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

- Quiz: 30%
  - **Individual**
  - Format: Either physical (closed book) or online (open book)
  - Question type: All kinds of questions

- Assignment: 40%
  - **Group (4 – 5 students)**
  - Work on pre-defined problems on real dataset with white space for creativity

- Directed reading and presentation: 30%
  - **Individual**
  - Find an interesting topic, and read at least 3 relevant papers
  - Summarize the main ideas in the papers and write a report
  - Record a video presentation

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

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

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

  - Programming skills (Python, Java, or others)

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 kind of computation?

- Some statistics!
- Some rules, based on linguistic theory

What to be covered (IR + NLP)

- This course covers **fundamental** techniques to manage and process **text** data. This course does NOT cover deep learning
  - AI6127 Deep Neural Networks for Natural Language Processing

- Text indexing and search
  - inverted index, query processing, ranking, and evaluation
  - (How does Google answer your queries)

- 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?

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

elasticsearch

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)

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

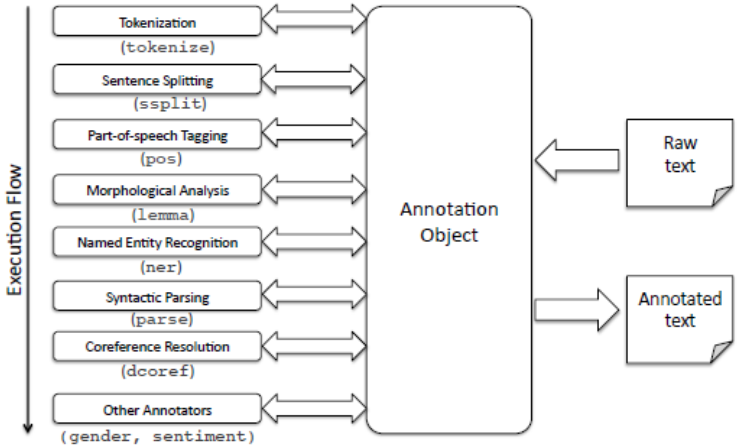
Steven J. Bethard
Computer and Information Sciences
U. of Alabama at Birmingham
bethard@cis.uab.edu

David McClosky
IBM Research
dmcclosky@us.ibm.com

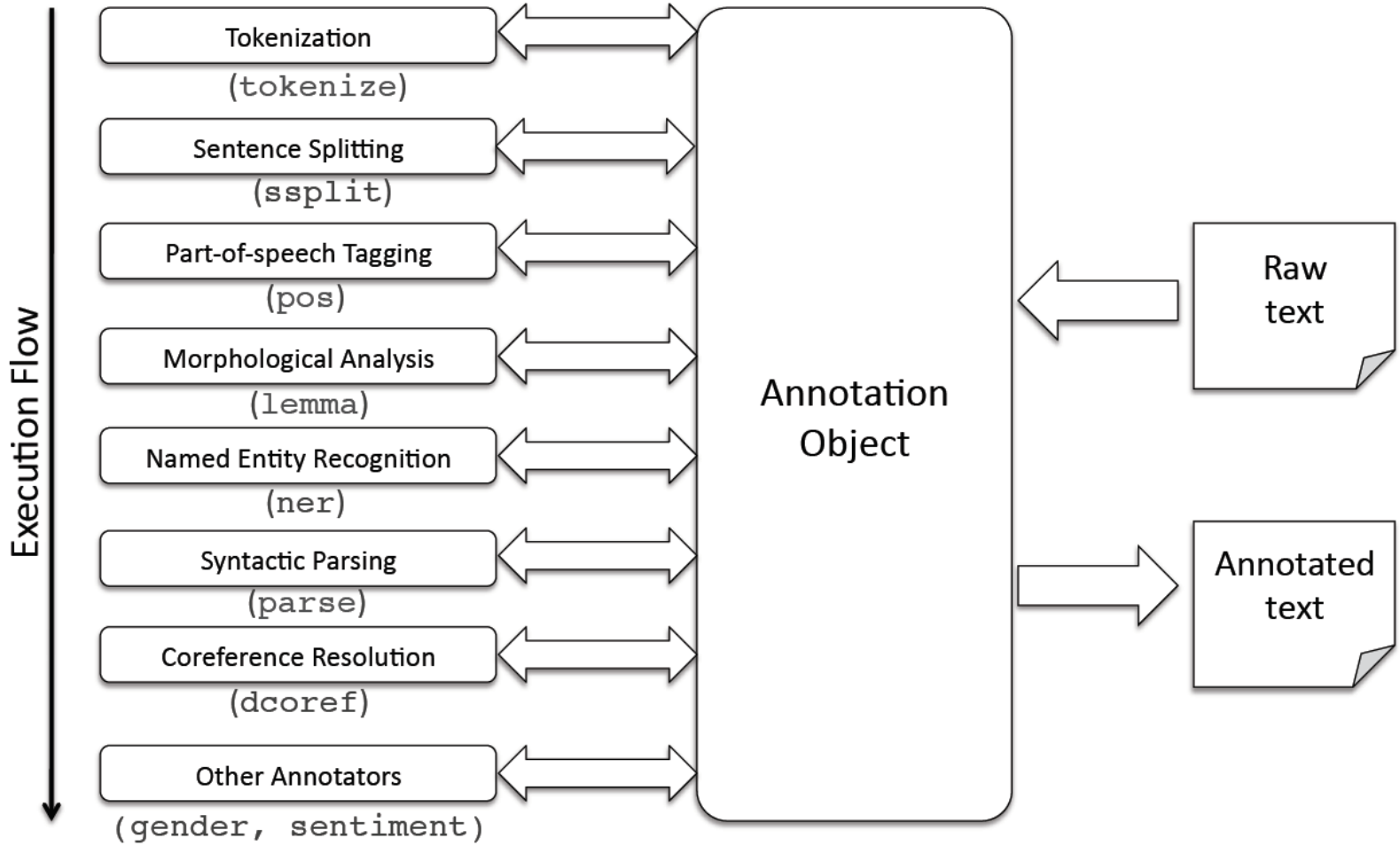
Abstract

We describe the design and use of the Stanford CoreNLP toolkit, an extensible pipeline that provides core natural lan-

[PDF] The **Stanford CoreNLP natural language processing toolkit**
CD Manning, M Surdeanu, J Bauer, JR Finkel... - Proceedings of 52nd ..., 2014 - aclweb.org
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 and also among commercial and government users of open source NLP technology. We suggest that this follows from a simple, approachable design, straightforward interfaces, the inclusion of robust and good quality analysis components, and not requiring use of a large amount of associated baggage.
Cited by 5942 Related articles All 25 versions



Stanford CoreNLP



The NLTK toolkit

NLTK 3.4.5 documentation

NEXT | MODULES | INDEX

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 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 ...

Cited by 3420 Related articles All 22 versions

Are these topics linked to the trending things?

ACL Anthology FAQ Corrections Submissions Search...

BERT Rediscovers the Classical NLP Pipeline

Ian Tenney, Dipanjan Das, Ellie Pavlick

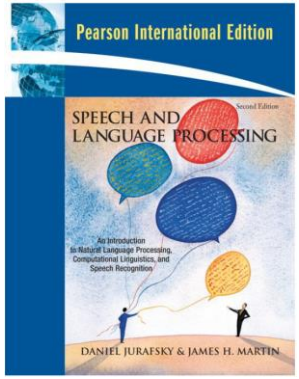
Abstract

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.

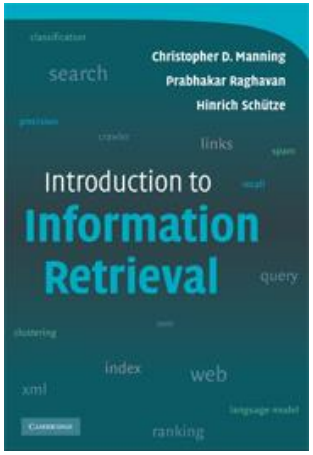
Anthology ID: P19-1452
Volume: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics
Month: July
Year: 2019
Address: Florence, Italy
Venue: ACL

Reference books

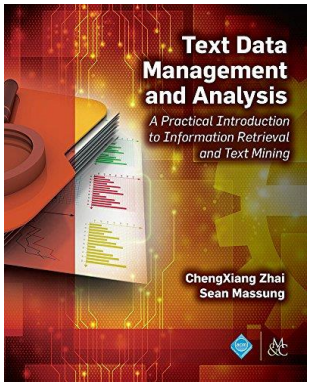
- Speech and Language Processing
  - Daniel Jurafsky and James H. Martin, 2nd edition, 2009
  - Draft of the 3rd edition: https://web.stanford.edu/~jurafsky/slp3/



- 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.



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.

- Microsoft Teams
  - AI6122 Team: **bi4k4ej**
  - Verification to be done at the end of add/drop period

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?

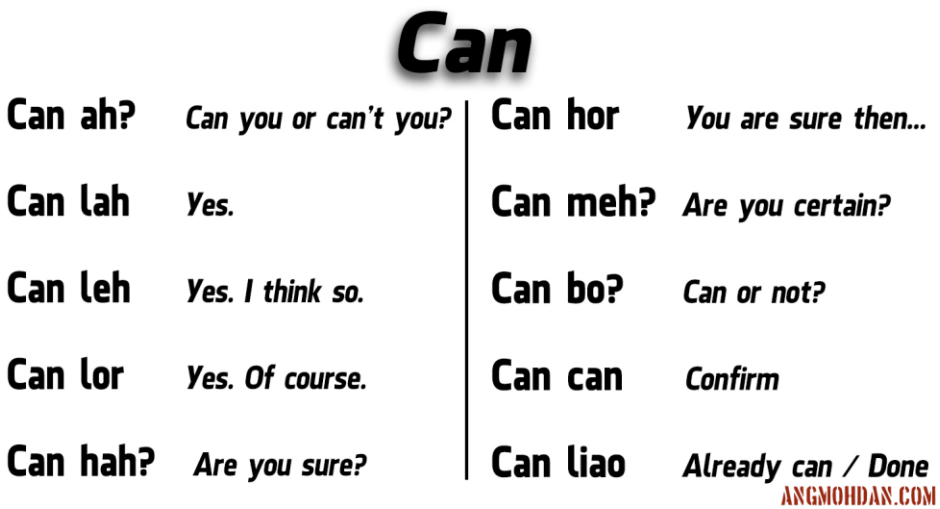
Problem, Dataset, Technique, Evaluation

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 text data in natural languages
  - You are expected to have knowledge of (basic) English grammar



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
  - Machine translation

- These days, the latter predominate
  - As NLP becomes increasingly practical, it is increasingly engineering oriented
  - It is hard to tell whether we are towards fully addressing the problem
    - https://thegradient.pub/machine-learning-wont-solve-the-natural-language-understanding-challenge/

Commercial world: Lots of exciting stuff going on

Google

Microsoft

IBM WATSON

THOMSON REUTERS

Chinese (Simplified)

知之为知之，不知为不知，是知也。 Edit

English

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

English

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

Chinese (Simplified)

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.

Open in Google Translate

"It is wise to hold what you know and admit what you don't know."

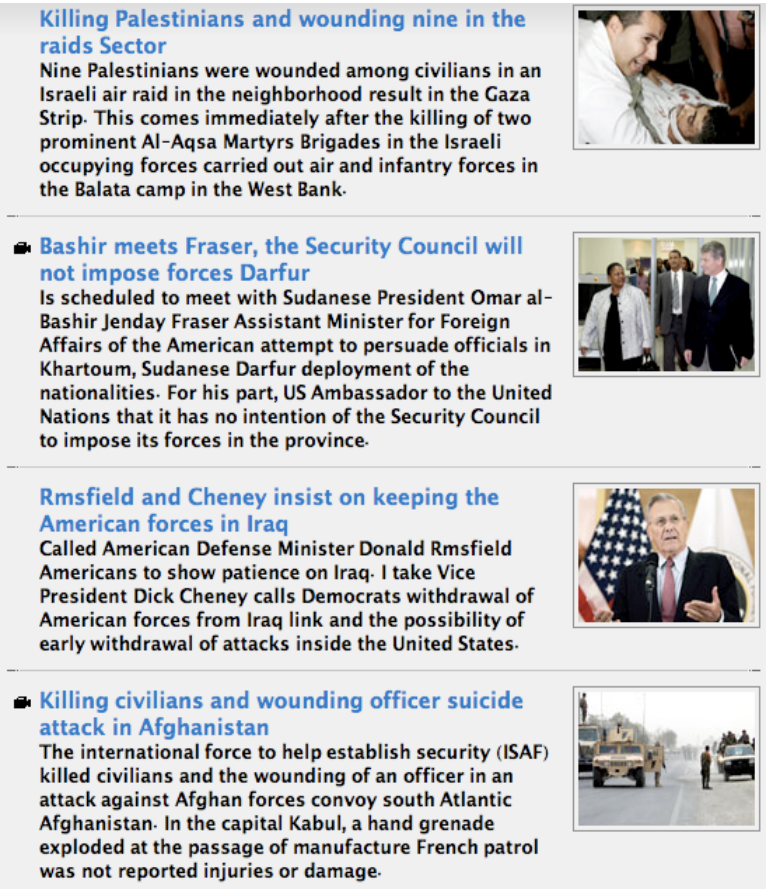
– Baidu Zhidao

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://explosion.ai/software#demos

Google Translate



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

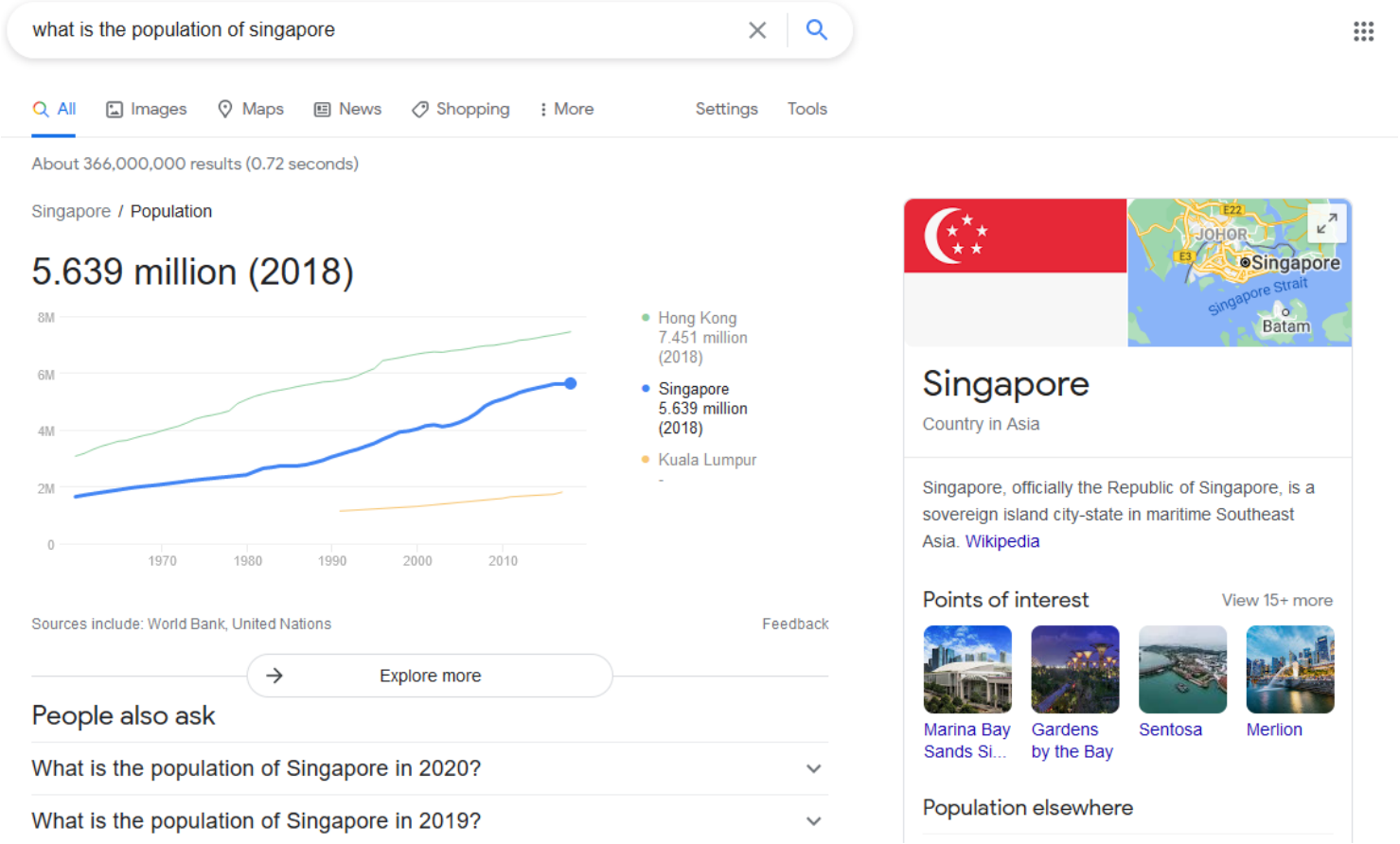
Called American Defense Minister Donald Rmsfield Americans to show patience on Iraq. I take Vice President Dick Cheney calls Democrats withdrawal of American forces from Iraq link and the possibility of early withdrawal of attacks inside the United States.

Killing civilians and wounding officer suicide attack in Afghanistan

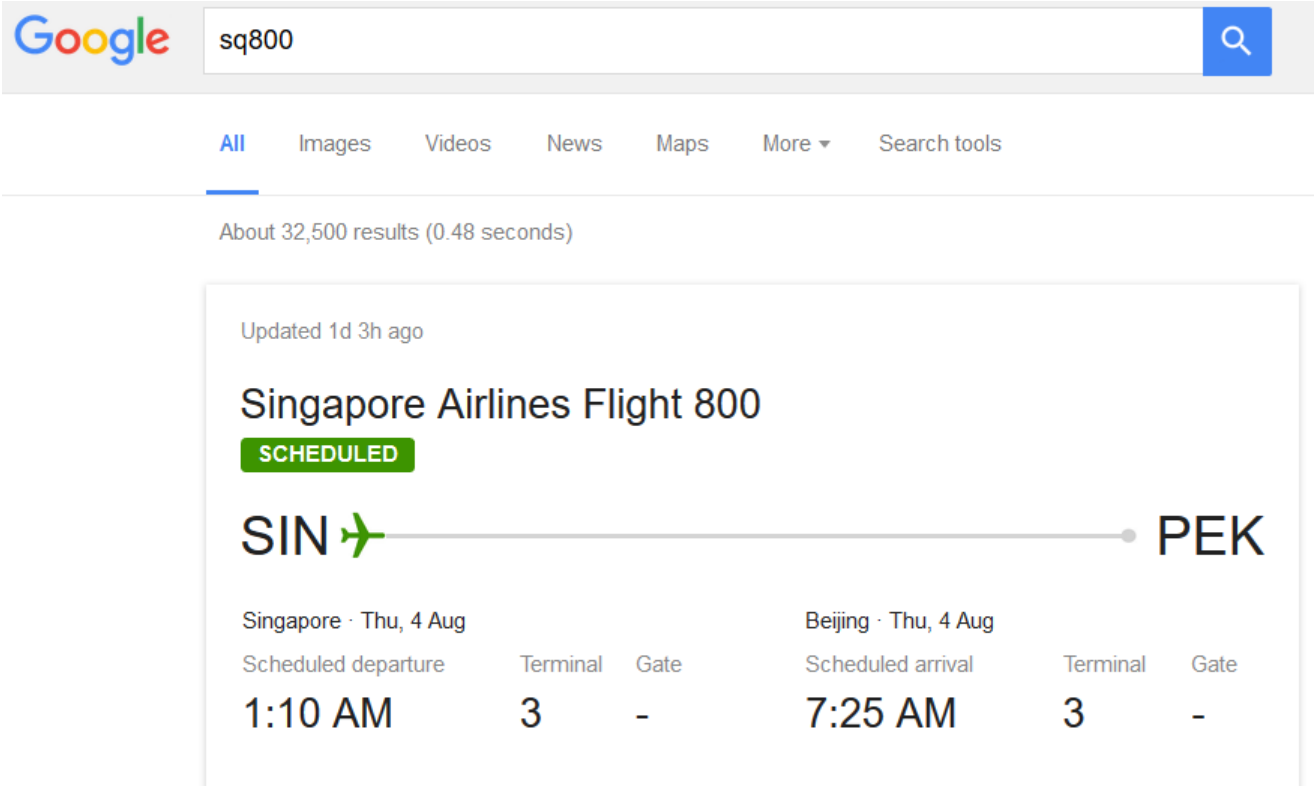
The international force to help establish security (ISAF) killed civilians and the wounding of an officer in an attack against Afghan forces convoy south Atlantic Afghanistan. In the capital Kabul, a hand grenade exploded at the passage of manufacture French patrol was not reported injuries or damage.

Web Q/A

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



Another example in Web search



Community based QnA

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: - ltp.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)

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?
  - 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**

<table>
<tr><th rowspan="2">Data</th><th rowspan="2">Finding Patterns</th><th colspan="2">Finding “Nuggets”</th></tr>
<tr><th>Novel</th><th>Non-Novel</th></tr>
<tr><td>Non-textual data</td><td>General Data Mining</td><td rowspan="2">Exploratory Data Analysis</td><td>Database Queries or other techniques</td></tr>
<tr><td>Textual data</td><td>Computational Linguistics</td><td>Information Retrieval</td></tr>
</table>

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

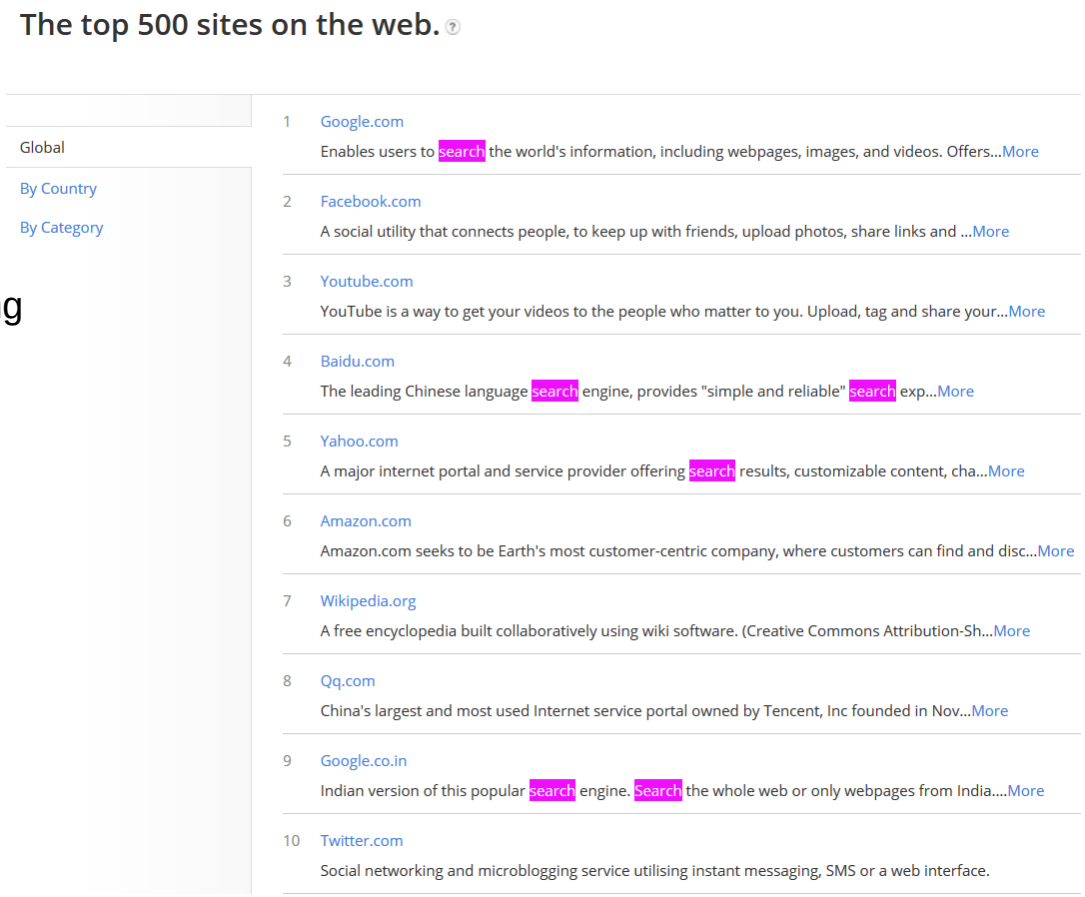
Metasearch engines

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



http://www.alexa.com/topsites

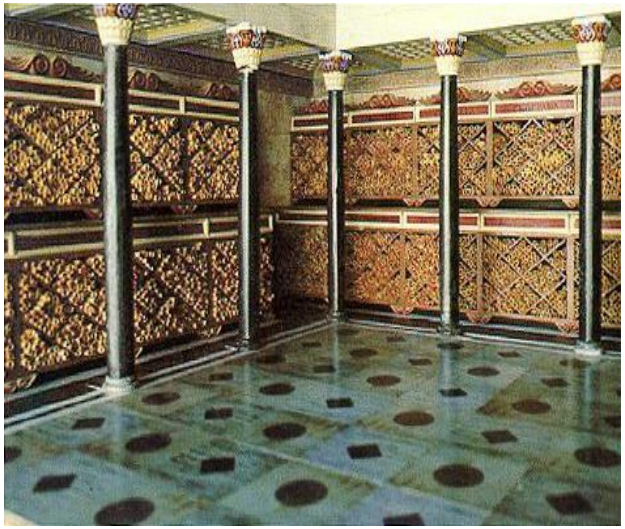
How to Retrieve Information?

- 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



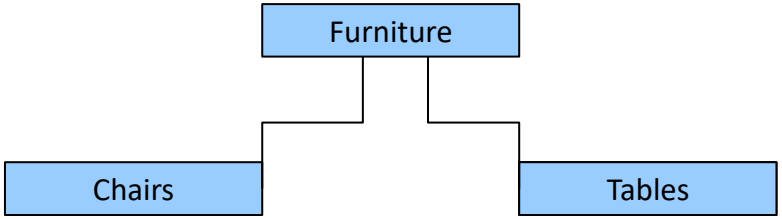
Classical Indexing

- **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.
- 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

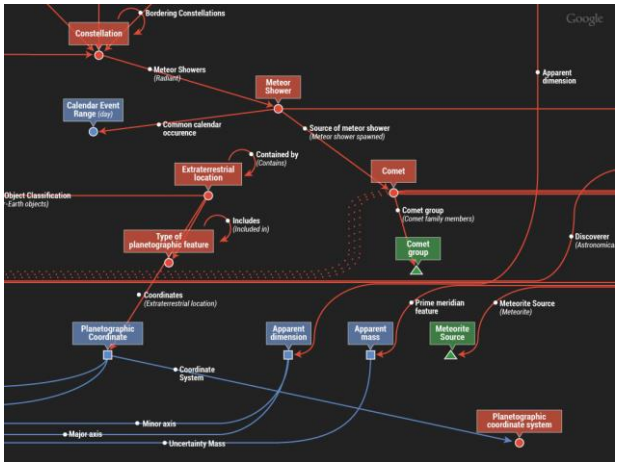
- Term **Specificity**
  - how broad/specific are the terms/keywords

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

Share Tweet 105

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