

Information Extraction

Jaime Arguello

INLS 613: Text Data Mining

jarguell@email.unc.edu

November 2, 2016

Information Extraction

- **Task:** extracting structured information from unstructured (and loosely formatted) natural language text

Information Extraction

named-entity recognition

Barack Obama was born in Honolulu, Hawaii. Obama is a graduate of Columbia University and Harvard Law School, where he was president of the Harvard Law Review. He was a community organizer in Chicago before earning his law degree. He worked as a civil rights attorney in Chicago and taught constitutional law at the University of Chicago Law School from 1992 to 2004. He served three terms representing the 13th District in the Illinois Senate from 1997 to 2004, running unsuccessfully for the United States House of Representatives in 2000.

Information Extraction

named-entity recognition

Barack Obama was born in Honolulu, Hawaii. Obama is a graduate of Columbia University and Harvard Law School, where he was president of the Harvard Law Review. He was a community organizer in Chicago before earning his law degree. He worked as a civil rights attorney in Chicago and taught constitutional law at the University of Chicago Law School from 1992 to 2004. He served three terms representing the 13th District in the Illinois Senate from 1997 to 2004, running unsuccessfully for the United States House of Representatives in 2000.

Information Extraction

entity linking or co-reference resolution

Barack Obama was born in Honolulu, Hawaii. Obama is a graduate of Columbia University and Harvard Law School, where he was president of the Harvard Law Review. He was a community organizer in Chicago before earning his law degree. He worked as a civil rights attorney in Chicago and taught constitutional law at the University of Chicago Law School from 1992 to 2004. He served the Illinois State House of Representatives from 1997 to 2004, running unsuccessfully for the U.S. House of Representatives in 2000.



Information Extraction

relation extraction

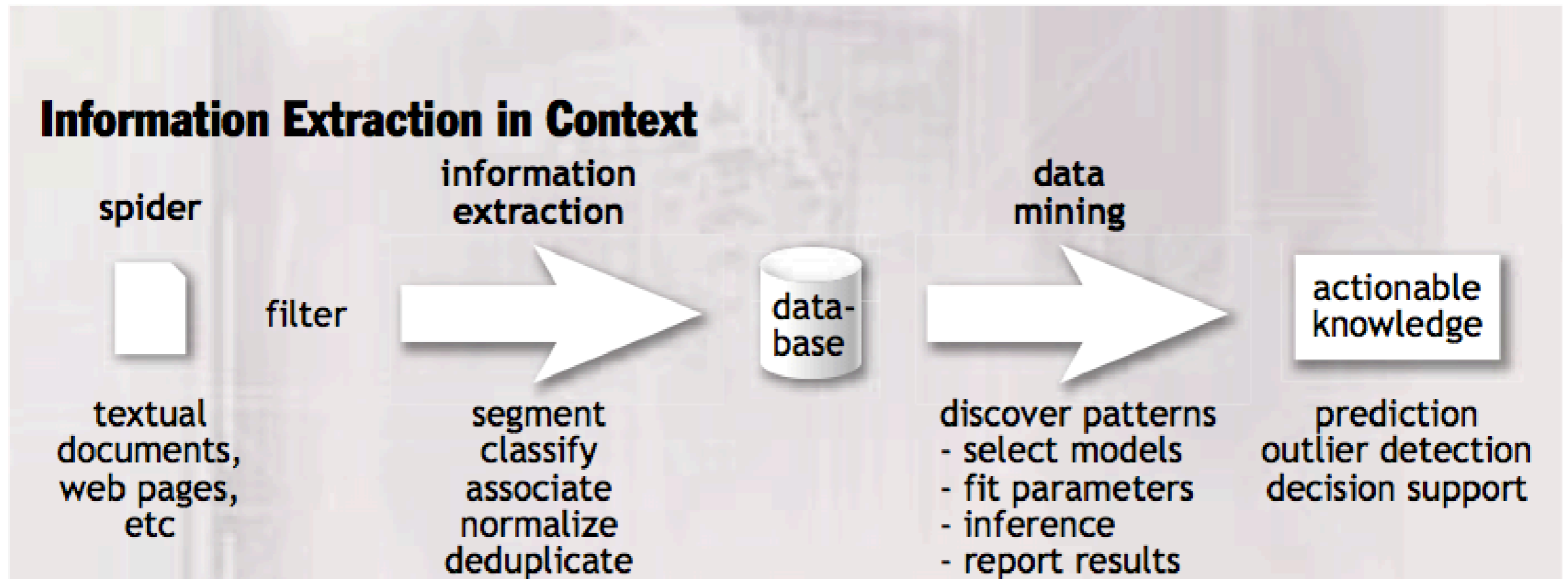
Barack Obama was born in Honolulu, Hawaii. Obama is a graduate of Columbia University and Harvard Law School, where he was president of the Harvard Law Review. He was a community organizer in Chicago before earning his law degree. He worked as a civil rights attorney in Chicago and taught constitutional law at the University of Chicago Law School from 1992 to 2004. He served three terms representing the 13th District in the Illinois Senate from 1997 to 2004, running unsuccessfully for the United States House of Representatives in 2000.

BornIn(Barack Obama, Honolulu)

GraduatedFrom(Obama, Columbia University)

Information Extraction

process



(Source: McCallum 2005)

Information Extraction

applications: search

chapel hill mexican food



Monterrey Chapel Hill

www.monterreychapelhill.com/

Score: **24** / 30 - 31 Google reviews

Bandido's

bandidoscafe.com/

Score: **14** / 30 - 14 Google reviews

Fiesta Grill

www.fiestagrill.us/

Score: **28** / 30 - 28 Google reviews

Cinco De Mayo

www.cincodemayorestaurants.net/

Score: **20** / 30 - 23 Google reviews

Los Potrillos

plus.google.com

Score: **17** / 30 - 13 Google reviews

La Hacienda

www.lahaciendaofchapelhill.com/

Score: **15** / 30 - 19 Google reviews

Qdoba Mexican Grill

www.qdoba.com/

Score: **11** / 30 - 24 Google reviews

A 237 South Elliot Road
Chapel Hill
(919) 969-8750

B 159 1/2 East Franklin
Street
Chapel Hill
(919) 967-5048

C 3307 North Carolina 54
Chapel Hill
(919) 928-9002

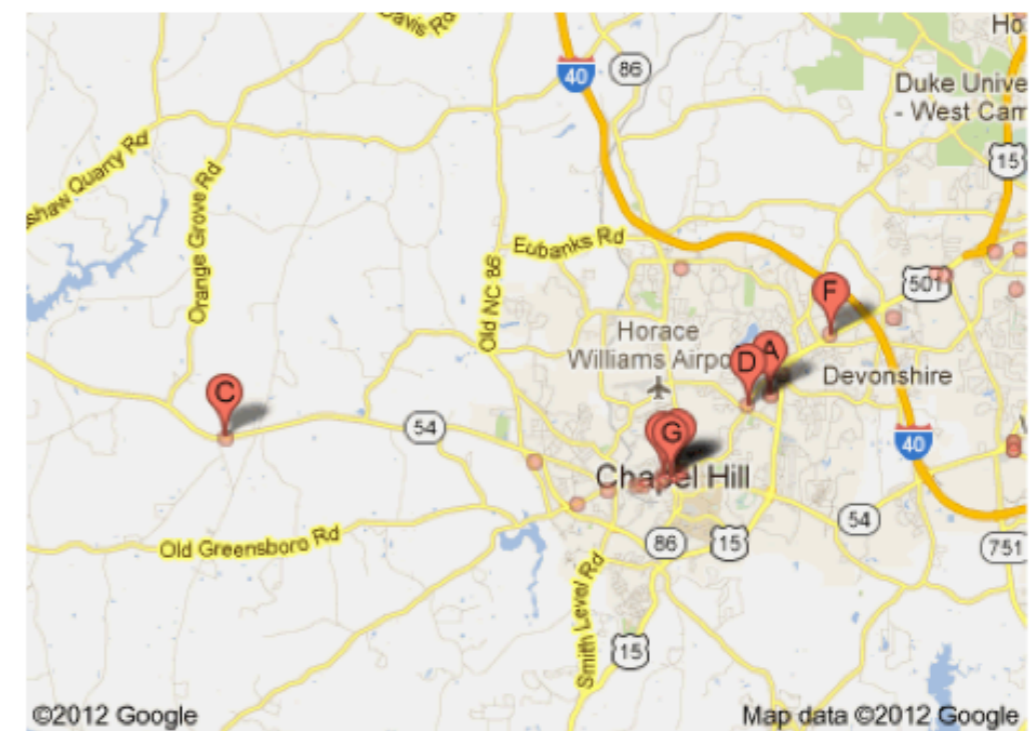
D 1502 East Franklin
Street
Chapel Hill
(919) 929-6566

E 220 West Rosemary
Street
Chapel Hill
(919) 932-4301

F 1813 Fordham
Boulevard
Chapel Hill
(919) 967-0207

G 100 West Franklin Street
Chapel Hill
(919) 929-8998





Map for chapel hill mexican food



Information Extraction

applications: search

[Shop for camera on Google](#) Sponsored ⓘ

			
Nikon Coolpix L26 16.1-me...	Nikon D3100 14.2MP Digit...	Nikon Coolpix L810 16.1-m...	Canon PowerShot S...
\$99.99	\$476.95	\$229.99	\$429.99
Crutchfield	Newegg.com	Crutchfield	Target

Shop by brand: [Canon](#) [Nikon](#) [Sony](#) [Panasonic Lumix](#) [Fujifilm](#)

Places for camera near Chapel Hill, NC

Best Buy - Durham-Raleigh stores.bestbuy.com Score: 8 / 30 - 16 Google reviews	A 5454 New Hope Commons Drive Durham (919) 403-2333
Best Buy - Durham stores.bestbuy.com 3 Google reviews	B 7001 Fayetteville Street Durham (919) 544-4354
Walgreens Store Chapel Hill www.walgreens.com Google+ page	C 1500 East Franklin Street Chapel Hill (919) 918-4392

[More results near Chapel Hill, NC »](#)

Information Extraction

applications: search

- Named entities are important index terms
- Document linking (e.g., to background page)
- Automatic summary generation
- In question answering, answers are often named entities
- Vertical search engine selection
 - ▶ local search, product search, calculator
-

Information Extraction

applications: data mining

- Citation/link analysis
- Co-occurrence mining
- Event detection
- Outlier detection
- Feature generation for predictive analysis

Named Entity Recognition

methods

- Simple regular expressions (e.g., email addresses)
 - ▶ `[A-Z0-9._%+-]+@[A-Z0-9.-]+\.[A-Z]{2,4}`
- Hand-written rules
 - ▶ send email to _____.
- Machine learning
 - ▶ use features generated from the candidate named entity, the formatting, and/or the context

Named Entity Recognition

classification

- Use noun-phrase chunking to first determine candidate entities (noun-phrases), e.g.,
- Obama is a graduate of Columbia University and Harvard Law School, where he was president of the Harvard Law Review.
- Train and apply independent binary classifiers (one per named-entity type) to predict whether a candidate entity belongs to a particular type.

Named Entity Recognition

features: combining representations

person?

A diagram consisting of a blue-outlined rectangular box at the top containing the text "person?". From the bottom center of this box, several blue arrows point downwards to various words in the sentence below. The arrows point to "Obama", "graduate", "Columbia", "University", "where", "he", "was", and "president".

- Obama is a graduate of Columbia University and Harvard Law School, where he was president of the Harvard Law Review.
- What types of features would you use?

Named Entity Recognition

features: combining representations

- Obama is a graduate of Columbia University and Harvard Law School, where he was president of the Harvard Law Review.
- **Capitalization:** Xx^* vs. XX^* vs. xx^*
- **Surrounding words:**
 - ▶ $w-2 = \langle \text{null} \rangle$ (adds $|V|$ features)
 - ▶ $w-1 = \langle \text{null} \rangle$ (adds $|V|$ features)
 - ▶ $w+1 = \text{is}$ (adds $|V|$ features)
 - ▶ $w+2 = \text{a}$ (adds $|V|$ features)

Named Entity Recognition

features: combining representations

- Obama is a graduate of Columbia University and Harvard Law School, where he was president of the Harvard Law Review.
- **Part-of-speech features:** Obama/NNP is/VBZ a/DT graduate/NN of/IN Columbia/NNP University/NNP and/CC Harvard/NNP Law/NNP School/NNP ,/, where/WRB he/PRP was/VBD president/NN of/IN the/DT Harvard/NNP Law/NNP Review/NNP ./.

Named Entity Recognition

features: combining representations

- Obama is a graduate of Columbia University and Harvard Law School, where he was president of the Harvard Law Review.
- Parse-tree Features:

```
(ROOT
 (S
  (NP (NNP Obama))
  (VP (VBZ is)
   (NP
    (NP (DT a) (NN graduate))
    (PP (IN of)
     (NP
      (NP (NNP Columbia) (NNP University))
      (CC and)
      (NP (NNP Harvard) (NNP Law) (NNP School))))))
  (, ,)
  (SBAR
   (WHADVP (WRB where))
   (S
    (NP (PRP he))
    (VP (VBD was)
     (NP
      (NP (NN president))
      (PP (IN of)
       (NP (DT the) (NNP Harvard) (NNP Law) (NNP Review))))))))
  (. .)))
```

Named Entity Recognition

classification

- Information extraction can be treated as a classification task
- Requires doing some preprocessing to identify candidate named-entities
- Requires training data in the form of annotations (is_person = yes, is_person = no)
- Requires deriving features from the entity itself, the local context, the “not-so-local” context (i.e., document category features), and the parse tree
- There is another form of extraction learning referred to as “bootstrapping”.