



NICF - TEXT ANALYTICS

MODULE 7: INFORMATION EXTRACTION

Leong Mun Kew
Institute of Systems Science
National University of Singapore
email: munkew@nus.edu.sg

© 2015 National University of Singapore. The contents contained in this document may not be reproduced in any form or by any means, without the written permission of ISS, NUS, other than for the purpose for which it has been supplied.



Outline for this Module

- **What is Information Extraction (IE)?**
- **How does an IE System work?**
 - Rule based methods for IE
 - Statistics based methods for IE
 - Coreference Resolution
- **Practical Information Extraction**
 - Example on how to extract concepts using a tool
 - Analysis of the tool and the results



FROM WORDS TO CONCEPTS



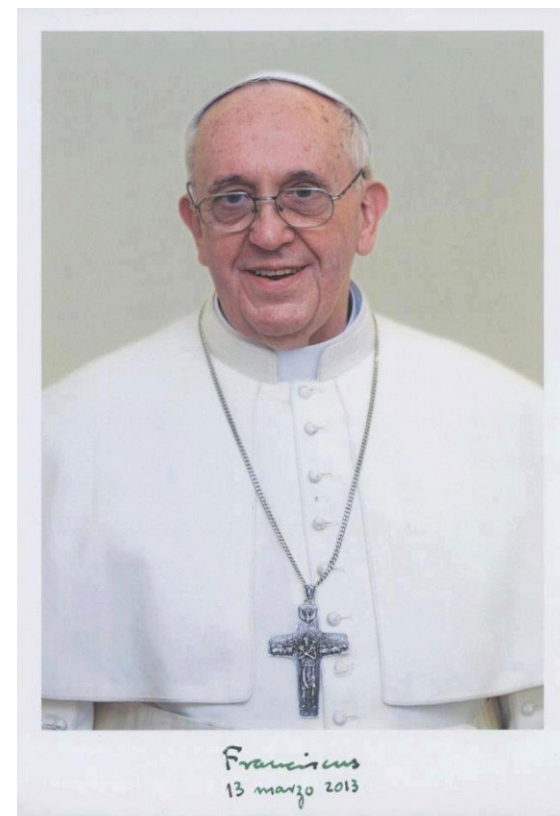
What's a word?

- A word is a bunch of characters separated by spaces or punctuation
 - “**The cat is sitting on the mat**” has 7 words
- How about “**The cat’s sitting on the mat**” – how many words?
- How about “**新加坡**” (Chinese) – how many words? Characters?
- How about “**memperkuasakan**” (Malay) – how many words?



What's a concept?

- **“Cardinal Jorge Mario Bergoglio of Buenos Aires was elected the 266th pope of the Catholic Church on 13 May 2013, taking the name Pope Francis”**
- **Cardinal: a concept = a prince of the Catholic Church**
- **Jorge Mario Bergoglio: a concept = Cardinal of Buenos Aires**
- **Buenos Aires: a concept = city in Argentina**
- **Pope: a concept = head of Catholic Church**
- **Catholic Church: a concept = religious organization**
- **13 May 2013: a concept = a date**
- **Pope Francis: a concept = current Pope**





Greedy heuristic 😊

Solve an optimization problem by finding locally optimal solutions.

How do you “automatically” create concepts

- Start with nouns or names or simple named entities (NEs)

- Identify NEs using dictionaries = lists = gazetteer = ...
- Eg, name lists: boy names, girl names, world names,...
- Eg, days, months, cities, countries, ... (can be multiwords)
- Proper nouns = words that start with Capitals (in the middle of sentence)

A gazetteer is a geographical dictionary or directory used in conjunction with a map or atlas.

- If nouns or names are contiguous, then aggregate them

- Jorge + Mario + Bergoglio = “Jorge Mario Bergoglio”
- Pope + Francis = “Pope Francis”
- Catholic + Church = “Catholic Church”

- Some lexical and syntactic markup can help

- Rule: if **X**_[title] + **Y**_[name] + “of.” + **Z**_[place] then **X Y of Z** is a concept
- Given:

Cardinal_[title]
Jorge Mario Bergoglio_[name]
Buenos Aires_[place]

- with the rule above: “Cardinal Jorge Mario Bergoglio of Buenos Aires” is a concept



Concept vs. Named Entity vs. Information

- **Name Entity = lowest level of recognition by an IE system**
 - Normally recognized by dictionaries or rules
- **Concept = rule or heuristic to create an abstraction**
 - Sometimes called a “natural class” = different people at different times and in different places would refer to the same referent with that concept
 - “president of the United States” vs. “president of the United Kingdom”
- **Information = words, named entities, concepts which fulfill a need**
 - So if you have a question, and a phrase answers that question, then that phrase is an example of information
 - Information is often regular, i.e., with a pattern
 - Eg, information about a person = name, age, sex, address, hp#, ...
 - Information about a company = name, address, stock symbol, Chairman, ...



EXERCISES

The National University of Singapore and The Hebrew University of Jerusalem (HUJ) are launching a Joint Doctor of Philosophy degree programme in biomedical science from August 2013. Professor Tan Eng Chye, NUS Deputy President and Provost, and Professor Menahem Ben-Sasson, President of HUJ signed the joint degree agreement at NUS, in the presence of Ambassador of Israel to Singapore Her Excellency Amira Arnon and about 30 invited guests.

Identify concepts from the text above, and categorize the concepts. Follow the examples below:

#	Concept	Category	#	Concept	Category
1	National University of Singapore	Place	7		
2	Tan Eng Chye	Name	8		
3			9		
4			10		
5			11		
6			12		

- Professor Tan Eng Chye, NUS Deputy President and Provost, and Professor Menahem Ben-Sasson, President of HUU signed the joint degree agreement at NUS, in the presence of Ambassador of Israel to Singapore Her Excellency Amira Arnon and about 30 invited guests.
- *The following is a simple lexical rule to recognize names of universities in the paragraph above:*

If “The” + $X_{[capitalized]}$ + “University” + “of” + $Y_{[place]}$

then “The X University of Y” is the name of a university

- *Write a rule similar in structure to recognize the names of persons mentioned in the paragraph above. You may assume reasonable dictionaries (lists) exist and that suitable parts of speech and meaningful markup has been done:*

If "Professor" + X + Y + Z Or "Excellency" + X + Y + Z
 Then X Y Z is names of persons



WHAT IS INFORMATION (CONCEPT) EXTRACTION



Contrast with a search engine

How tall is the Eiffel Tower?

- A search engine returns a list of documents which (hopefully) includes the answer
- The user needs to open/read the document to get the answer

bing Beta

WEB IMAGES NEWS MORE

how tall is the eiffel tower

4,850,000 RESULTS Narrow by language ▼ Narrow by region ▼

[How Tall is the Eiffel Tower? - Buzzle](#)
www.buzzle.com/articles/how-tall-is-the-eiffel-tower.html ▼
How Tall is the Eiffel Tower? Eiffel Tower is the tallest man-made marvel in France. This massive latticework structure is owned by the city of Paris and is a ...

[Eiffel Tower - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/Eiffel_Tower ▼
History · Design of the tower · Tourism · Attempted relocation · Economics
The tower stands 320 metres (1,050 ft) tall, about the same height as an 81-story building. During its construction, the Eiffel Tower surpassed the Washington Monument to ...

[How tall is The Eiffel Tower? | Height of the Celebrities](#)
howtallis.info/eiffel-tower.html ▼
Want to know How tall is The Eiffel Tower? Ok, here we are. Find the answer of this question. Eiffel Tower height is 324 metres (1,063 ft)

[How tall is the Eiffel tower? - Yahoo! Answers India](#)
in.answers.yahoo.com/question/index?qid=20071125233320AAnuotX ▼
Resolved · 7 total answers
26/11/2007 · Best Answer: According to The Oxford Dictionary of Phrase and Fable: "Eiffel Tower a wrought-iron structure erected in Paris for the World Exhibition of ...

[Eiffel Tower Height, How Tall Is The Eiffel Tower?](#)
myeiffeltower.com/eiffel-tower-height ▼
Find out the exact Eiffel Tower height here! Discover how tall is the Eiffel Tower - the most visited monument in the world!

[How tall is the Eiffel Tower? - Yahoo! Answers](#)
answers.yahoo.com/question/index?qid=20070216082426AA7wNpQ ▼
Resolved · 3 total answers
16/2/2007 · Best Answer: Eiffel Tower (French: La Tour Eiffel, /tur ɛfɛl/) is an iron tower built on the Champ de Mars beside the River Seine in Paris, France. It ...



Information extraction version



how tall is the eiffel tower



Mun Kew Leong

3

+ Share

Search

About 2,420,000 results (0.26 seconds)

Web

Images

Maps

Videos

News

Shopping

More

Show search tools

324 m

Eiffel Tower, Height

[Hide details](#)

[How tall is the Eiffel Tower - answers.com](#)

The Eiffel tower is located in Paris, France. It is **324 m** ...

[Eiffel Tower Height, How ... - myeiffeltower.com](#)

Eiffel Tower height - **324 m**. of iron parts! In 1889 ...

[How tall is the eifel tower? - Yahoo! Answers - yahoo.com](#)

... the structure is **324 m** (1058 ft) ...

Is this accurate? [Yes](#) - [No](#)

[Eiffel Tower - Wikipedia, the free encyclopedia](#)

[en.wikipedia.org/wiki/Eiffel_Tower](#)

Jump to [Lattice towers taller than the Eiffel Tower](#): The **Eiffel Tower** is a puddled iron lattice tower located on the Champ de Mars in Paris, named ...

[History](#) - [Design of the tower](#) - [Tourism](#) - [Attempted relocation](#)

[How tall is the Eiffel Tower](#)

[wiki.answers.com](#) > ... > [France](#) > [Paris](#) > [Eiffel Tower](#)

The **Eiffel Tower** is 100 meters **tall**. **How tall is the Eiffel tower** and which city is it in? The **Eiffel tower** is located in Paris, France. It is **324 m** (1,063 ft) **tall**.

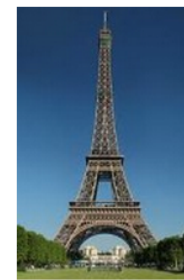
[Would Shard critics once have slammed the Eiffel Tower? | Daniel ...](#)

[www.guardian.co.uk/commentisfree/.../shard-critics-eiffel-tow...](#) [Share](#)

5 Jul 2012 – But while the main body of the **Eiffel Tower** is only 300 metres **tall**, it stands at 324 metres once you add the base and television antenna. >>

Eiffel Tower

[Directions](#)



[en.wikipedia.org](#)

The Eiffel Tower is a puddled iron lattice tower located on the Champ de Mars in Paris, named after the engineer Gustave Eiffel, whose company designed and built the tower. [Wikipedia](#)

Height: 324 m

Construction started: January 28, 1887

Hours: Mon-Sun 9:30am–11:45pm

Address: Avenue Gustave Eiffel, 75007 Paris, France

Phone: 0892 70 12 39

Architect: [Stephen Sauvestre](#)

People also search for



How tall is the eiffel tower?

324 m

Eiffel Tower, Height
[Hide details](#)

Answer

[How tall is the Eiffel Tower](#) - [answers.com](#)

The Eiffel tower is located in Paris, France. It is **324 m** ...

[Eiffel Tower Height, How ...](#) - [myeiffeltower.com](#)

Eiffel Tower height - **324 m**. of iron parts! In 1889 ...

[How tall is the eifel tower?](#) - [Yahoo! Answers](#) - [yahoo.com](#)

... the structure is **324 m** (1058 ft) ...

Is this accurate? [Yes](#) - [No](#)

Sources
(evidence)

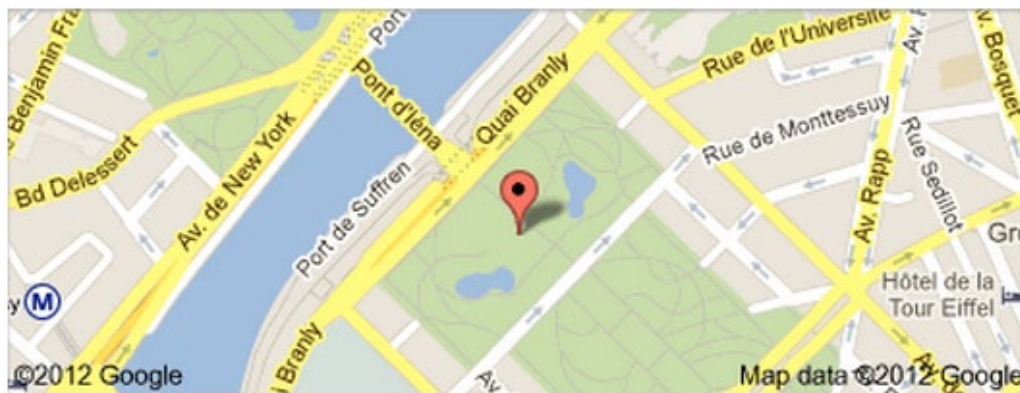


Template version

- **Automatic template filling**
 - Different fields for different places
 - Depends on sources; not fixed template
- **Information Extraction**
- **High degree of accuracy**
- **Human checking?**

Eiffel Tower

[Directions](#)



The Eiffel Tower is a puddled iron lattice tower located on the Champ de Mars in Paris, named after the engineer Gustave Eiffel, whose company designed and built the tower. [Wikipedia](#)

Height: 324 m

Construction started: January 28, 1887

Hours: Mon-Sun 9:30am–11:45pm

Address: Avenue Gustave Eiffel, 75007 Paris, France

Phone: 0892 70 12 39

Architect: [Stephen Sauvestre](#)

Lincoln Memorial

[Directions](#)



en.wikipedia.org

The Lincoln Memorial is an American national monument built to honor the 16th President of the United States, Abraham Lincoln. It is located on the National Mall in Washington, D.C. across from the Washington Monument. [Wikipedia](#)

Height: 30 m

Area: 66 ha

Address: 2 Lincoln Memorial Cir NW, Washington, DC 20037

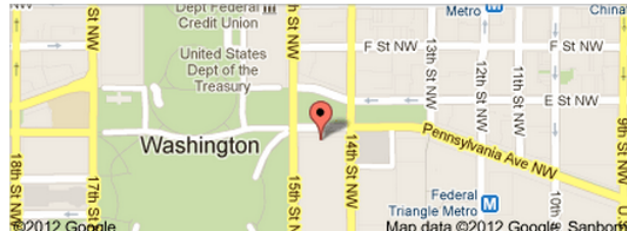
Hours: Mon-Sun Open 24 hours

Architectural styles: Beaux-Arts architecture, Doric order

Architect: [Henry Bacon](#)

White House

[Directions](#)



plus.google.com

The White House is the official residence and principal workplace of the President of the United States. [Wikipedia](#)

Construction started: 1792

Address: Alexander Hamilton Place Northwest, Washington, DC 20005

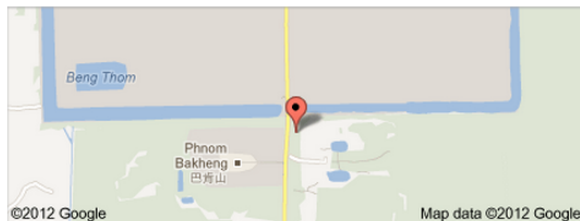
Phone: (202) 208-1631

Hours: Mon-Sun 7:30am–4pm

Architectural styles: Neoclassicism, Palladian architecture, Neoclassical architecture

Architect: [James Hoban](#)

Angkor Wat



en.wikipedia.org

Angkor Wat is the largest Hindu temple complex in the world. The temple was built by King Suryavarman II in the early 12th century in Yasodharapura, the capital of the Khmer Empire, as his state temple and eventual mausoleum. [Wikipedia](#)

Construction started: 1125

Address: Angkor Wat, Siem Reap 82720, Cambodia

Phone: (0)63 764444

Function: Place of worship

Architectural styles: Dravidian architecture, Khmer architectural style

Machu Picchu



en.wikipedia.org

Machu Picchu is a pre-Columbian 15th-century Inca site located 2,430 metres above sea level. Machu Picchu is located in the Cusco Region of Peru, South America. [Wikipedia](#)

Address: Muro de la Ciudad, Peru

Management: Government of Peru

Phone: (0)1 610 0400



What is information extraction?

- **The automatic extraction of (possibly pre-specified) information from natural language documents**
 - Facts about types of entities, events, relationships
- **The automatic population of a structured information source (template) from natural language documents (i.e., create a table!)**
 - Documents may be semi-structured (eg., patents), unstructured (e.g., websites) or free text (e.g., documents)



Things to note...

- **IE does not specify the documents to be read; rather it extracts the salient information from the documents and presents just those information to the user**
- **Normally maintains links from facts to source documents**
 - Allows evidence trail, context, further discovery
- **Accuracy is variable – both in completeness and reliability**
- **Certain information (e.g., person names; organizations) have high reliability. Other information (e.g., sentiment) have very poor accuracy.**



Drilling down on information...

- **Entity/Concept:**
 - an object of interest (date, person, building,...)
- **Attribute:**
 - property of an entity (name, height, type)
- **Fact:**
 - a predicate about an entity, e.g., tall(X)
 - a relationship between two or more entities, e.g., married(X,Y)
- **Event**
 - an abstraction over a possibly sequential relationship of several entities e.g., terrorist attack, birthday party, etc.

Example (Entities and Facts)

- Fletcher Maddox, former Dean of the UCSD Business School, announced the formation of La Jolla Genomatics together with his two sons. La Jolla Genomatics will release its product Geninfo in June 1999. Geninfo is a turnkey system to assist biotechnology researchers in keeping up with the voluminous literature in all aspects of their field.
- Dr. Maddox will be the firm's CEO. His son, Oliver, is the Chief Scientist and holds patents on many of the algorithms used in Geninfo. Oliver's brother, Ambrose, follows more in his father's footsteps and will be the CFO of L.J.G. headquartered in the Maddox family's hometown of La Jolla, CA.

Persons:	Organizations:	Locations:	Artifacts:	Dates:
Fletcher Maddox	UCSD Business School	La Jolla	Geninfo	June 1999
Dr. Maddox	La Jolla Genomatics	CA	Geninfo	
Oliver	La Jolla Genomatics			
Oliver	L.J.G.			
Ambrose				
Maddox				

PERSON	Employee_of	ORGANIZATION
Fletcher Maddox	Employee_of	UCSD Business School
Fletcher Maddox	Employee_of	La Jolla Genomatics
Oliver	Employee_of	La Jolla Genomatics
Ambrose		La Jolla Genomatics
ARTIFACT	Product_of	ORGANIZATION
Geninfo	Product_of	La Jolla Genomatics
LOCATION	Location_of	ORGANIZATION
La Jolla	Location_of	La Jolla Genomatics
CA	Location_of	La Jolla Genomatics

Events and Attributes

COMPANY-FORMATION_EVENT:

COMPANY:	La Jolla Genomatics
PRINCIPALS:	Fletcher Maddox Oliver Ambrose
DATE:	
CAPITAL:	

RELEASE-EVENT:

COMPANY:	La Jolla Genomatics
PRODUCT:	Geninfo
DATE:	June 1999
COST:	

NAME:	Fletcher Maddox Maddox
DESCRIPTOR:	former Dean of the UCSD Business School his father the firm's CEO
CATEGORY:	PERSON
NAME:	Oliver
DESCRIPTOR:	His son Chief Scientist
CATEGORY:	PERSON
NAME:	Ambrose
DESCRIPTOR:	Oliver's brother the CFO of L.J.G.
CATEGORY:	PERSON
NAME:	UCSD Business School
DESCRIPTOR:	
CATEGORY:	ORGANIZATION
NAME:	La Jolla Genomatics L.J.G.
DESCRIPTOR:	
CATEGORY:	ORGANIZATION
NAME:	Geninfo
DESCRIPTOR:	its product
CATEGORY:	ARTIFACT
NAME:	La Jolla
DESCRIPTOR:	the Maddox family's hometown
CATEGORY:	LOCATION
NAME:	CA
DESCRIPTOR:	
CATEGORY:	LOCATION



Unstructured Text

POLICE ARE INVESTIGATING A ROBBERY THAT OCCURRED AT A 7-ELEVEN STORE LOCATED AT 2545 LITTLE RIVER TURNPIKE, LINCOLNIA AREA ABOUT 12:30 AM FRIDAY. A 21-YEAR-OLD ALEXANDRIA AREA EMPLOYEE WAS APPROACHED BY TWO MEN WHO DEMANDED MONEY. SHE RELINQUISHED AN UNLabeled BAG OF CASH AND THE MEN LEFT. NO ONE WAS INJURED. THE SUSPECTS DESCRIBED AS BLACK, IN THEIR MID TWENTIES, 5 FEET NINE INCHES TALL, WITH MEDIUM BUILD, DARK HAIR, CLEAN SHAVEN. THEY WERE BOTH WEARING BLACK COATS. ANYONE WITH INFORMATION ABOUT THE ROBBERY OR THE SUSPECTS INVOLVED IS ASKED TO CALL 5555.

Structured (Desired) Information

Crime	Address	Town	Time	Day
ABDUCTION	8700 BLOCK OF LITTLE RIVER TURNPIKE,	ANNANDALE	11:30 PM	SUNDAY
...
ROBBERY	7-ELEVEN STORE LOCATED AT 2545 LITTLE RIVER TURNPIKE,	LINCOLNIA	12:45 AM	FRIDAY
ROBBERY	7-ELEVEN STORE LOCATED AT 5624 OX ROAD,	FAIRFAX	3:00 AM	FRIDAY

From: Ron Feldman, *Information Extraction: Theory and Practice*, <http://cs.fit.edu/~pkc/icdm03/printing/tutorials/extraction/extraction.tutorial.pdf>



QUICK EXERCISE

Read this text:

- During the Boston Marathon on April 15, 2013, two pressure cooker bombs exploded at 2:49 pm EDT killing 3 people and injuring 264 others. The bombs exploded about 13 seconds and 210 yards (190 m) apart, near the finish line on Boylston Street. The Federal Bureau of Investigation (FBI) took over the investigation, and on April 18, released photographs and surveillance video of two suspects. The suspects were identified later that day as the Chechen brothers Dzhokhar and Tamerlan Tsarnaev.

(from:
https://en.wikipedia.org/wiki/Boston_Marathon_bombings)

Fill up this template:

Attribute	Value
Occurrence	
Event	
Location	
Start Date	
Start Time	
End Date	
End Time	
Method	
Incident_1	
Incident_2	
Incident_3	
Perpetrators	
Caught (Y/N)	

Switzerland, Sweden and Singapore are the three most innovative countries in the world, according to the Global Innovation Index (GII) 2012, a ranking of 141 countries co-produced by INSEAD and the World Intellectual Property Organization (WIPO, a specialised agency of the United Nations) published in Geneva on July 3, 2012. Currently in its fifth year, the GI measures the degree to which countries and economies integrate innovation into their political business and social spheres. Knowledge partners for the GI 2012 are Alcatel-Lucent, Booz & Company and the Confederation of Indian Industry (CII).

- This is the second year running that Switzerland, Sweden and Singapore have been in the top three positions. The rest of the top ten this year are: Finland, the United Kingdom, the Netherlands, Denmark, Hong Kong (China), Ireland and the United States. Canada dropped out of the top ten this year, while the U.S. fell to tenth position from number seven last year, changes which the report attributes to cutbacks in spending on, and support of, education and research and development.

(from: <http://knowledge.insead.edu/innovation/global-innovation-index-2012-481>)

- Read the text above and fill in the table below:**

Country	2012 Rank	2011 Rank	Country	2012 Rank	2011 Rank
	1			6	
	2			7	
	3			8	
	4			9	
	5			10	



HOW DOES AN IE SYSTEM WORK?



Types of IE systems

- **Rule-based Systems**
 - Hand-coded rules
 - Coded by linguists, with domain input
 - Iterative method based on document inspection
 - Slow but very good results
 - Induced (machine learning) rules
 - Fully machine learning
 - Given an annotated corpus, derive a basis set of rules that cover a pre-determined % of the annotated examples (and only the annotated examples)
 - Heuristic approach: one rule at a time!
 - Hybrid systems – machine learning to fine-tune the rules

A heuristic technique, often called simply a heuristic, is any approach to problem solving, learning, or discovery that employs a practical method not guaranteed to be optimal or perfect, but sufficient for the immediate goals.

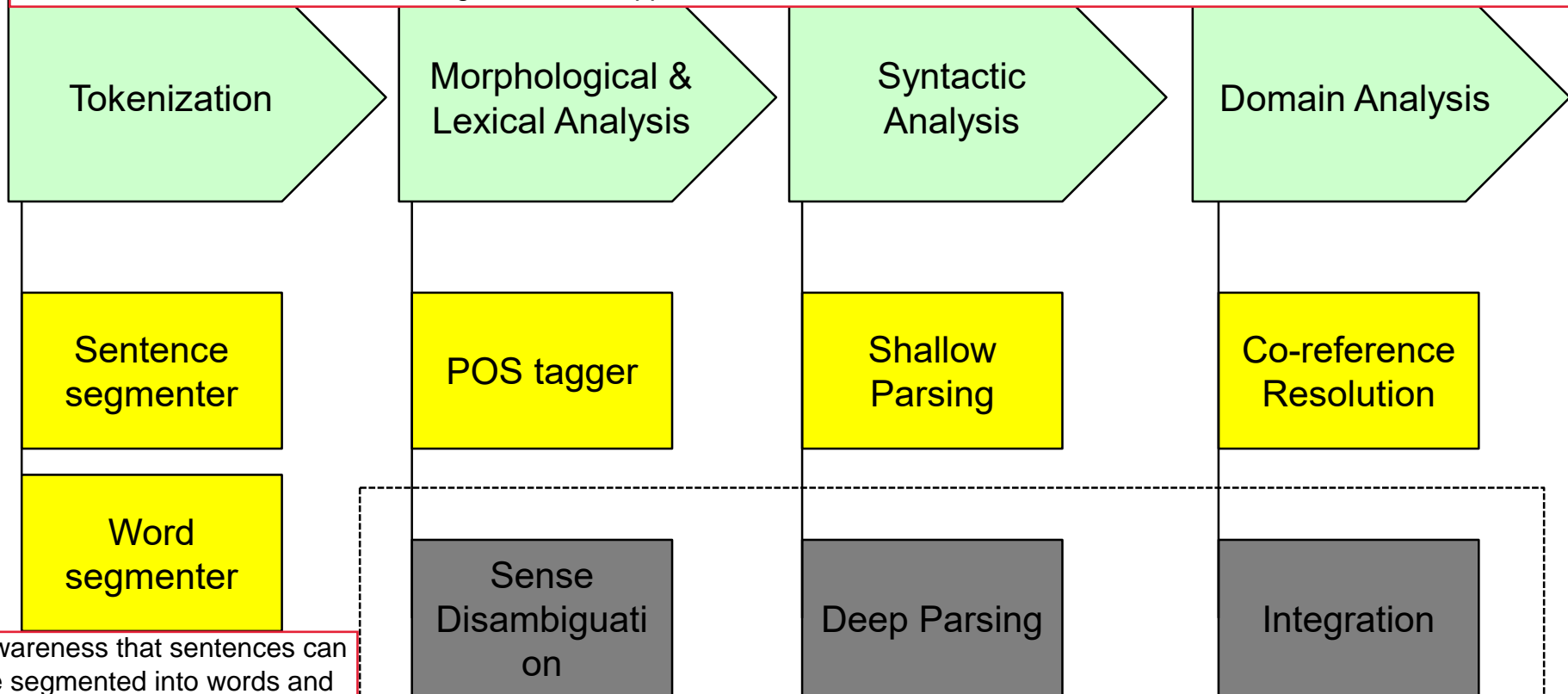
- **Statistics-based Systems**

- Start with a well-annotated corpus
- Depending on the method (e.g., Hidden markov models), derive statistical rules to create a model that generates the examples
- Advantages compared to Rule based systems
 - Language independent (within representational limits)
 - No linguistic or domain knowledge needed in the team
 - Relatively small effort in creating the models
- Issues
 - The complexity moves to the corpus – must be well annotated and must cover the full space of possibilities
 - Requires very large number of training examples to get good results

Hidden Markov Models (HMMs) are a class of probabilistic graphical model that allow us to predict a sequence of unknown (hidden) variables from a set of observed variables. A simple example of an HMM is predicting the weather (hidden variable) based on the type of clothes that someone wears (observed).

Main components of an IE system

Shallow parsing is the task of parsing a limited part of the syntactic information from the given text. While deep parsing is required for more complex NLP applications, such as dialogue systems and summarization, shallow parsing is more suited for information extraction and text mining varieties of applications.



Nice to have but many good systems work without them

WSD is basically solution to the ambiguity which arises due to different meaning of words in different context. For example, consider the two sentences. "The bank will not be accepting cash on Saturdays." "The river overflowed the bank."

Awareness that sentences can be segmented into words and that words can be segmented into syllables. Orally segmenting sentences into words helps children understand that sentences are composed of separate words

Have you ever watched a
semantic extraction engine at work?

0:02 / 2:20

Semantic Extraction in Slow Motion



joinvision · 2 videos



Subscribe

4

3,083

4 1

Like

About

Share

Add to



Uploaded on 28 Jul 2011

Have you ever seen a semantic extraction engine in action? This video demonstrates how JoinVision's CVlizer application tags and extracts all relevant information from a common CV step by step in slow motion. The process of information extraction, usually lasting for less than a second, is broken down to its elemental operations,

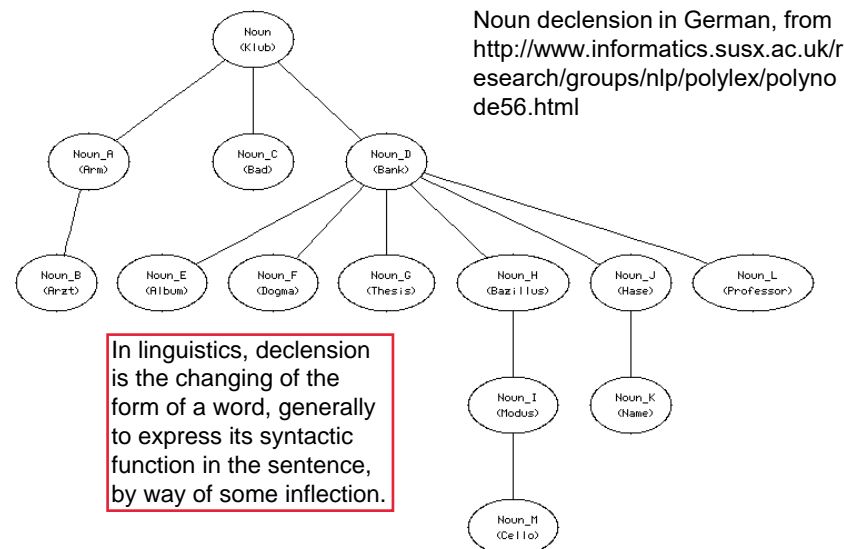
Buy "Peer Gynt Suite No 1 Op 46: In the Hall of the Mountain King" on

From: <http://www.youtube.com/watch?v=gsAATmT0qEc>

Issues in IE - Tokenization

Language Considerations

- Morphology is easy in English, much harder in German, Hebrew
- Word boundary easy in roman script, ambiguous in character scripts
- English is one-dimensional, other languages have more than one



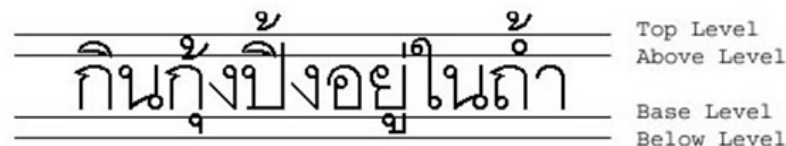
original, un-segmented text

再往远些看，随着汉字识别和语音识别技术的发展，中文计算机用户将跨越语言差异的鸿沟，在录入上走向中西文求同的道路。

separated word entities after segmentation

再往远些看，随着汉字识别和语音识别技术的发展，中文计算机用户将跨越语言差异的鸿沟，在录入上走向中西文求同的道路。

Word segmentation in Chinese, from <http://blog.faroo.com/2009/06/09/lightweight-chinese-word-segmentation/>



Word composition in Thai, from <http://www-01.ibm.com/software/globalization/topics/thai/character.html>

Issues in IE – Language in use (pragmatics)

Pragmatics refers to how words are used in a practical sense.
For example, words that attempt to explain abstract concepts-freedom, beauty-have no meaning in and of themselves. Instead, someone who looks at pragmatics would attempt to understand how they are being used in a given, concrete, practical situation.

Examples of difficulties:

- **Sarcasm**

- John is really smart...like a coconut!

- **Negation and ambiguity**

- “*All that glitters is not gold*”
 - *For all (x), if x glitters, then x is (not gold)
 - It is not the case that (for all (x), if x glitters, then x is gold)

- **Prosody**

- John and Anne are *married*? I thought the wedding was next year!
- John and Anne *are* married? I thought they were just going out!
- John and *Anne* are married? I thought John was interested in Mary!
- *John* and Anne are married? I thought Anne was going out with Tom!
- John *and* Anne are married? I thought they were both single!

- **Colloquialisms**

- “Referee *kayu*! Blind like donkey, is it?! So obvious ball out one leh!”

a word or phrase that is not formal or literary and is used in ordinary or familiar conversation.



Issues in IE – Structure

- **Genres (“styles” of documents)**

- Newspapers, journal articles, patents
 - Semi-structured – there is a pattern to how artifacts are created
- Emails, tweets, SMSs
 - Semi-structured, well defined fields
 - Short or very short “sentences”
- Speech transcripts (e.g., court documents, Hansard, etc.)
 - Natural speech is rarely fluent
 - Interjections, repetitions, hesitations, malapropisms, indexicals, etc.

Hansard is the traditional name of the transcripts of Parliamentary debates in Britain and many Commonwealth countries.

- **Structure**

- Information in tables, in captions
- Precedence (implied order)

Malapropism is the act of using an incorrect word in place of one that is similar in pronunciation.



RULE-BASED IE METHODS



When to use rule-based systems?

- **Rule-based systems can and do work well**
 - Corpus is relatively static (in terms of vocabulary, language structure, etc.)
 - Can be fast especially in well-defined limited domains
(compared to annotating training examples)
- **A typical rule-based system comprises**
 - Set of rules
 - Policies to control when and how (multiple) rules are applied, e.g., order, looping.

What does a rule look like?

- **Form:**
 - Match(pattern) then Do(action)

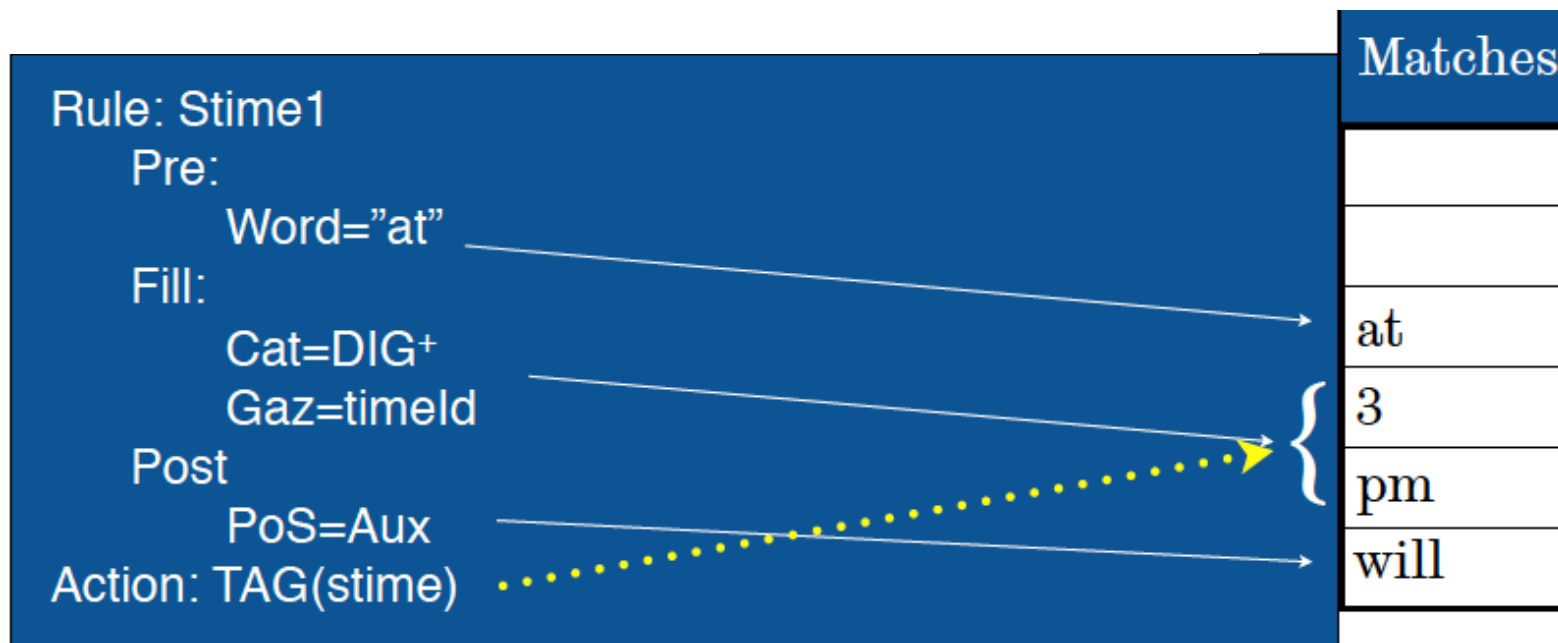
Rule: Company1 An orthography is a set of conventions for writing a language. from gate.ac.uk

```
( ( {Token.orthography == upperInitial} )+
  {Lookup.kind == companyDesignator}
):match
-->
:match.NamedEntity = { kind=company, rule="Company1" }
```

Types of IE recognition rules

- **Lexical pattern matching – whole entity patterns**
 - Classic rule: (left context + filler + right context) patterns
 - Simplistic: entities are considered independent
 - More natural for hand-coded rules

In lexicography, a lexical item is a single word, a part of a word, or a chain of words that forms the basic elements of a language's lexicon. Examples are cat, traffic light, take care of, by the way, and it's raining cats and dogs



Types of IE recognition rules

- **Lexical pattern matching – boundary patterns**
 - Start boundary is distinct from end boundary
 - Different rules are required to recognize the entity
 - Developed primarily for machine learning of rules

Rule: Stime1

Pre:

Word="seminar":

Word="at":

Post:

Cat=DIG+

Gaz=timeId

Action: TAG(<stime>)

Matches

The

seminar

at

3

pm

Types of IE recognition rules

- **Structural pattern recognition**
 - Typically used only for highly structured text
 - Can recognize more than one entity at a time

Example:

<p> Capitol Hill- **1** br twnhme. D/W W/D. Pkg incl
\$**675**. **3** BR upper flr no gar. \$**995**. (206)999-9999

3 Bedroom upper
floor no garden ?

Rule:

ID:7

Pattern: * ('Capitol Hill') * (*Digit*) * '\$' (*Number*)

Output: Rental {Neighborhood \$1} {Bedrooms \$2} {Price \$3}

Rule from: STEPHEN SODERLAND:

Learning Information Extraction Rules for Semi-structured and Free Text,
Machine Learning 1, 440



STATISTICS BASED IE METHODS



When to use statistics based systems?

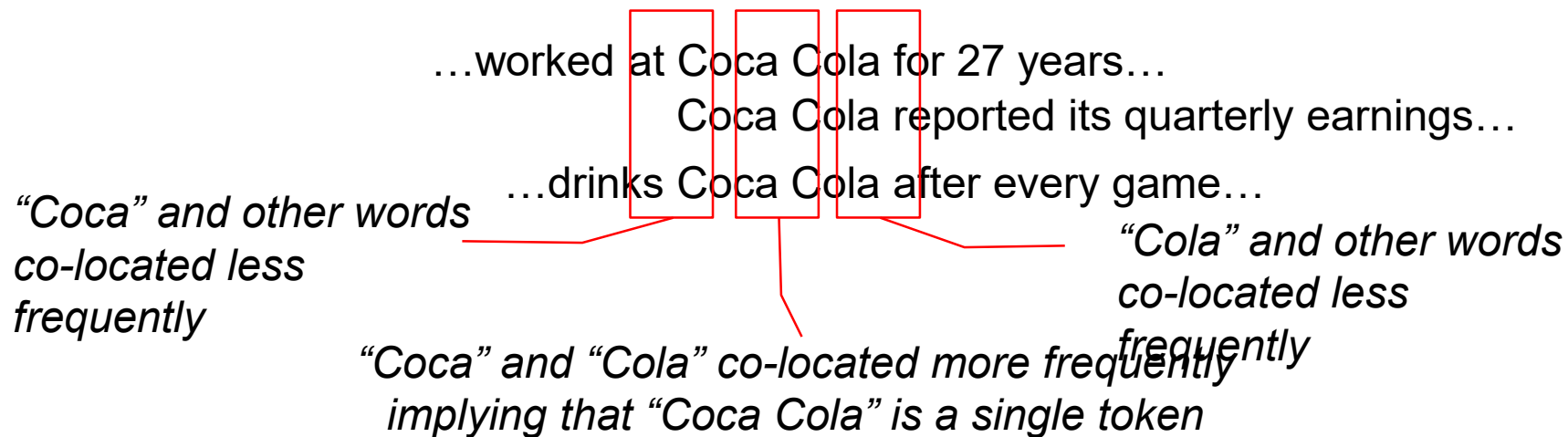
- **Current top performing systems are statistics based**
 - Machine learning (ML) on very large corpora is state-of-the-art
 - Beyond the scope of this course
- **Annotation based corpora for training**
 - You have a well annotated corpora with many features
 - Various ML techniques from simple to sophisticated
 - Relatively homogeneous real data (not training data) in any given domain. Note that models don't transfer well across domains
 - You don't have domain or language resources in that area



Simple model is at token level

- Text is a linear sequence of tokens (such as words)
- Token boundaries can be fairly easily derived in some languages, e.g., space & punctuation for English, but much harder for others, e.g., Chinese
- Simple tokenization
 - Dictionary based
 - Colocation frequencies (see next page)
- Alternatives
 - Ignore multi-unit tokens
 - Bi-grams, tri-grams, multi-grams

Colocation frequencies identify boundaries



- Colocation frequencies (given sufficient examples) can very accurately identify token boundaries.
- Greedy algorithm – the longer the token, the better
- Works on non-English languages, e.g., Chinese



Token model IE

Apple Inc	(AAPL)	reported	its	earnings	for	Q3	2010
Organization	Stock symbol	Null	Null	Null	Null	Quarter	Year

Neil	Armstrong	died	August	25	,	2012	in	Cincinnati
Null	Null	Null	dateStart	dateContinue	DateContinue	DateEnd	Null	Null

- The IE task is to assign a pre-defined entity label to each token in the stream
- Two examples above – different methods
- Treated as a generative model
- “Null” is the output when the model fails



Popular models

- **Hidden Markov Models (HMM)**
 - Simple, joint probability
- **Conditional Random Fields (CRF)**
 - Conditional probability
 - Considers features of current token, and of preceding n tokens (window= n)
- **Similarity algorithms**
 - Measure distance of group of words to a dictionary list
 - Works especially well for jargon and other terminology
- **Support Vector Machines (SVM)**
 - Training method for standard perceptron
 - Optimize the points to determine the hyperplane dividing the positive training samples from the negative ones

A Perceptron is an algorithm used for supervised learning of binary classifiers.



More about Co-occurring Words

- Words appearing together in the same document
- Words appearing consecutively together, or in close vicinity in a defined relation, in the same document (expanded definition of *collocations*)

Great location with a little bit of history, the staff make this hotel though
Have a drink in the Long Bar, throw your nutshells on the floor, then go to the Tiffin Room for the best curry in the world. About £40a head for food but the choice is brilliant and when my wife mentioned it was her birthday at the end of the meal a cake was presented, what amazing service.

Great location with a little bit of history, the staff make this hotel though
Have a drink in the Long Bar, throw your nutshells on the floor, then go to the Tiffin Room for the best curry in the world. About £40a head for food but the choice is brilliant and when my wife mentioned it was her birthday at the end of the meal a cake was presented, what amazing service.



Detecting Collocation

- The collocations can be discovered by comparing the number of times two or more words appear together with the number of times they appear in other contexts.
- The challenge lies in separating those from words that randomly appear together.
- Common strategies:
 - Statistical hypothesis testing
 - Information theoretic analysis
 - Combination of a part of speech tagger and a simple frequency filter



Collocation Detection Strategies (1)

- Statistical hypothesis testing
 - We formulate a **null hypothesis** H_0 (no association beyond chance) and calculate the probability that a pair of words would co-occur if H_0 were true.
 - Then **reject** H_0 if p is too low, otherwise retain H_0 as possible.
 - The t-test, the chi-square test, and the likelihood ratio
 - Based on contingency table where the cells of the table contain the counts of times each appears both in isolation and together.
 - Collocations are those that are statistically unlikely to have occurred together by chance according to the test.



Collocation Detection Strategies (2)

- Information theoretic analysis
 - Based on the principles of information theory
 - Using *pointwise mutual information* metric as the scoring function
 - The amount of information provided by the occurrence of the event y about the occurrence of the event x
 - The amount of reduction in **uncertainty**

$$I(x, y) = \log_2 \frac{P(xy)}{P(x)P(y)} = \log_2 \frac{P(y|x)}{P(y)}$$

- If the appearance of the first word provides a strong indication that the second word will follow, then those two words are marked as a collocation.



Collocation Detection Strategies (3)

- Combination of a part of speech tagger and a simple frequency filter
 - Based on linguistic patterns of collocations, usually as combinations of nouns and adjectives
 - A part-of-speech tagger is required to find POS for each word, and identify phrases following such patterns.
 - Then apply sorting based on frequency to identify collocations
 - Examples of POS patterns
 - Adjective + noun** *third quarter*
 - Adjective + adjective + noun** *executive vice president*
 - Noun + preposition + noun** *earnings per share*



COREFERENCE RESOLUTION



What is coreference?

- **Coreference resolution**

- Determine relationship between entities which are related
 - Identity relation (morning star vs. evening star)
 - Whole-part relation
- Simple version
 - Determine entities which have the same referent
 - Anaphora (Pronouns)

The poem was a great example of anaphora as it started each line with the same three words.
 - Proper names, proper nouns, noun phrases,...
 - Definite descriptions (may be time dependent)
 - Usain Bolt & “the fastest man in the world”



Examples & Discussion

Coreference Examples

What do you need to know to do accurate co-reference?

- Anaphora
 - The elephant stepped on the rabbit and it died.
 - The elephant stepped on the landmine and it died.
- Proper nouns
 - John Smith and Mary Brown were married this morning. The groom was dressed in a white tuxedo while the bride was...
- Definite descriptions
 - Usain Bolt has won the Olympic 100m gold medal. The fastest man in the world successfully defended his title last night.



PRACTICAL INFORMATION EXTRACTION



How to get started

- **IE is the first step in further processing**
- **You need a representation structure**
- **You may need a reasoning engine**
- **You need to be able to automate**
- **You need to be able to experiment**
- **You need to iterate – try, try, try again**
- **Make sure you know how to code...**



Tools and resources

- **Many tools and resources on the internet**
- **Many libraries and tools are freely available, but may have restrictions for commercial use**
- **Some are tied to platforms**
- **It's never an “either or” situation**
It's always “and” – you *will* use R, Python, commercial software, databases, etc.



Reference & Resources

- **Ron Feldman**, *Information Extraction: Theory and Practice*, <http://cs.fit.edu/~pkc/icdm03/printing/tutorials/extraction/extraction.tutorial.pdf>
- **Fabio Ciravegna**, *Tutorial on Information Extraction from Text*, [http://www.isweb.uni-koblenz.de/files/ssms09/SSMS_Slides/ciravegna-IE text.pdf](http://www.isweb.uni-koblenz.de/files/ssms09/SSMS_Slides/ciravegna-IE_text.pdf)
- **Chris Manning & Hinrich Schutze**, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999
- **NLP resources**: <http://nlp.stanford.edu/links/statnlp.html>