	ACL

Reference books

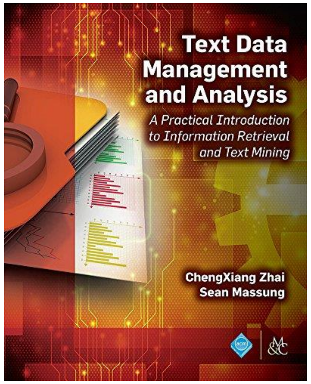
- Speech and Language Processing
  - Daniel Jurafsky and James H. Martin, 2nd edition, 2009
  - Draft of the 3rd edition: https://web.stanford.edu/~jurafsky/slp3/


Pearson International Edition
Second Edition
SPEECH AND LANGUAGE PROCESSING
An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition
DANIEL JURAFSKY & JAMES H. MARTIN

- Introduction to Information Retrieval
  - Christopher D. Manning, Prabhakar Raghavan, Hinrich Schutze
  - Cambridge University Press. 2008.
  - http://nlp.stanford.edu/IR-book/



- Text Data Management and Analysis
  - ChengXiang Zhai, Sean Massung
  - ACM and Morgan & Claypool Publishers, July 2016.


Text Data Management and Analysis
A Practical Introduction to Information Retrieval and Text Mining
ChengXiang Zhai
Sean Massung

Some of the slides are adopted from these books/authors

Course Web Page

- The course web page can be found at **NTULearn**

- It will have the lecture notes, announcements, etc.
  - Slides cannot replace the textbook.
  - They are at most a guideline.

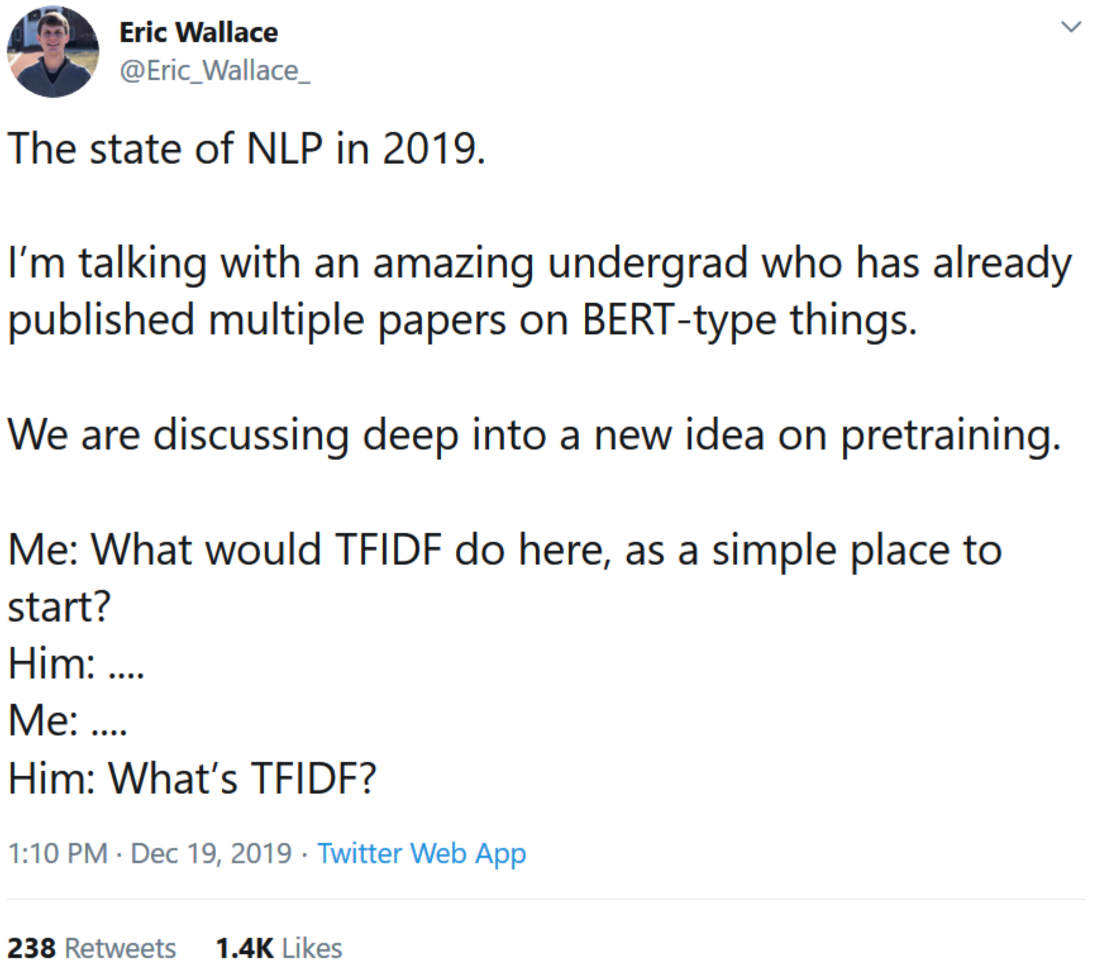
Expectations

- You are **willing to learn NLP and IR**
  - for text data management & analysis

- You are expected to participate.

- You are expected to
  - **Read** lecture slides for reference only
  - **Read** necessary chapters in the reference books
  - **Enjoy** assignment!

Traditional techniques vs deep learning



Eric Wallace
@Eric_Wallace_

The state of NLP in 2019.

I'm talking with an amazing undergrad who has already published multiple papers on BERT-type things.

We are discussing deep into a new idea on pretraining.

Me: What would TFIDF do here, as a simple place to start?
Him: ....
Me: ....
Him: What's TFIDF?

1:10 PM · Dec 19, 2019 · Twitter Web App

238 Retweets **1.4K** Likes

Do we understand our task?

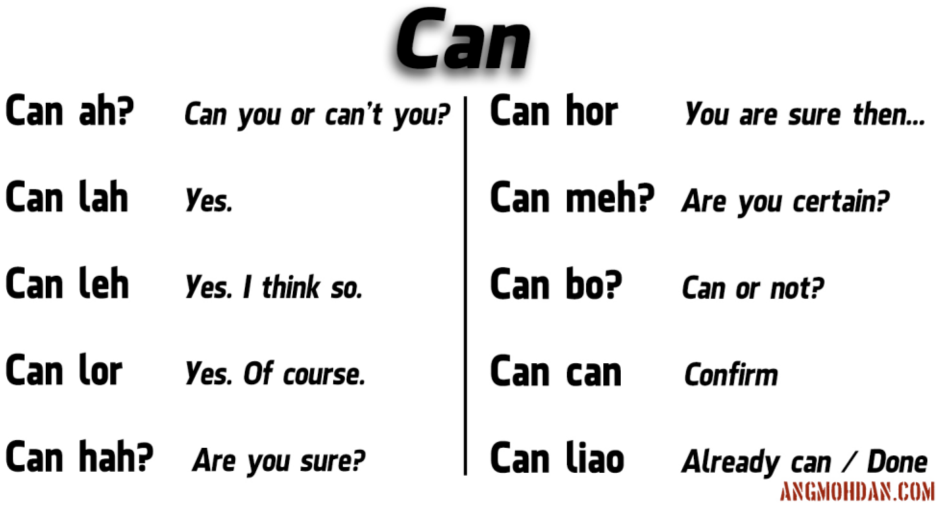
Do we understand language?

https://twitter.com/Eric_Wallace_/status/1207528697239982080

Language processing is probably hard

- We learn techniques which can be used in **practical**, robust systems that can (partly) understand human language

- This is **not** a language course
  - Computational methods of processing natural languages
  - You are expected to have knowledge of (basic) English grammar



Can

Can ah?	*Can you or can't you?*	**Can hor**	*You are sure then...*
Can lah	*Yes.*	**Can meh?**	*Are you certain?*
Can leh	*Yes. I think so.*	**Can bo?**	*Can or not?*
Can lor	*Yes. Of course.*	**Can can**	*Confirm*
Can hah?	*Are you sure?*	**Can liao**	*Already can / Done*

ANGMOHDAN.COM

Text Data Management and Processing

- Natural Language Processing (NLP)

- Information Retrieval (IR)

- Linguistics

Goals of the field of NLP

- We hope computers could
  - handle our email, do our library research, chat to us…
  - Google: google booking demo
    - https://www.youtube.com/watch?v=D5VN56jQMWM

- Then come the hard problems!
  - How can we tell computers about language?
  - Or help them learn it as kids do?

- In this course
  - We identify many open research problems in NLP
  - We aims to understand language from computing perspective

What/where is NLP?

- Goals can be very far reaching …
  - True text understanding
  - Real-time participation in spoken dialogs

- Or very down-to-earth
  - Finding the price of products on the web
  - Sentiment detection about products or stocks
  - Extracting facts or relations from documents

- These days, the latter predominate (as NLP becomes increasingly practical, it is increasingly engineering oriented)

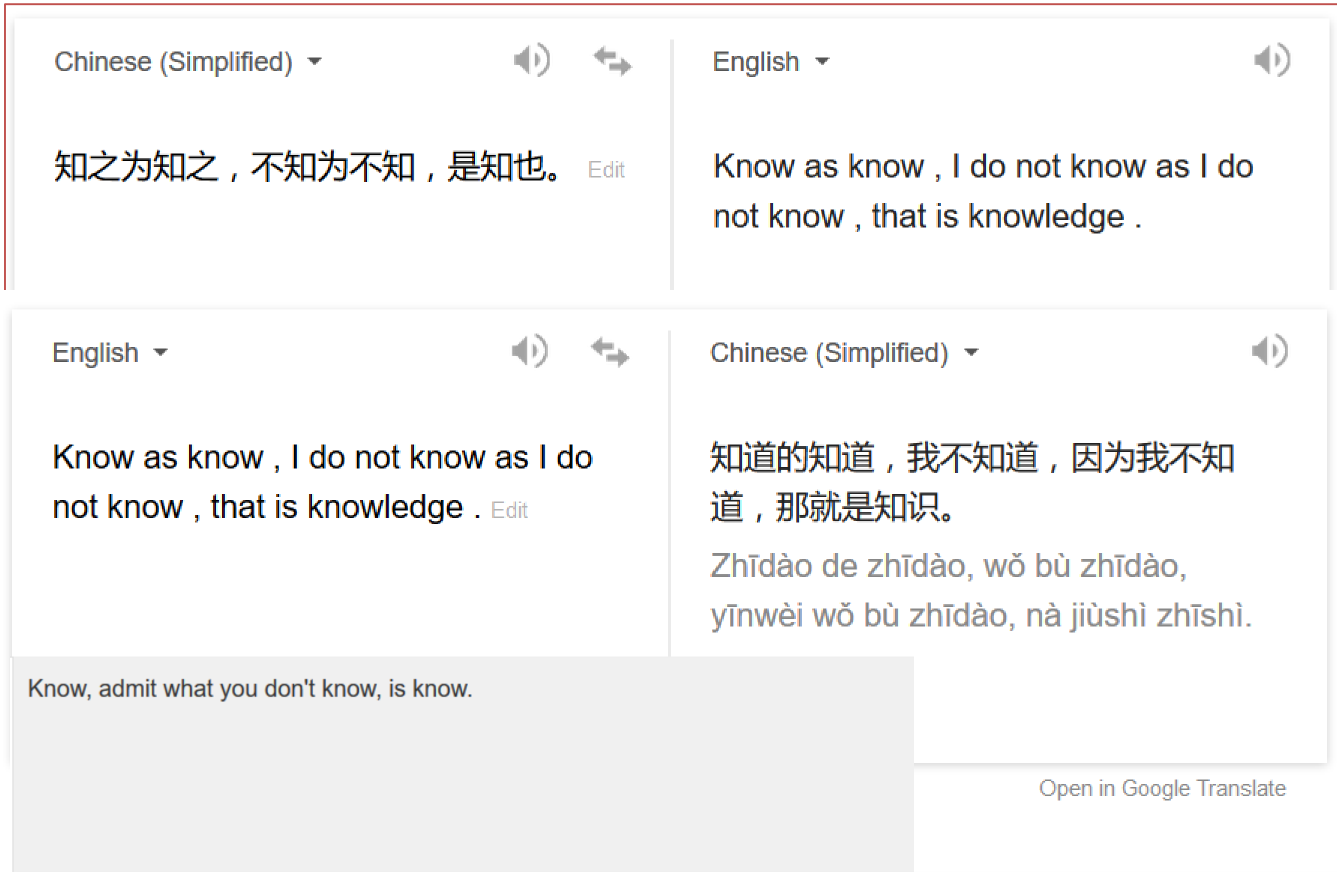
Commercial world: Lots of exciting stuff going on

Google™

Microsoft®

IBM WATSON

THOMSON REUTERS



知之为知之，不知为不知，是知也。

Know as know , I do not know as I do not know , that is knowledge .

Know as know , I do not know as I do not know , that is knowledge .

知道的知道，我不知道，因为我不知道，那就是知识。

Zhīdào de zhīdào, wǒ bù zhīdào, yīnwèi wǒ bù zhīdào, nà jiùshì zhīshì.

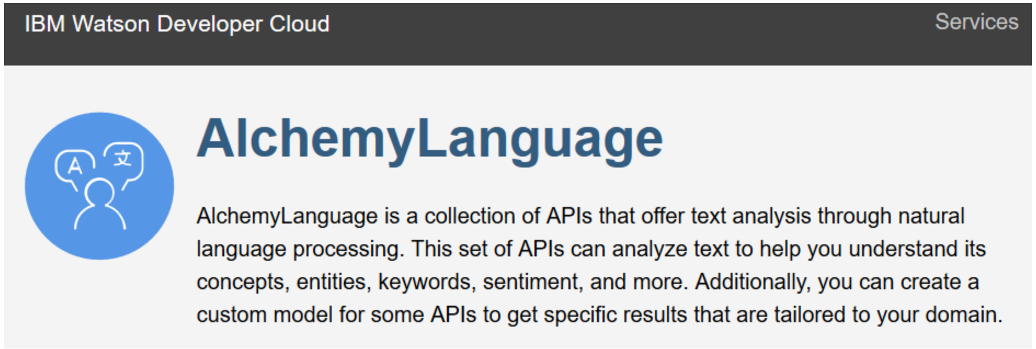
Know, admit what you don't know, is know.

> "It is wise to hold what you know and admit what you don't know."
>
> – Baidu Zhidao

Example of down-to-earth Applications

- Some deployed applications
  - Machine translation: Chinese < == > English
  - Question answering: Yahoo! Answer, Baidu Zhidao
  - Information extraction: Extracting product information from the Web
  - Text analytics: Sentiment Analysis

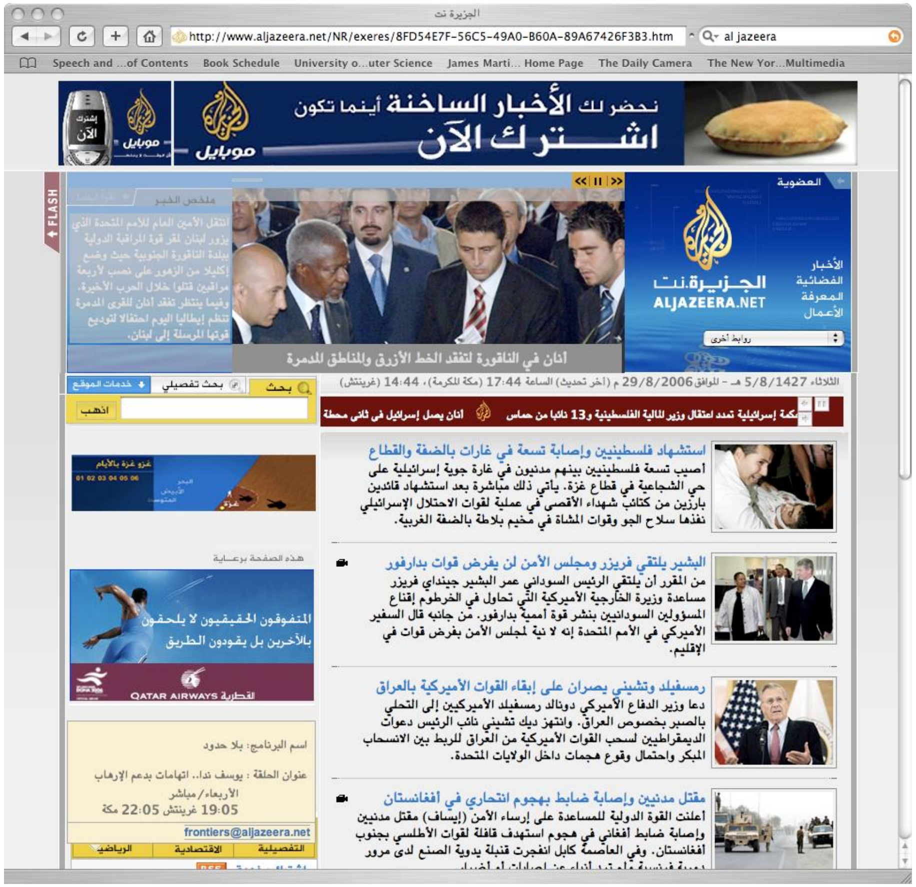
- Example https://alchemy-language-demo.mybluemix.net/



AlchemyLanguage

AlchemyLanguage is a collection of APIs that offer text analysis through natural language processing. This set of APIs can analyze text to help you understand its concepts, entities, keywords, sentiment, and more. Additionally, you can create a custom model for some APIs to get specific results that are tailored to your domain.

Google Translate



انتقل الأمين العام للأمم المتحدة الذي يزور لبنان لمقر قوة المراقبة الدولية ببلدة الناقورة الجنوبية حيث وضع إكليلا من الزهور على نصب لأربعة مراقبين قتلوا خلال الحرب الأخيرة. وفيما ينتظر تفقد أنان للقرى المدمرة تنظم إيطاليا اليوم احتفالا لتوديع قوتها المرسلة إلى لبنان.

أنان في الناقورة لتفقد الخط الأزرق والمناطق المدمرة

الثلاثاء 5/8/1427 هـ - الموافق 29/8/2006 م (آخر تحديث) الساعة 17:44 (مكة المكرمة)، 14:44 (غرينتش)

كمة إسرائيلية تمدد لاعتقال وزير المالية الفلسطينية و13 نائبا من حماس

أنان يصل إسرائيل في ثاني محطة

استشهاد فلسطينيين وإصابة تسعة في غارات بالضفة والقطاع

أصيب تسعة فلسطينيين بينهم مدنيون في غارة جوية إسرائيلية على حي الشجاعية في قطاع غزة. يأتي ذلك مباشرة بعد استشهاد قائدين بارزين من كتائب شهداء الأقصى في عملية لقوات الاحتلال الإسرائيلي نفذها سلاح الجو وقوات المشاة في مخيم بلاطة بالضفة الغربية.

البشير يلتقي فريزر ومجلس الأمن لن يفرض قوات بدارفور

من المقرر أن يلتقي الرئيس السوداني عمر البشير جينداي فريزر مساعدة وزيرة الخارجية الأميركية التي تحاول في الخرطوم إقناع المسؤولين السودانيين بنشر قوة أممية بدارفور. من جانبه قال السفير الأميركي في الأمم المتحدة إنه لا نية لمجلس الأمن بفرض قوات في الإقليم.

رمسفيلد وتشيني يصران على إبقاء القوات الأميركية بالعراق

دعا وزير الدفاع الأميركي دونالد رمسفيلد الأميركيين إلى التحلي بالصبر بخصوص العراق. وانتهز ديك تشيني نائب الرئيس دعوات الديمقراطيين لسحب القوات الأميركية من العراق للربط بين الانسحاب المبكر واحتمال وقوع هجمات داخل الولايات المتحدة.

مقتل مدنيين وإصابة ضابط بهجوم انتحاري في أفغانستان

أعلنت القوة الدولية للمساعدة على إرساء الأمن (إيساف) مقتل مدنيين وإصابة ضابط أفغاني في هجوم استهدف قافلة لقوات الأطلسي بجنوب أفغانستان. وفي العاصمة كابل انفجرت قنبلة يدوية الصنع لدى مرور دورية فرنسية ولم ترد أنباء عن إصابات أو أضرار

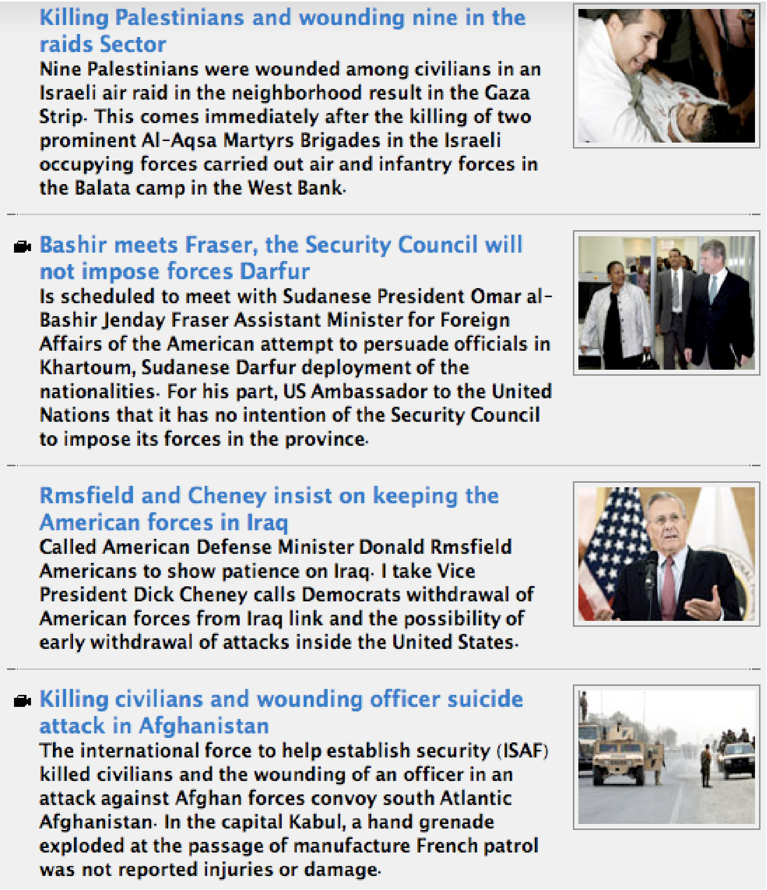
اسم البرنامج: بلا حدود

عنوان الحلقة : يوسف ندا.. اتهامات بدعم الإرهاب

الأربعاء/ مباشر

19:05 غرينتش 22:05 مكة

frontiers@aljazeera.net



Killing Palestinians and wounding nine in the raids Sector

Nine Palestinians were wounded among civilians in an Israeli air raid in the neighborhood result in the Gaza Strip. This comes immediately after the killing of two prominent Al-Aqsa Martyrs Brigades in the Israeli occupying forces carried out air and infantry forces in the Balata camp in the West Bank.

Bashir meets Fraser, the Security Council will not impose forces Darfur

Is scheduled to meet with Sudanese President Omar al-Bashir Jenday Fraser Assistant Minister for Foreign Affairs of the American attempt to persuade officials in Khartoum, Sudanese Darfur deployment of the nationalities. For his part, US Ambassador to the United Nations that it has no intention of the Security Council to impose its forces in the province.

Rmsfield and Cheney insist on keeping the American forces in Iraq

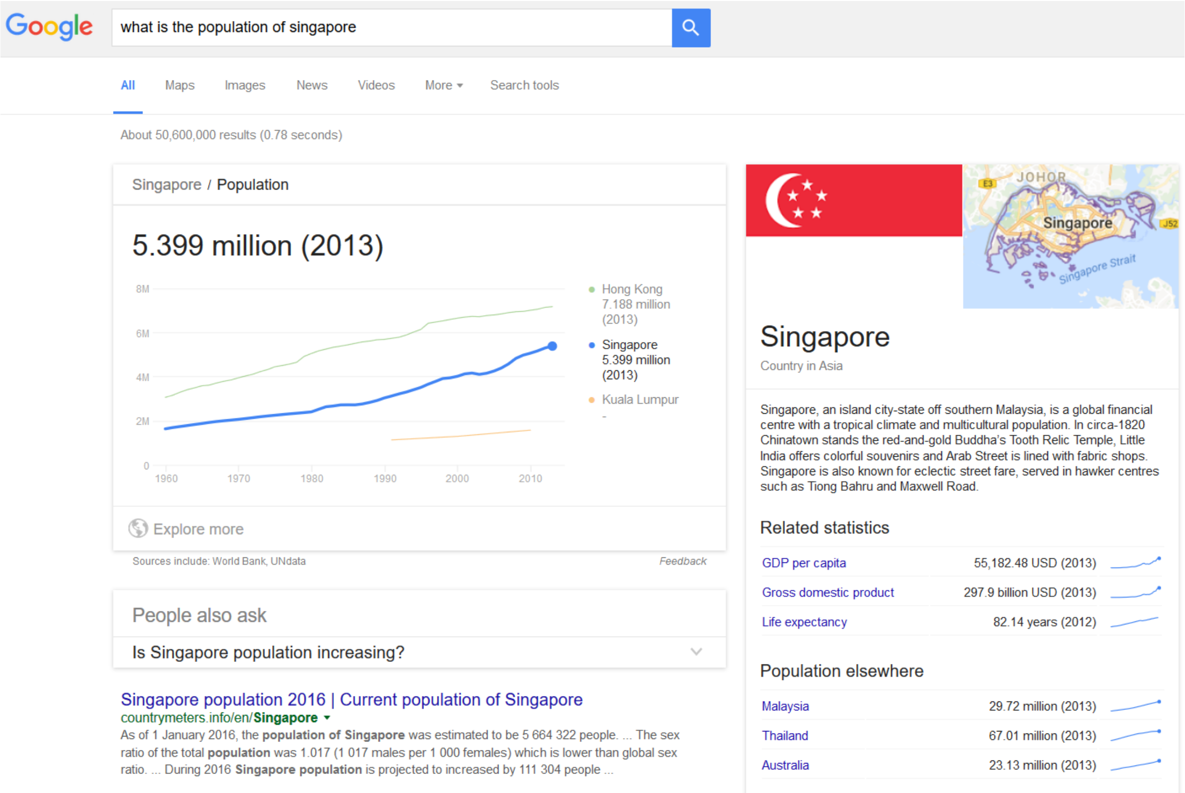
Called American Defense Minister Donald Rmsfield Americans to show patience on Iraq. I take Vice President Dick Cheney calls Democrats withdrawal of American forces from Iraq link and the possibility of early withdrawal of attacks inside the United States.

Killing civilians and wounding officer suicide attack in Afghanistan

The international force to help establish security (ISAF) killed civilians and the wounding of an officer in an attack against Afghan forces convoy south Atlantic Afghanistan. In the capital Kabul, a hand grenade exploded at the passage of manufacture French patrol was not reported injuries or damage.

Web Q/A

- Get answers directly
  - Without the need of clicking, Recommend related questions, Retrieve relevant information



Google | what is the population of singapore

All | Maps | Images | News | Videos | More ▾ | Search tools

About 50,600,000 results (0.78 seconds)

Singapore / Population

5.399 million (2013)

8M
6M
4M
2M
0
1960
1970
1980
1990
2000
2010
● Hong Kong 7.188 million (2013)
● Singapore 5.399 million (2013)
● Kuala Lumpur -

Explore more

Sources include: World Bank, UNdata

Feedback

People also ask

Is Singapore population increasing?

Singapore population 2016 | Current population of Singapore
countrymeters.info/en/**Singapore** ▾
As of 1 January 2016, the **population of Singapore** was estimated to be 5 664 322 people. ... The sex ratio of the total **population** was 1.017 (1 017 males per 1 000 females) which is lower than global sex ratio. ... During 2016 **Singapore population** is projected to increased by 111 304 people ...

Singapore
Country in Asia

Singapore, an island city-state off southern Malaysia, is a global financial centre with a tropical climate and multicultural population. In circa-1820 Chinatown stands the red-and-gold Buddha's Tooth Relic Temple, Little India offers colorful souvenirs and Arab Street is lined with fabric shops. Singapore is also known for eclectic street fare, served in hawker centres such as Tiong Bahru and Maxwell Road.

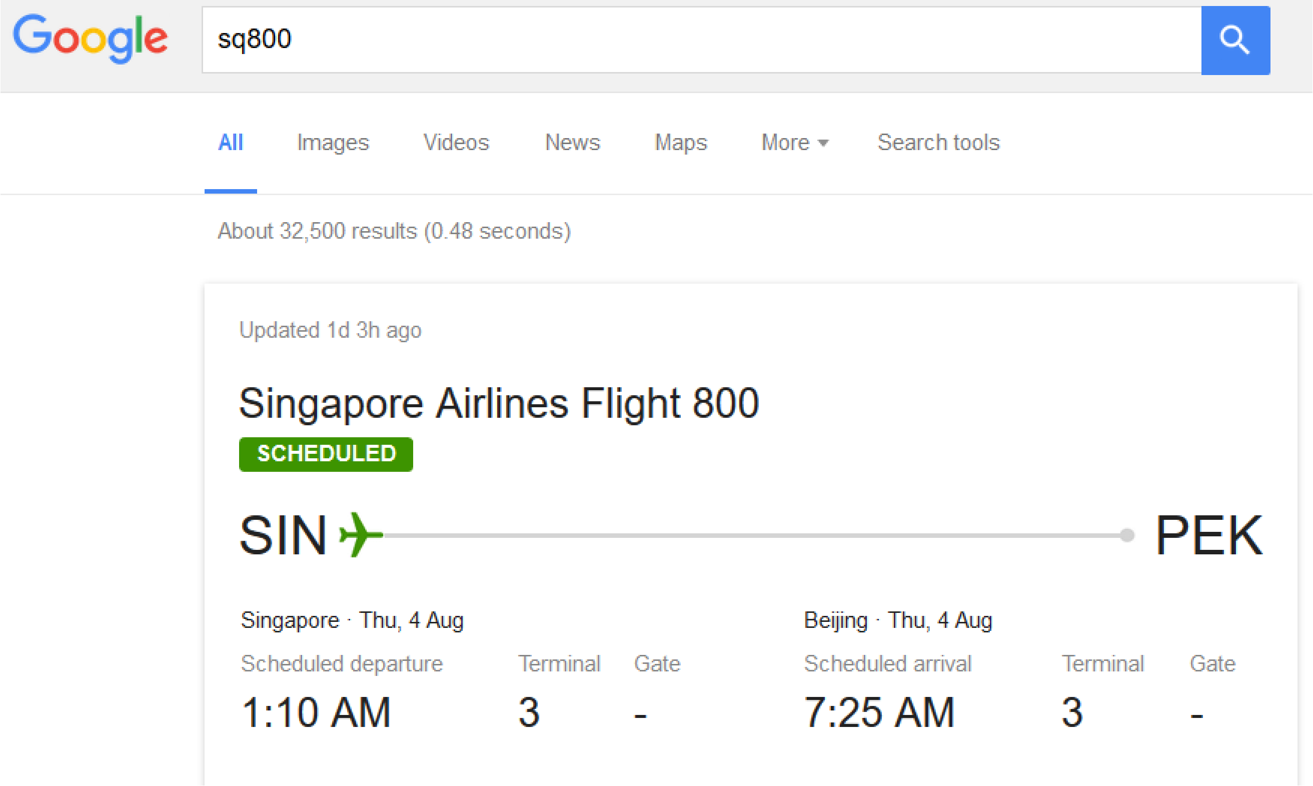
Related statistics

GDP per capita	55,182.48 USD (2013)
Gross domestic product	297.9 billion USD (2013)
Life expectancy	82.14 years (2012)

Population elsewhere

Malaysia	29.72 million (2013)
Thailand	67.01 million (2013)
Australia	23.13 million (2013)

Another example in Web search



sq800

About 32,500 results (0.48 seconds)

Updated 1d 3h ago

Singapore Airlines Flight 800

SCHEDULED

SIN ✈ PEK

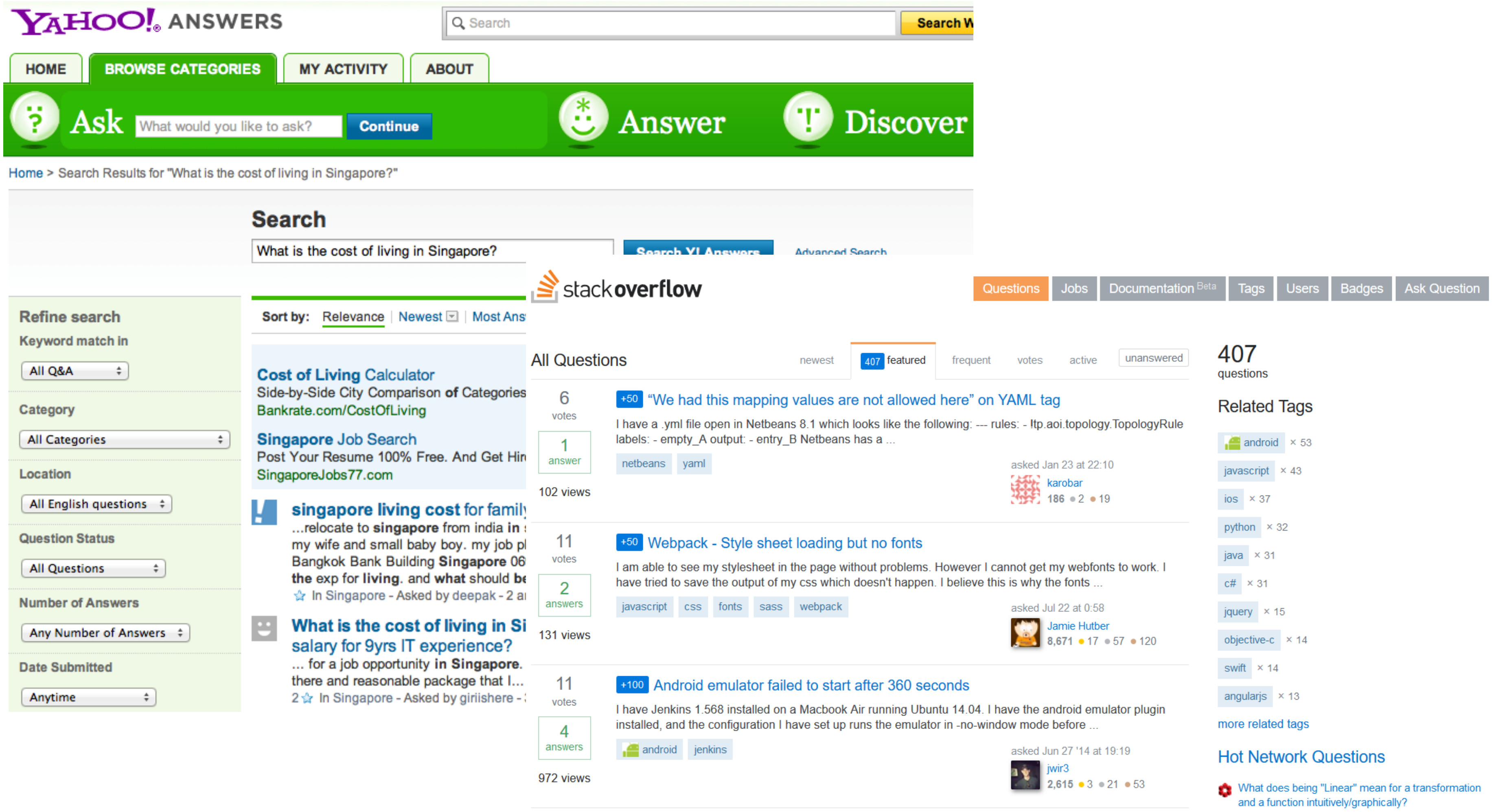
Singapore · Thu, 4 Aug

Scheduled departure	Terminal	Gate
1:10 AM	3	-

Beijing · Thu, 4 Aug

Scheduled arrival	Terminal	Gate
7:25 AM	3	-

Community based QnA



YAHOO! ANSWERS

Search | Search W

HOME | BROWSE CATEGORIES | MY ACTIVITY | ABOUT

Ask What would you like to ask? Continue | Answer | Discover

Home > Search Results for "What is the cost of living in Singapore?"

Search

What is the cost of living in Singapore? | Search Y! Answers | Advanced Search

Refine search

Keyword match in: All Q&A

Category: All Categories

Location: All English questions

Question Status: All Questions

Number of Answers: Any Number of Answers

Date Submitted: Anytime

Sort by: Relevance | Newest | Most Ans

Cost of Living Calculator
Side-by-Side City Comparison of Categories
Bankrate.com/CostOfLiving

Singapore Job Search
Post Your Resume 100% Free. And Get Hir
SingaporeJobs77.com

singapore living cost for famil
...relocate to singapore from india in s
my wife and small baby boy. my job pl
Bangkok Bank Building Singapore 06
the exp for living. and what should be
In Singapore - Asked by deepak - 2 a

What is the cost of living in Si
salary for 9yrs IT experience?
... for a job opportunity in Singapore.
there and reasonable package that I...
2 In Singapore - Asked by girishere -

stackoverflow

Questions | Jobs | Documentation Beta | Tags | Users | Badges | Ask Question

All Questions — newest | 407 featured | frequent | votes | active | unanswered

6 votes | 1 answer | 102 views
+50 "We had this mapping values are not allowed here" on YAML tag
I have a .yml file open in Netbeans 8.1 which looks like the following: --- rules: - !tp.aoi.topology.TopologyRule labels: - empty_A output: - entry_B Netbeans has a ...
netbeans yaml
asked Jan 23 at 22:10 karobar 186 ● 2 ● 19

11 votes | 2 answers | 131 views
+50 Webpack - Style sheet loading but no fonts
I am able to see my stylesheet in the page without problems. However I cannot get my webfonts to work. I have tried to save the output of my css which doesn't happen. I believe this is why the fonts ...
javascript css fonts sass webpack
asked Jul 22 at 0:58 Jamie Hutber 8,671 ● 17 ● 57 ● 120

11 votes | 4 answers | 972 views
+100 Android emulator failed to start after 360 seconds
I have Jenkins 1.568 installed on a Macbook Air running Ubuntu 14.04. I have the android emulator plugin installed, and the configuration I have set up runs the emulator in -no-window mode before ...
android jenkins
asked Jun 27 '14 at 19:19 jwir3 2,615 ● 3 ● 21 ● 53

407 questions

Related Tags

android	× 53
javascript	× 43
ios	× 37
python	× 32
java	× 31
c#	× 31
jquery	× 15
objective-c	× 14
swift	× 14
angularjs	× 13

more related tags

Hot Network Questions

What does being "Linear" mean for a transformation and a function intuitively/graphically?

The hidden structure of language

- We’re going beneath the surface…
  - Not just string processing
  - Not just keyword matching in a search engine
  - Search Google on “tennis racquet” and “tennis racquets” or “laptop” and “notebook” and the results are quite different … though these days Google does lots of subtle stuff beyond keyword matching itself

- We want to recover and manipulate at least some aspects of language structure and meaning

Example tasks (1)

- Word-level processing
  - Task 1: Locate all verbs and verbs only
    - E.g. the tower collapsed as a result of safety violations
    - *Is 'result' here a noun or a verb?*

- Syntactic processing
  - Task 2: Answer "Who killed John?"
    - E.g. "Mary killed John."
    - E.g. "John was killed by Mary."
    - E.g. "The guy who loved Mary killed John."
    - E.g. "Mary is not sure of who killed John."

    - *Hint: find subject of 'killed' whose object is 'John'*

Example tasks (2)

- Semantic processing
  - Task 3: Answer “Who killed John?”
    - E.g. Mary assassinated John.
  - Task 4: Answer “Who snores?”
    - E.g. Everyone who smokes snores and John smokes.

- Discourse analysis
  - Task 5: Answer “Who killed John?”
    - E.g. Mary threw John into sea. He drowned.

Learning Objective for the NLP topics

- You will learn natural language processing at a basic level, establishing a solid understanding on the theory of morphological, syntactic, and semantic analysis.

- With that, you will gain skills to apply the NLP techniques to real-world problems by using NLP packages and toolkits.

- Upon completion of the course, you should be able to:
  - Understand and analyze the linguistic characteristics of written English
  - Design and develop a NLP system to analyze and process a general corpus
  - Troubleshoot for domain-specific NLP applications

Caveat

- Why NLP is difficult? NLP has an AI aspect to it.
  - The language is hugely **ambiguous**
  - We don’t often come up with exact solutions/algorithms

- Example
  - Time *flies* like an arrow.
  - Fruit *flies* like a banana.

- What is “Java”?
  - https://en.wikipedia.org/wiki/Java_(disambiguation)

An archer about to launch an arrow

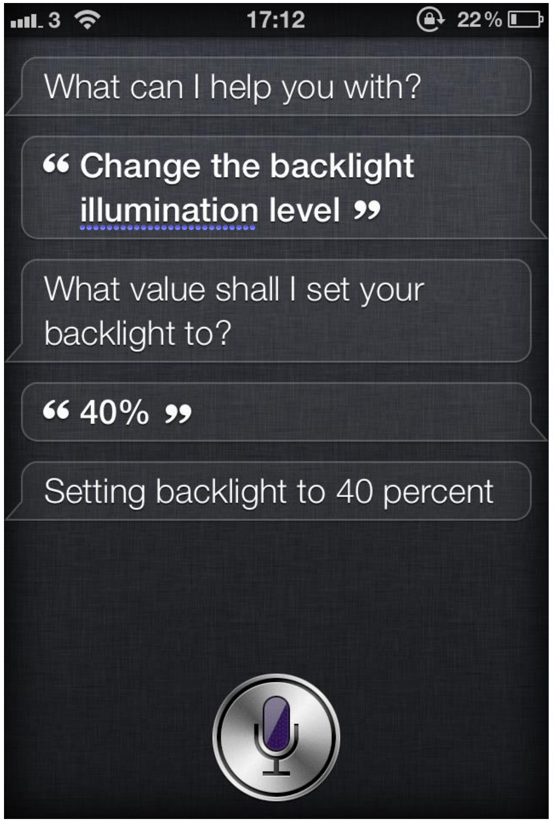
A fruit fly on a banana peel

Ambiguity is Pervasive

- Find at least 5 meanings of this sentence: **I made her duck**
  - "**duck**" (lexical category): can be a noun or verb
  - "**her**" (lexical category): can be a possessive ("of her") or dative ("for her") pronoun
  - "**make**" (lexical semantics): can mean "create" or "cook", and about 100 other things as well

✓ I cooked waterfowl for her
✓ I cooked waterfowl belonging to her
✓ I created the (plaster?) duck she owns
✓ I caused her to quickly lower her head and body
✓ I waved my magic wand and turned her into undifferentiated waterfowl

Language is still the ultimate UI (Example: Siri)



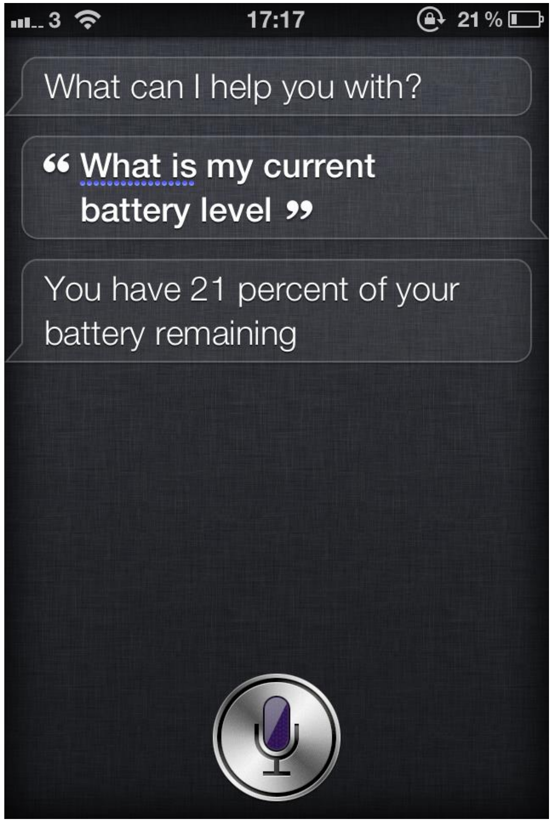
What can I help you with?

“ Change the backlight illumination level ”

What value shall I set your backlight to?

“ 40% ”

Setting backlight to 40 percent



What can I help you with?

“ What is my current battery level ”

You have 21 percent of your battery remaining

Learning Objective for the IR topics

- How to **build** your own search engine, or **customize** an existing text search engine

- How to enhance applications using IR, e.g.,
  - Cluster text-like information such as microarray data
  - Find similar actions / data / objects
  - Analyze text/dialogues (e.g., Facebook posts, Twitter, comments)

- How to build your own n-th Generation IR killer app
  - Matching people based on their preferences
  - Recommending similar products through keywords or content

This course will NOT cover

- Non-text data
  - Image
  - Video

- Semi-structured data and NoSQL databases

- Structured Data Retrieval
  - SQL

What is IR?

- What to retrieve?
  - – bookmarks like del.icio.us
  - – people, like linkedIn, facebook
  - – books (in library or on Amazon)
  - – text (web pages, medical reports, assignment reports)
  - – image (photos, flickr)
  - – video (home movies, youtube)

- Information Retrieval vs. Text Mining

What is Text Mining?

- “The objective of Text Mining is to exploit information contained in textual documents in various ways, including …discovery of patterns and trends in data, associations among entities, predictive rules, etc.”

  - Grobelnik et al., 2001

- “Another way to view text data mining is as a process of exploratory data analysis that leads to heretofore unknown information, or to answers for questions for which the answer is not currently known”

  -Hearst, 1999

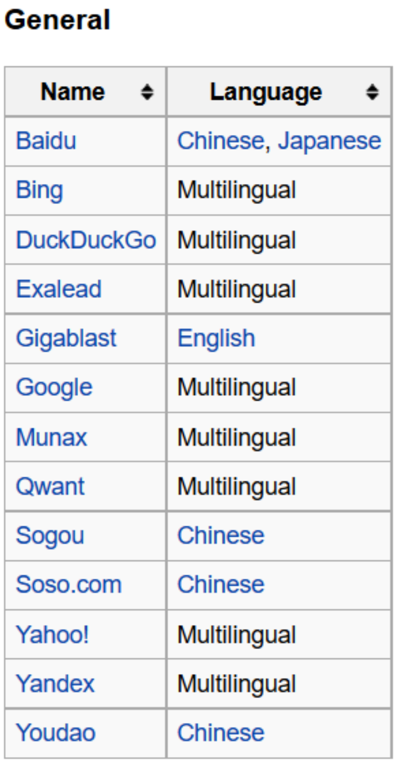
Text vs Data Mining

- When it comes to finding **novel** Nuggets, data and text mining share many of the **same techniques**

Data	Finding Patterns	Finding “Nuggets”	
		Novel	**Non-Novel**
Non-textual data	General Data Mining	Exploratory Data Analysis	Database Queries or other techniques
Textual data	Computational Linguistics		**Information Retrieval**

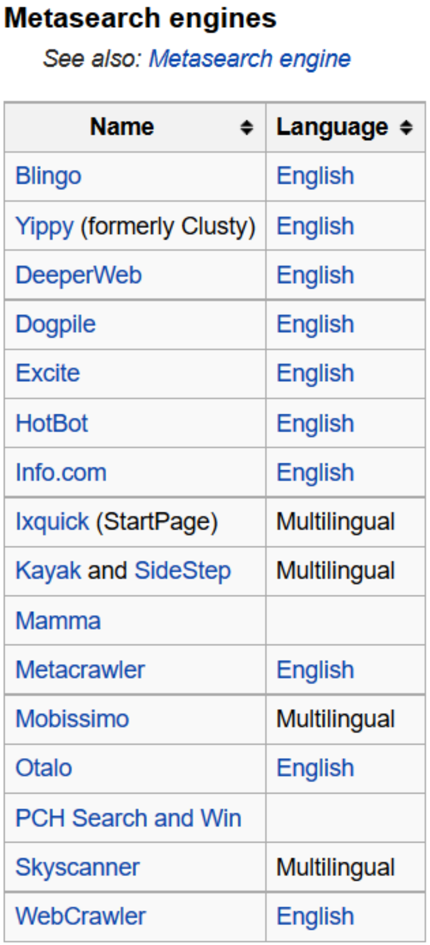
Is IR relevant to you?

- You are given a computer
  - Without Internet connection
  - With Internet connection, but
    - Search engines blocked
    - Search button blocked



General

Name	Language
Baidu	Chinese, Japanese
Bing	Multilingual
DuckDuckGo	Multilingual
Exalead	Multilingual
Gigablast	English
Google	Multilingual
Munax	Multilingual
Qwant	Multilingual
Sogou	Chinese
Soso.com	Chinese
Yahoo!	Multilingual
Yandex	Multilingual
Youdao	Chinese



Metasearch engines

See also: Metasearch engine

Name	Language
Blingo	English
Yippy (formerly Clusty)	English
DeeperWeb	English
Dogpile	English
Excite	English
HotBot	English
Info.com	English
Ixquick (StartPage)	Multilingual
Kayak and SideStep	Multilingual
Mamma	
Metacrawler	English
Mobissimo	Multilingual
Otalo	English
PCH Search and Win	
Skyscanner	Multilingual
WebCrawler	English

Text Mining Research Areas

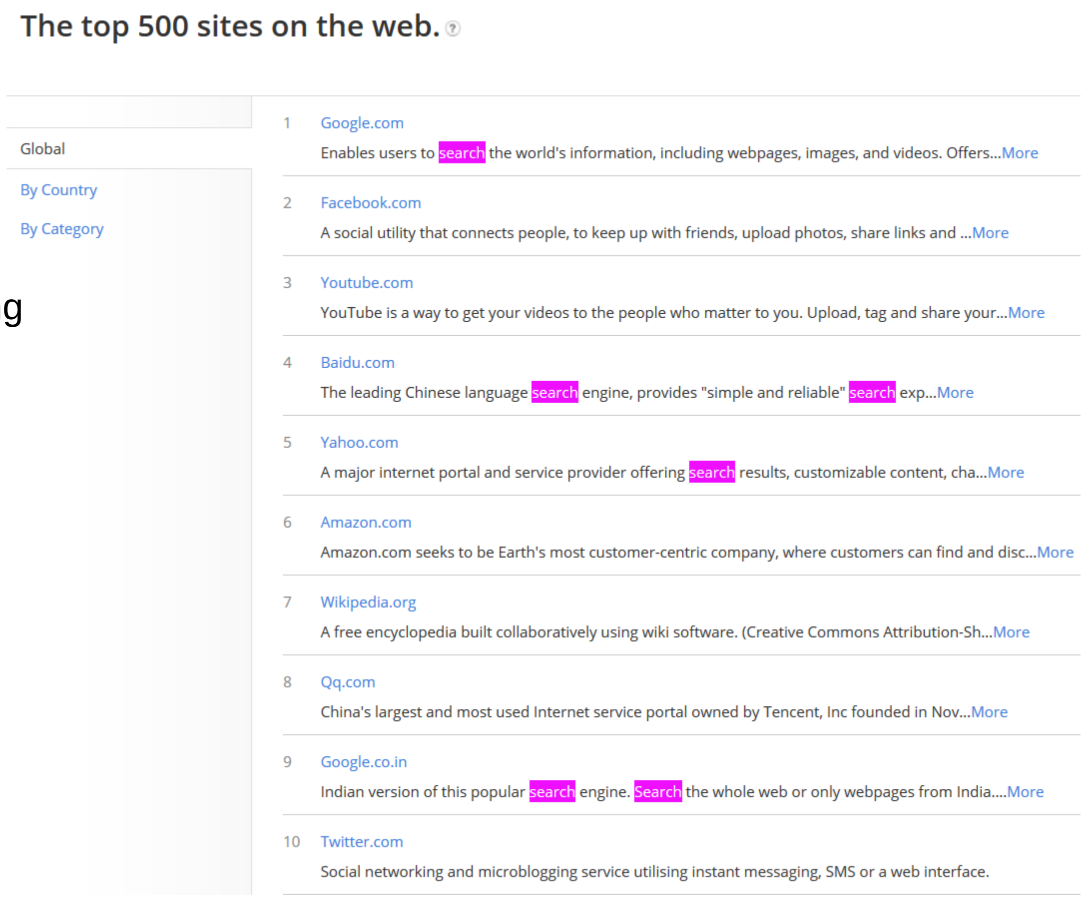
- Information Retrieval (IR)
  - Search Engines
  - Classification
  - Recommendation

- Information Extraction (IE)
  - Product Information (e.g. price) scraping
  - Name entity recognition

- Information Understanding
  - Natural Language Processing (NLP)
  - Question Answering
  - Concept Extraction from Newsgroup
  - Visualization, Summarization

- Cross-Lingual Text Mining

- Trend Detection
  - Outlier Detection
  - Event Detection



The top 500 sites on the web.

1 Google.com
Enables users to search the world's information, including webpages, images, and videos. Offers...More

2 Facebook.com
A social utility that connects people, to keep up with friends, upload photos, share links and ...More

3 Youtube.com
YouTube is a way to get your videos to the people who matter to you. Upload, tag and share your...More

4 Baidu.com
The leading Chinese language search engine, provides "simple and reliable" search exp...More

5 Yahoo.com
A major internet portal and service provider offering search results, customizable content, cha...More

6 Amazon.com
Amazon.com seeks to be Earth's most customer-centric company, where customers can find and disc...More

7 Wikipedia.org
A free encyclopedia built collaboratively using wiki software. (Creative Commons Attribution-Sh...More

8 Qq.com
China's largest and most used Internet service portal owned by Tencent, Inc founded in Nov...More

9 Google.co.in
Indian version of this popular search engine. Search the whole web or only webpages from India....More

10 Twitter.com
Social networking and microblogging service utilising instant messaging, SMS or a web interface.

http://www.alexa.com/topsites

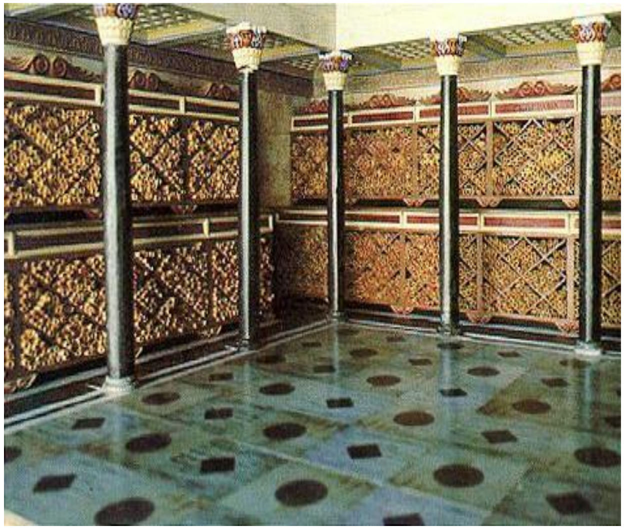
How to Retrieve Information?

- Example
  - Scan through every book in library/store bookshelf
  - View every image/video

- To speed up IR:
  - Must scan every piece of information before retrieving
    - Google/Bing tries to download the entire Web
  - **Indexing** = Scan everything = remember **where each information is located**
    - “1984” located at Level 2 Shelf 34 of National Library
    - List of documents containing “1984” stored on harddrive /dev/sda

Let’s start with some history (not covered in exam)

- 300B.C.: Great Library of Alexandria, Egypt
  - Most books stored in armaria (closed, labeled cupboards) that were still used for book storage in medieval times



Classical Indexing

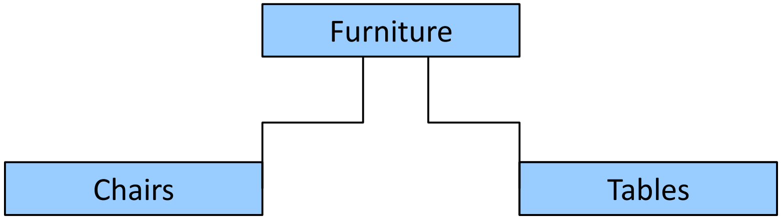
- **Indexing**
  - **Human Librarians** construct document surrogates by assigning identifiers to text items.

- Includes
  - Keyword Indexing
    - Similar to Modern Day's Search Engine Index

  - Subject Indexing
    - Similar to Modern Day's Classification Engine

Subject Indexing - Classification

- Hierarchical structure
  - Similar Subjects @ same level



- Goals of Classification
  - Collocate subjects
    - group all documents of same subject together on shelves & put them next to related subjects.
  - Define & Assign code (Call Number) to document
    - to facilitate identification from the catalogue and to shelf location

Dewey Decimal Classification (DDC)

- Most widely used
  - Used by > 135 countries

- Translated into more than 30 languages
  - Arabic, Chinese, French, Greek, Hebrew, Icelandic, Russian, Spanish.

MELVIL DEWEY.

- Universe of knowledge divided into 10 main classes.
  - Each class divided into 10 main divisions, …
  - until all disciplines, subjects and concepts are defined.

- Currently: 23rd edition (2011)

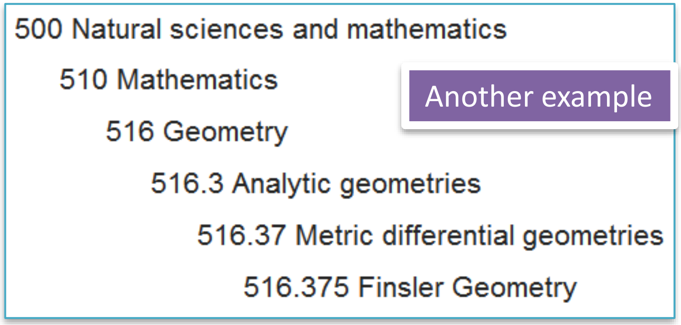
http://en.wikipedia.org/wiki/Dewey_Decimal_Classification

DDC Example

000	Generalities
100	Philosophy, paranormal phenomena, psychology
200	Religion
300	Social sciences
400	Language
500	Natural sciences and mathematics
600	Technology (Applied sciences)
700	The arts
800	Literature
900	Geography, history, and auxiliary disciplines

600	Technology (applied sciences)
610	Medical sciences
620	Engineering and allied operations
630	Agriculture and related technologies
640	Home economics and family living
650	Management and auxiliary services
660	Chemical engineering and related technologies
670	Manufactures
680	Manufacture of products for specific uses
690	Buildings

620	Engineering & allied operations
621	Applied physics
622	Mining and related operations
623	Military and nautical engineering
624	Civil engineering
625	Engineering of railroads and roads
626	[not used]
627	Hydraulic engineering
628	Sanitary and municipal engineering
629	Other branches of engineering



500 Natural sciences and mathematics

- 510 Mathematics
  - 516 Geometry
    - 516.3 Analytic geometries
      - 516.37 Metric differential geometries
        - 516.375 Finsler Geometry

Another example

DDC Pain

- DDC Classification **Guidelines**
  - **Determine** the subject of a work
  - **Determine** the disciplinary focus of a work
  - Refer to the **schedules**

- Rules to handle a document in multiple classes
  - **First-of-two Rule**: When two subjects receive equal treatment, classify the work with the subject whose number comes first in the schedules
  - **Rule of Application**: Classify a work dealing with interrelated subjects with the subject that is acted upon

Classical Indexing

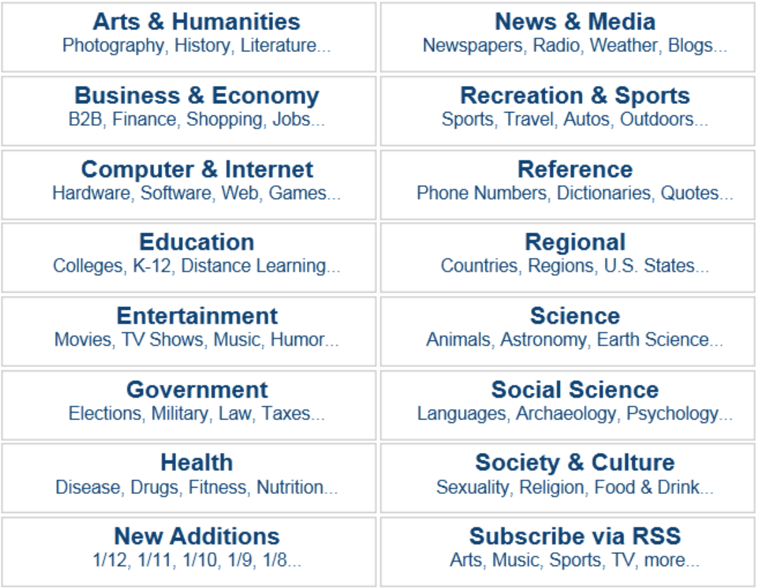
The Natural Language problem:

- **Low consistency**:
  - People use **different** words to refer to same things
  - People use same words to refer to **different** things

- Objective in IR:
  - Search & retrieval of documents (or records) require some level of intellectual control over the item and its **contents**, at the same time, recognizing the need for **flexibility**

Classical Indexing

- **Keyword** indexing (Google)
  - Index entries generated from the title and/or keywords from the text.
  - **No** intellectual process of **text analysis** or **abstraction**
- **Subject** indexing (Yahoo)
  - Involves analysis of the subject by humans / computers



Arts & Humanities Photography, History, Literature...	**News & Media** Newspapers, Radio, Weather, Blogs...
Business & Economy B2B, Finance, Shopping, Jobs...	**Recreation & Sports** Sports, Travel, Autos, Outdoors...
Computer & Internet Hardware, Software, Web, Games...	**Reference** Phone Numbers, Dictionaries, Quotes...
Education Colleges, K-12, Distance Learning...	**Regional** Countries, Regions, U.S. States...
Entertainment Movies, TV Shows, Music, Humor...	**Science** Animals, Astronomy, Earth Science...
Government Elections, Military, Law, Taxes...	**Social Science** Languages, Archaeology, Psychology...
Health Disease, Drugs, Fitness, Nutrition...	**Society & Culture** Sexuality, Religion, Food & Drink...
New Additions 1/12, 1/11, 1/10, 1/9, 1/8...	**Subscribe via RSS** Arts, Music, Sports, TV, more...

Classical Indexing Problems

Effectiveness of indexing depends on:

- Indexing **Exhaustiveness**
  - extent to which the subject matter of a given document has been reflected through the index entries

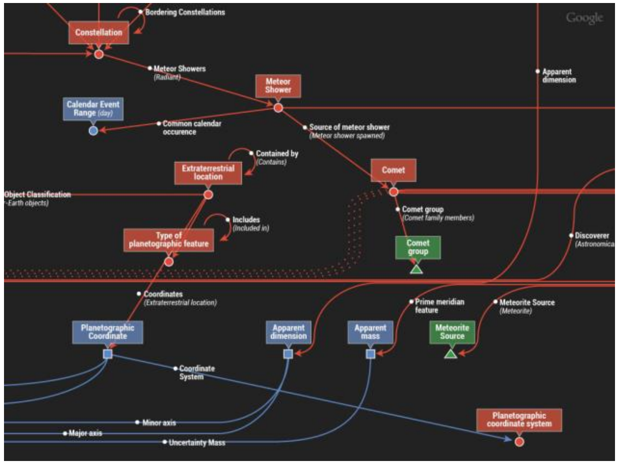
- Term **Specificity**
  - how broad/specific are the terms/keywords

Vocabulary Control: Controlled vs Natural language indexing

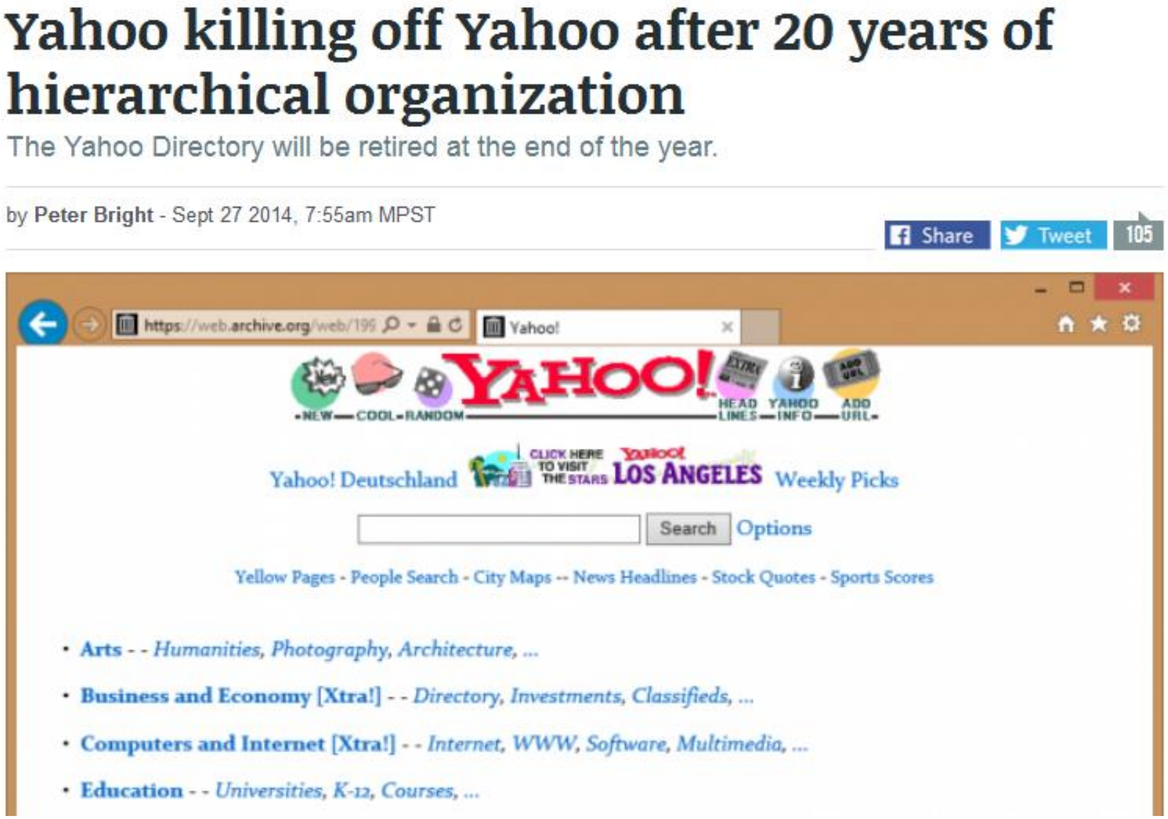
- Controlled language
  - Use of **vocabulary control** tool in indexing
  - Semantic Web
  - Dublin Core
  - XML Ontologies

- **Natural** language (free text)
  - Any term in the document may be an index term. No mechanism controls the indexing process
  - Modern Search Engine

How about Google Knowledge Graph?



Results?



Yahoo killing off Yahoo after 20 years of hierarchical organization

The Yahoo Directory will be retired at the end of the year.

by **Peter Bright** - Sept 27 2014, 7:55am MPST

Yahoo! Deutschland   Yahoo! LOS ANGELES   Weekly Picks

Search   Options

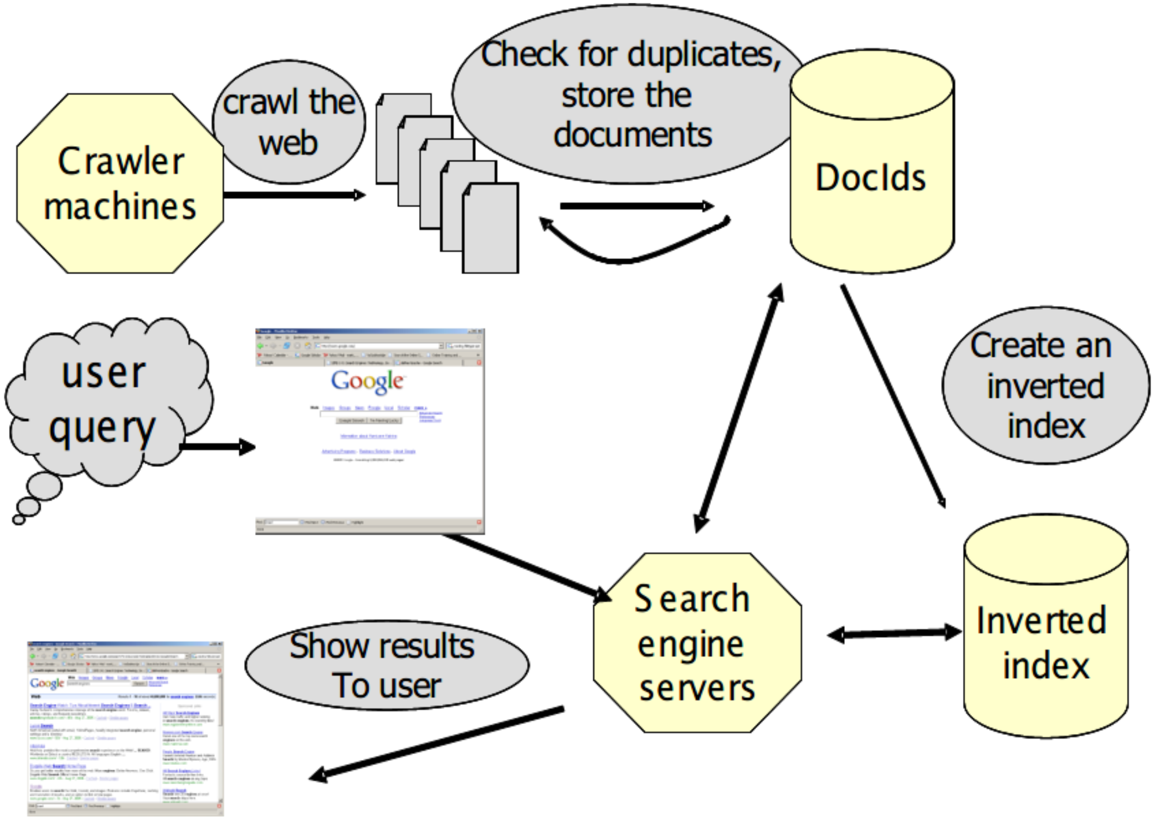
Yellow Pages - People Search - City Maps -- News Headlines - Stock Quotes - Sports Scores

- **Arts** - - *Humanities, Photography, Architecture, ...*
- **Business and Economy [Xtra!]** - - *Directory, Investments, Classifieds, ...*
- **Computers and Internet [Xtra!]** - - *Internet, WWW, Software, Multimedia, ...*
- **Education** - - *Universities, K-12, Courses, ...*

http://arstechnica.com/information-technology/2014/09/yahoo-killing-off-yahoo-after-20-years-of-hierarchical-organization/

A Modern IR System (Search Engine)

- Crawler
- Indexer
- Searcher



Crawler machines

crawl the web

Check for duplicates, store the documents

DocIds

Create an inverted index

Inverted index

Search engine servers

user query

Show results To user