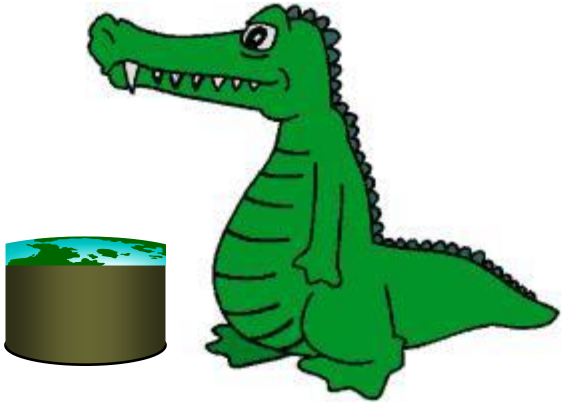
CAP4770/5771
Introduction to Data Science
Fall 2016

University of Florida, CISE Department
Prof. Daisy Zhe Wang



Logistics

- Lab 4 grades and keys released
- Lab 5 material released last Friday, due this Thursday 11:59pm
  - JAVA + AWS/EMR
- Lab 5 in class this Wed.
  - Extra prep: AWS, EMR setup and tutorial
- NIST DSE Introduction + QA this Friday
- No office hour this Wed.

Review

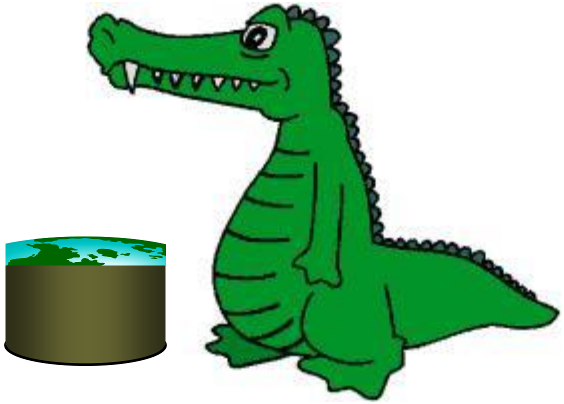
- Distributed/Parallel Computing
- Distributed Files Systems
- Map-Reduce Programming Model

Text Processing, Classification and NLP

Basic Text Processing

Classification

Information Extraction



Definitions

- Information Extraction (IE)
  - Extracting existing facts from unstructured or loosely structured text into a structured form (e.g., KBP)
- Information Retrieval (IR)
  - Finding documents relevant to a user query (e.g., Search)
- Named Entity Recognition (NER)
  - Discovery of groups of textual mentions that belong to certain semantic class (e.g., PER, ORG)
- Natural Language Processing (NLP)
  - Computational methods for text processing based on linguistically sound principles
  - Clinical NLP – NLP for the clinical narrative
  - Biomedical NLP – NLP for the clinical narrative and biomedical literature

Problems that can be solved by NLP Techniques

- Mostly Solved
  - Spam filtering, POS tagging, NER
  - IR, spelling correction
- Making Good Progress
  - Sentiment analysis, IE (relations, events)
  - WSD, Parsing, MT, coreference
- Still Hard
  - QA, Dialog
  - Paraphrase, summarization

Main classes of NLP techniques

- Basic Text Processing: Segmentation, Normalization, POS Tagging…
- Classification: Bag of Words, Ngrams Model, Naïve Bayes…
- Information Extraction: Named Entity Recognition, Relation Extraction, Event Extraction…

POS Tagging: Example

I	saw	the	nurse	with	the	doctor.
w1	w2	w3	w4	w5	w6	w7
pronoun	verb	article	noun	prep	article	noun

Mr. Smith feels dizzy and has a cold sweat.

Classification: Example

- Effectiveness of medication:
  - Classes: High, Moderate, Low

"I feel much better after taking the medicine for 1 month with more energy and less pain…

"I do not feel any better after taking the medicine. I have stopped using it…

"I feel a little better with the medication and I will try it a little longer to see…

Information Extraction: Example

- "Tamoxifen 20 mg po daily started on March 1, 2005."
  - Drug (with predefined schema/attributes)
    - Text: Tamoxifen
    - Associated code: C0351245
    - Strength: 20 mg
    - Start date: March 1, 2005
    - End date: null
    - Dosage: 1.0
    - Frequency: 1.0
    - Frequency unit: daily
    - Duration: null
    - Route: Enteral Oral
    - Form: null
    - Status: current
    - Change Status: no change
    - Certainty: null

NLP methods

- Rule-based
  - Regular Expression Pattern matching (e.g., “\b(lipitor|Lipitor)\b”)
  - Dictionaries (e.g., drug names, ICD10 codes)
- Statistical Machine Learning
  - HMM: Hidden Markov Models
  - Linear-CRF: Linear-Chain Conditional Random Fields
  - Viterbi algorithms
- Hybrid

Outline

- NLP Tasks and Techniques
  - Regular Expression
  - Normalization, POS, NER, etc.
  - Sequence Labeling

Regular expressions

- A formal language for specifying text strings
- How can we search for any of these?
  - woodchuck
  - woodchucks
  - Woodchuck
  - Woodchucks

Regular Expressions: Disjunctions

- Letters inside square brackets []

Pattern	Matches
[wW]oodchuck	Woodchuck, woodchuck
[1234567890]	Any digit

- Ranges [A-Z]

Pattern	Matches	
[A-Z]	An upper case letter	Drenched Blossoms
[a-z]	A lower case letter	my beans were impatient
[0-9]	A single digit	Chapter 1: Down the Rabbit Hole

Regular Expressions: Negation in Disjunction

- Negations `[^Ss]`
  - Carat means negation only when first in []

Pattern	Matches	
`[^A-Z]`	Not an upper case letter	`Oyfn pripetchik`
`[^Ss]`	Neither 'S' nor 's'	`I have no exquisite reason`
`[^e^]`	Neither e nor ^	`Look here`
`a^b`	The pattern a carat b	`Look up a^b now`

Regular Expressions: More Disjunction

- Woodchucks is another name for groundhog!
- The pipe | for disjunction

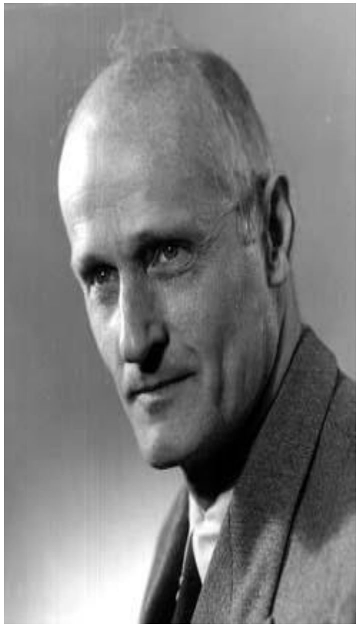
<table>
<tr><th>Pattern</th><th>Matches</th></tr>
<tr><td>groundhog|woodchuck</td><td></td></tr>
<tr><td>yours|mine</td><td>yours
mine</td></tr>
<tr><td>a|b|c</td><td>= [abc]</td></tr>
<tr><td>[gG]roundhog|[Ww]oodchuck</td><td></td></tr>
</table>

Photo D. Fletcher

Wildcards in Regular Expressions:

? * + .

Pattern	Matches	
colou?r	Optional previous char	color colour
oo*h!	0 or more of previous char	oh! ooh! oooh! ooooh!
o+h!	1 or more of previous char	oh! ooh! oooh! ooooh!
baa+		baa baaa baaaa baaaaa
beg.n		begin begun begun beg3n



Stephen C Kleene

Kleene *, Kleene +

Anchors in Regular Expressions: ^ $

Pattern	Matches
`^[A-Z]`	Palo Alto
`^[^A-Za-z]`	1 "Hello"
`\.$`	The end.
`.$`	The end? The end!

Example

- Find me all instances of the word “the” in a text.

  `the`

  Misses capitalized examples

  `[tT]he`

  Incorrectly returns `other` or `theology`

  `[^a-zA-Z][tT]he[^a-zA-Z]`

Errors

- The process we just went through was based on fixing two kinds of errors
  - Matching strings that we should not have matched (there, then, other)
    - False positives (Type I)
  - Not matching things that we should have matched (The)
    - False negatives (Type II)

Errors cont.

- In NLP/ML we are always dealing with these kinds of errors.
- Reducing the error rate for an application almost always involves both:
  - Increasing accuracy or precision (minimizing false positives)
  - Increasing coverage or recall (minimizing false negatives).

Regular Expression vs. Machine Learning in NLP

- Regular expressions play a surprisingly large role
  - Sophisticated sequences of regular expressions are often the first model for any text processing text
- For many hard tasks, we use machine learning classifiers
  - But regular expressions are used as features in the classifiers
  - Can be very useful in capturing generalizations

Basic Text Processing: Text Normalization

- Every NLP task needs to do text normalization:
  1. Segmenting/tokenizing words in running text
  2. Normalizing word formats (e.g., lemmatization, stemming)
  3. Segmenting sentences in running text
     1. **Regular Expression**
     2. **Classifier (e.g., decision tree)**

Basic Text Processing: POS Tagging and Chunking

- UPenn Treebank POS: noun, verb, adjective, adverb, pronoun, preposition, conjunction
  - https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
- Regular Expression and Dictionary Based POS Tagging
- Machine Learning Sequence Models (e.g., CRF, HMM)

Information Extraction and Named Entity Recognition

Introducing the tasks:

Getting simple structured information out of text

Information Extraction

- Information extraction (IE) systems
  - Find and understand limited relevant parts of texts
  - Gather information from many pieces of text
  - Produce a structured representation of relevant information:
    - *relations* (in the database sense), a.k.a.,
    - a *knowledge base*
  - Goals:
    1. Organize information so that it is useful to people
    2. Put information in a semantically precise form that allows further inferences to be made by computer algorithms

Information Extraction (IE)

- IE systems extract clear, factual information
  - Roughly: *Who did what to whom when?*
- E.g.,
  - Gathering earnings, profits, board members, headquarters, etc. from company reports
    - The headquarters of BHP Billiton Limited, and the global headquarters of the combined BHP Billiton Group, are located in Melbourne, Australia.
    - headquarters("BHP Biliton Limited", "Melbourne, Australia")
  - Learn drug-gene-product interactions from medical research literature

6 cTakes object templates with their attributes

Medication CEM template

associatedCode
Change_status
Conditional
Dosage
Duration
End_date
Form
Frequency
Generic
Negation_indicator
Route
Start_date
Strength
Subject
Uncertainty_indicator

Sign/Symptom CEM template

Alleviating_factor
associatedCode
Body_laterality
Body_location
Body_side
Conditional
Course
Duration
End_time
Exacerbating_factor
Generic
Negation_indicator
Relative_temporal_context
Severity
Start_time
Subject
Uncertainty_indicator

Disease/Disorder CEM template

Alleviating_factor
Associated_sign_or_symptom
associatedCode
Body_laterality
Body_location
Body_side
Conditional
Course
Duration
End_time
Exacerbating_factor
Generic
Negation_indicator
Relative_temporal_context
Severity
Start_time
Subject
Uncertainty_indicator

Procedure CEM template

associatedCode
Body_laterality
Body_location
Body_side
Conditional
Device
End_date
Generic
Method
Negation_indicator
Relative_temporal_context
Start_date
Subject
Uncertainty_indicator

Lab CEM template

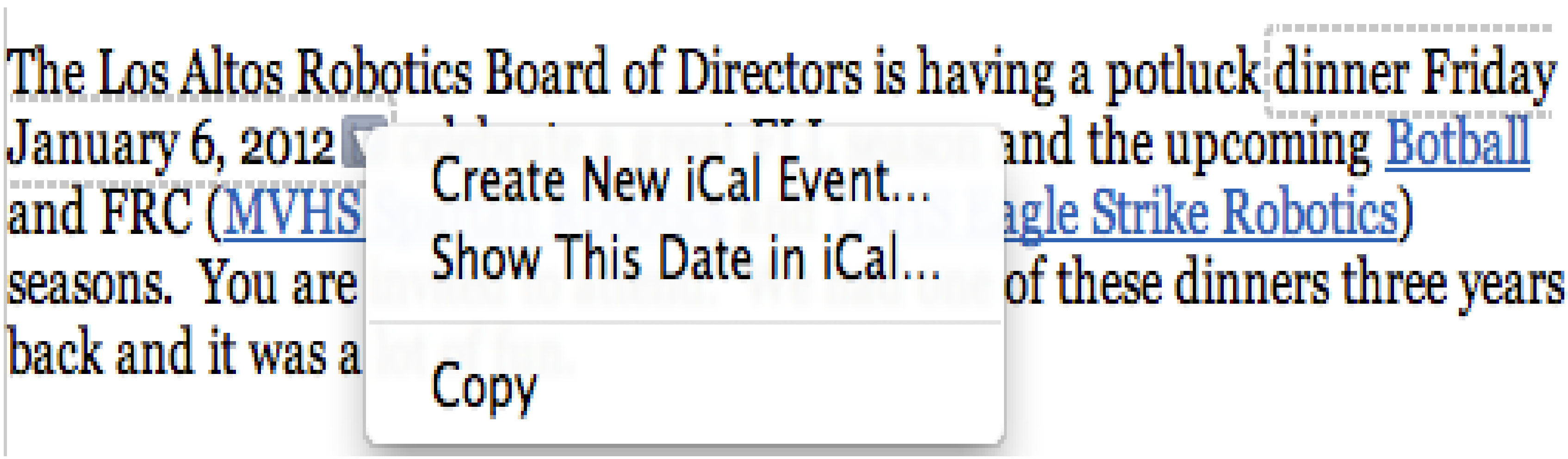
Abnormal_interpretation
associatedCode
Conditional
Delta_flag
Estimated_flag
Generic
Lab_value
Negation_indicator
Ordinal_interpretation
Reference_range_narrative
Subject
Uncertainty_indicator

Anatomical Site CEM template

associatedCode
Body_laterality
Body_site
Conditional
Generic
Negation_indicator
Subject
Uncertainty_indicator

Low-level information extraction

- Is now available – and I think popular – in applications like Apple or Google mail, and web indexing



- Often seems to be based on regular expressions and name lists

Low-level information extraction

Named Entity Recognition (NER)

- A very important sub-task: find and classify names in text, for example:
  - The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Named Entity Recognition (NER)

- A very important sub-task: find and classify names in text, for example:
  - The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Named Entity Recognition (NER)

- A very important sub-task: find and classify names in text, for example:
  - The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Person

Location

Organi-
zation

Named Entity Recognition (NER)

- The uses:
  - Named entities can be indexed, linked off, etc.
  - Sentiment can be attributed to companies or products
  - A lot of IE relations are associations between named entities
  - For question answering, answers are often named entities.
- Concretely:
  - Many web pages tag various entities, with links to bio, wikipedia or topic pages, etc.
    - Reuters' OpenCalais, Evri, AlchemyAPI, Yahoo's Term Extraction, …
  - Apple/Google/Microsoft/… smart recognizers for document content

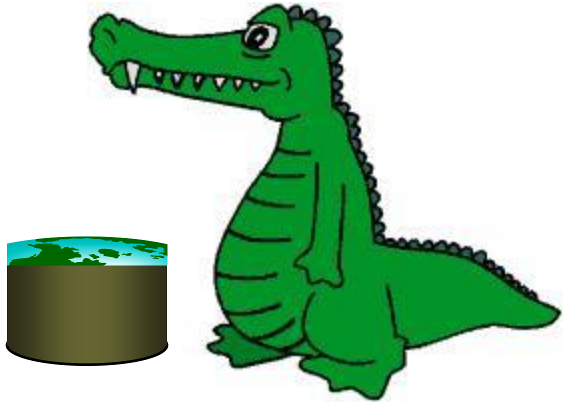
The Named Entity Recognition Task

- Task: Extract named entities in a text
- Sequence labeling: label each token with ORG/PER/.../O
- One way you could evaluate is per-token

Foreign	ORG
Ministry	ORG
spokesman	O
Shen	PER
Guofang	PER
told	O
Reuters	ORG
:	:

Standard evaluation is per entity, *not* per token

Sequence Models for Named Entity Recognition



The ML sequence model approach to NER

Training

1. Collect a set of representative training documents
2. Label each token for its entity class or other (O)
3. Design feature extractors appropriate to the text and classes
4. Train a sequence classifier to predict the labels from the data

Testing

1. Receive a set of testing documents
2. Run sequence model inference to label each token
3. Appropriately output the recognized entities
4. Evaluate precision/recall i.e., type I/II errors

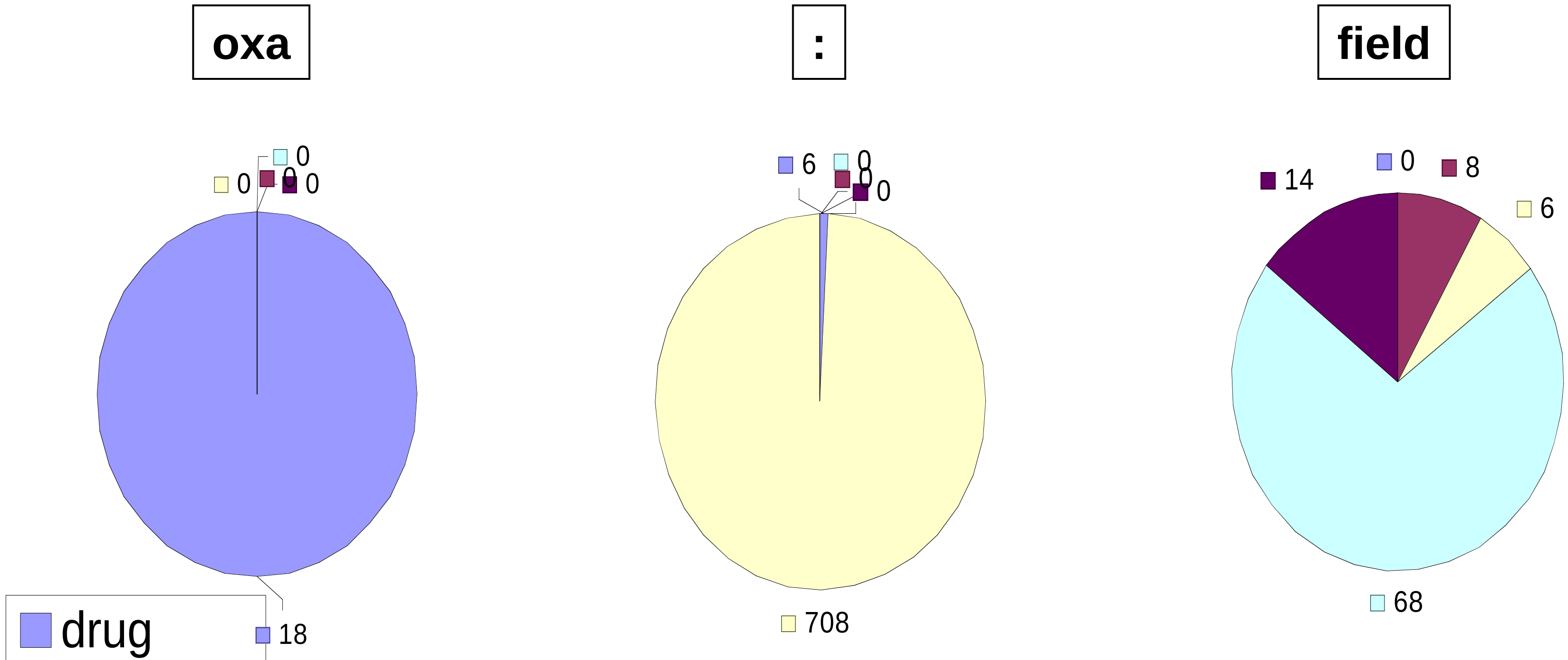
Encoding classes for sequence labeling

	IO encoding	IOB encoding
Fred	PER	B-PER
showed	O	O
Sue	PER	B-PER
Mengqiu	PER	B-PER
Huang	PER	I-PER
‘s	O	O
new	O	O
painting	O	O

Features for sequence labeling

- Words
  - Current word: caps, regular expressions, digits, dictionaries, substrings
  - Previous/next word (context)
- Other kinds of inferred linguistic classification
  - Part-of-speech tags
- Label context
  - Previous (and perhaps next) label

Features: Word substrings



Cotrimoxazole

Wethersfield

Alien Fury: Countdown to Invasion

Features: Word shapes

- Word Shapes
  - Map words to simplified representation that encodes attributes such as length, capitalization, numerals, Greek letters, internal punctuation, etc.

Varicella-zoster	Xx-xxx
mRNA	xXXX
CPA1	XXXd