NATURAL LANGUAGE PROCESSING

LU XIAO
LXIA004@SYR.EDU
213 HINDS HALL

ADOPTED SOME MATERIALS DEVELOPED IN PREVIOUS COURSES BY NANCY MCCRACKEN, LIZ LIDDY AND OTHERS; AND SOME INSTRUCTOR RESOURCES FOR THE BOOK "SPEECH AND LANGUAGE PROCESSING" BY DANIEL JURAFSKY AND JAMES H. MARTIN

NATURAL LANGUAGE PROCESSING (NLP)

- A range of computational techniques:
  - for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis
  - for the purpose of achieving human-like language processing
  - for a range of particular tasks or applications.

- Computational Linguistics – doing linguistics on computers
  - Closely related, often treated as synonymous with NLP

WHERE IS NLP NOW?

- Goals can be far-reaching
  - True text understanding
  - Reasoning about knowledge in text
  - Real-time participation in spoken dialogs
- Or very down-to-earth
  - Finding the price of products on the web
  - Context-sensitive spell-checking
  - Analyzing authorship or opinions statistically
  - Extracting facts or relations from documents
  - Remembering previous searches and contexts to guide future interactions
- Currently, NLP is providing these practical applications (yet still dreaming of the AI goals)

NLP APPLICATION AREAS

- Information Retrieval / Search Engines – provision of documents containing requested information
  - Google, many other search engines
  - Use lowest levels of NLP to stem words, find phrases for indexing documents
  - Users conform to keyword query restriction, instead of natural language queries

- Machine Translation – conversion of text from one language to another
  - Usefulness of Parallel Corpora
  - Often statistically based patterns of word usage and context
  - Google, Yahoo and Bing all have language translators
  - MT techniques use context , not just word for word substitution

GOOGLE TRANSLATE

https://www.nounou-top.fr/conversations/1437830

This page is in French Would you like to translate it? Nope Translate Options

Camille

il y a plus d'un an

Bonjour,
Je me propose pour garder vos enfants, et je vous invite à aller voir mon profil. Cependant, je ne suis disponible qu'en juin et juillet, mais pas en août. Si cela vous intéresse quand même, merci de me répondre, mais je trouverais ça normal si vous choisissiez une babysitter disponible les deux mois.

Cordialement,
Camille

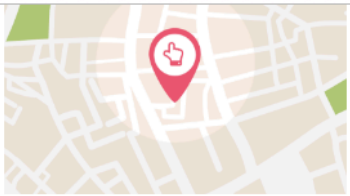
Lu

il y a plus d'un an

Hi Camille,

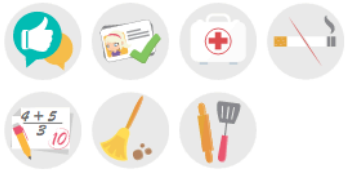
Thank you for contacting me. It's ok if you can only help in July. From your profile, it seems that you charge a rate of 5.00 euro per hour for regular service, right? I hope to have someone to take care of the children from 8:30 - 4:30 (8hours) per day, and 4 days a week. So it will be 32 hours per week, 128 hours in July. With your rate, it will be 640 Euro for the month. Could you confirm this? You seem to live in a different place (we will live in the city of Grenoble). The public transportation is very convenient, I was told. But I am not sure if I have the budget to pay for your transportation cost. Would this be ok? thanks.

Vous



(470 m de chez vous)

Ses plus :



Voir le profil complet

NLP APPLICATION AREAS

- Information Extraction / Text-mining – populating a structured database with specific bits of information found in text
  - Competitive Intelligence analyzes news text and web blogs for
    - Names of people, companies and other entities
    - Relations between them, e.g. corporate roles, or events such as mergers

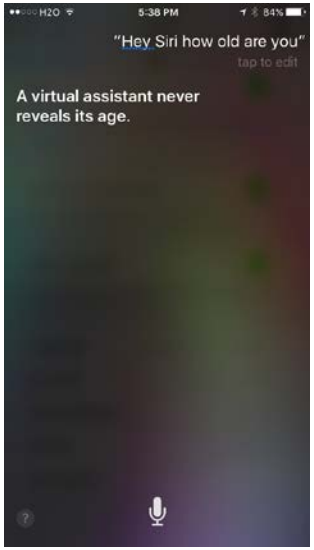
Weblog Analytics

Data-mining of Weblogs, discussion forums, message boards, user groups, and other forms of user generated media

- Product marketing information
- Political opinion tracking
- Social network analysis
- Buzz analysis (what's hot, what topics are people talking about right now).

NLP APPLICATION AREAS

- Human-computer Interfaces – NLP assistants, chatbots, interactive querying of databases



- Summarization – abstraction and condensation of text's major points
  - Current systems select a set of significant sentences from the document as a summary
  - Example summarizer: http://textsummarization.net/text-summarizer

NLP APPLICATION AREAS

- Question & Answering Systems – focused information provision

- Metadata Generation – assignment of values for metadata elements in a particular standard

Elizabeth D. Liddy, Eileen Allen, Sarah Harwell, Susan Corieri, Ozgur Yilmazel, N. Ercan Ozgencil, Anne Diekema, Nancy McCracken, Joanne Silverstein, and Stuart Sutton. 2002. Automatic metadata generation & evaluation. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '02). ACM, New York, NY, USA, 401-402. DOI=http://dx.doi.org/10.1145/564376.564464

THE TRENDS

1. An enormous amount of knowledge is now available in machine readable form as natural language text
2. Conversational agents are becoming an important form of human-computer communication
3. Much of human-human communication is now mediated by computers

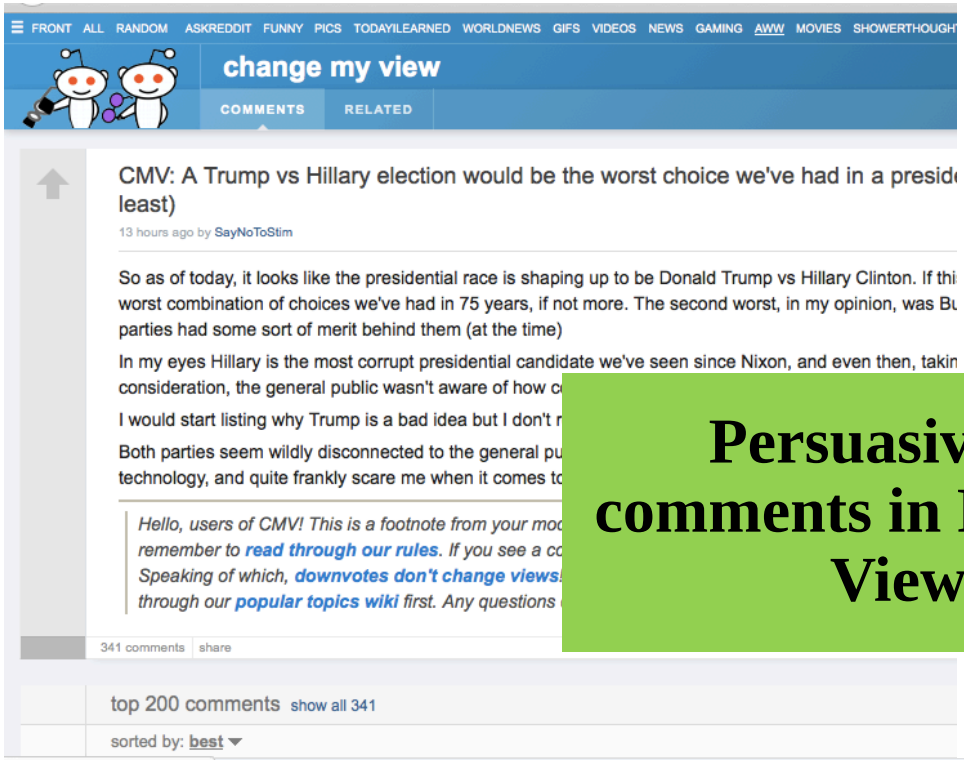
NEED FOR (MORE) NLP WORK

- Huge amounts of data
  - Internet
  - Intranet
- Applications for processing large amounts of texts require NLP expertise
- Data Science/Text Mining

Classify text into categories
Index and search large texts
Automatic translation of web documents in different languages
Speech understanding
Understand phone conversations
Information extraction
Extract useful information from resumes
Automatic summarization
Condense 1 book into 1 page
Daily news summaries
Question answering
Knowledge acquisition
Text generations / dialogues

Detect Representative Rationales and Imperatives from Wikipedia's Article for Deletion (AfD) Discussions

Persuasive power of online comments in Reddit's "Change My View" discussions



WHY IS NLP SO HARD?

- **Seems pretty simple for humans**
  - Usually quite unaware of the complexity of the language tasks they perform so effortlessly
- **Some reasons are**
  - Subleties of meaning
    - Irony, sarcasm, humor, metaphor
  - Ambiguity
    - Ambiguity is a fundamental problem of computational linguistics
    - Resolving ambiguity is a crucial goal

AMBIGUITY

- Find at least 5 meanings of this sentence:
  - **I made her duck**

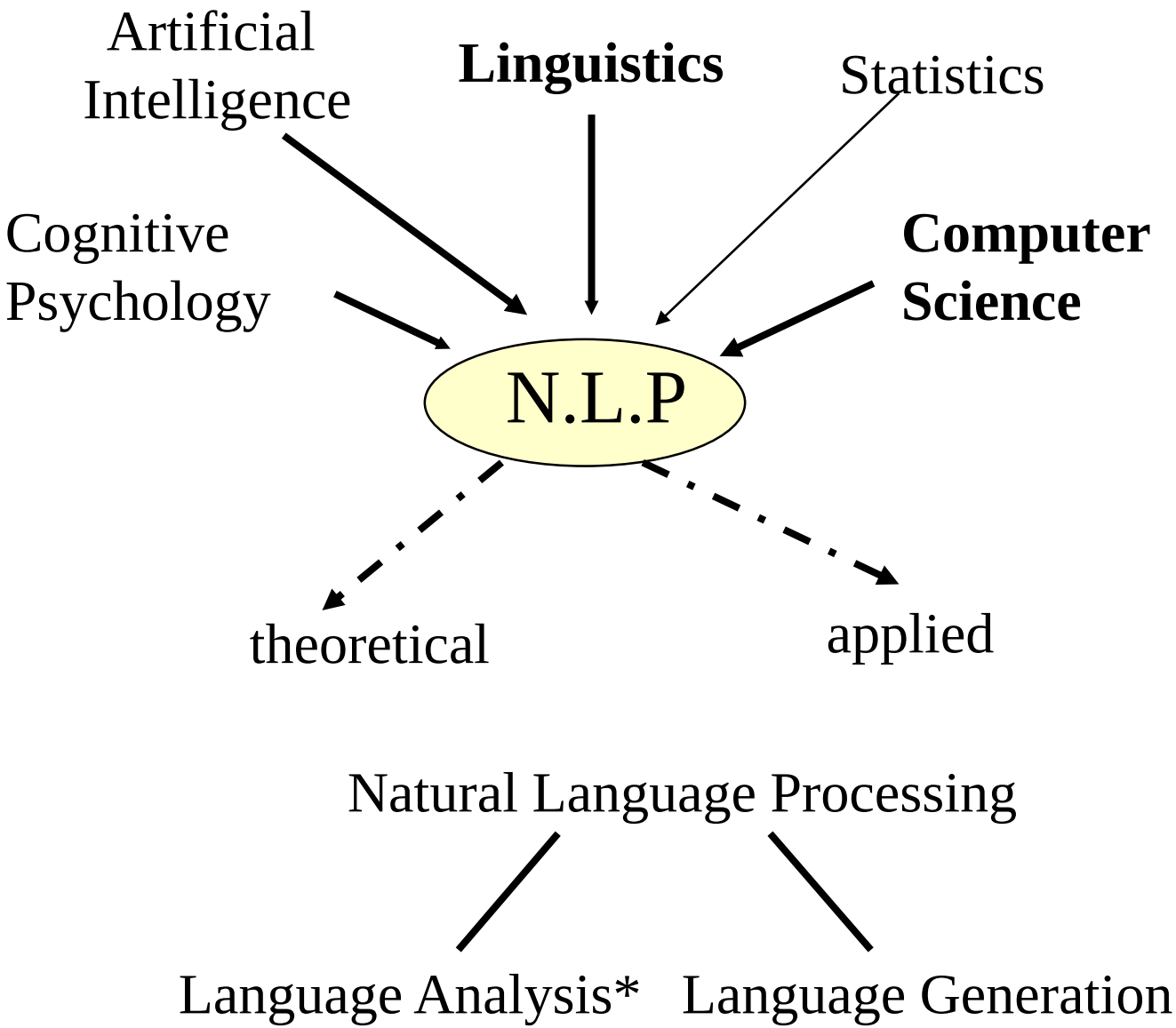
- I cooked waterfowl for her benefit (to eat)
- I cooked waterfowl belonging to her
- I created the (plaster?) duck she owns
- I caused her to quickly lower her head or body
- I waved my magic wand and turned her into undifferentiated waterfowl

AMBIGUITY IS PERVASIVE

- I caused her to quickly lower her head or body
  - **Lexical category**: "duck" can be a N or V
- I cooked waterfowl belonging to her.
  - **Lexical category**: "her" can be a possessive ("of her") or dative ("for her") pronoun
- I made the (plaster) duck statue she owns
  - **Lexical Semantics**: "make" can mean "create" or "cook"

AMBIGUITY IS PERVASIVE

- **Grammar**: Make can be:
  - **Transitive: (verb has a noun direct object)**
    - I cooked [waterfowl belonging to her]
  - **Ditransitive: (verb has 2 noun objects)**
    - I made [her] (into) [undifferentiated waterfowl]
  - **Action-transitive (verb has a direct object and another verb)**
  - I caused [her] [to move her body]



*Main emphasis in this course

CATEGORIES OF KNOWLEDGE

- Phonology
- Morphology
- Syntax
- Semantics
- Pragmatics
- Discourse

COURSE OUTLINE

- Lectures
  - Mar. 26 – away for a conference trip; Ph.D. student Jenna Kim will deliver the lecture and hold the lab session
- Jenna will also hold office hours in the two weeks before the assignment due date to help you on the homework
  - **Wednesdays: 1:00 – 5:00 pm**
  - **Thursdays: 10:00 am – 5:00 pm**

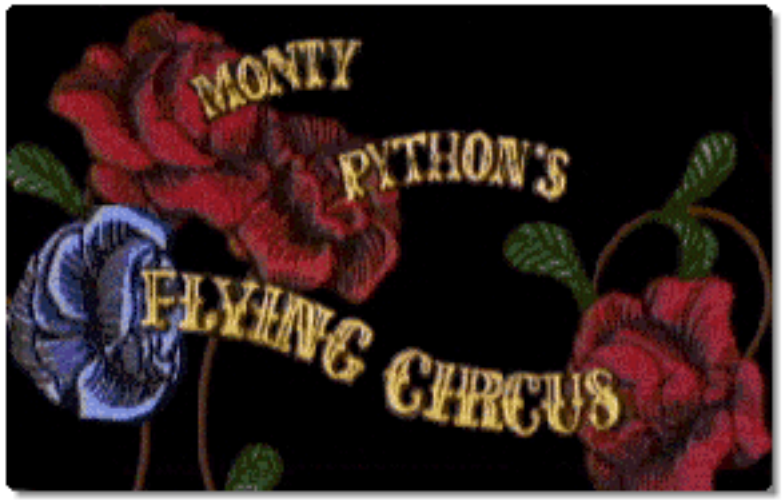
METHODS OF ASSESSMENT

Assessment method	Due Dates (submission will be due by 11:59 pm. of the day unless otherwise noted)
Class Participation	Notes: this includes participation in weekly lecture and Python exercises, and contributions to class discussions
Homework assignments	1st: Feb. 18 2nd: Mar. 18 3rd: Apr. 1 4th: Apr. 22
NLP Application Investigations	Group report: Apr. 29 Group poster: Apr. 29

NATURAL LANGUAGE TOOLKIT (NLTK) AND PYTHON

PYTHON AND NLP

- Python is freely available for many platforms from the Python Software Foundation:
  - http://www.python.org/
  - Named for the group Monty Python
  - We are using Python version 3.x
    - (not backward compatible with Python 2.x)



The group in 1969

CHARACTERISTICS OF PYTHON

- Easy-to-learn scripting language, similar in many aspects to Perl, but with WYSIWYG block structure
- Object-oriented, with modules, classes, exceptions, high-level dynamic data types, similar to Java
- Strongly typed, but without type declarations (dynamic typing)
- Regular Expressions and other string processing features
- Many libraries offer wide functionality:
  - https://xkcd.com/353/

[optional reading about strongly typed and dynamic typed:
http://stackoverflow.com/questions/11328920/is-python-strongly-typed

https://wiki.python.org/moin/Why%20is%20Python%20a%20dynamic%20language%20and%20also%20a%20strongly%20typed%20language

GETTING STARTED IN PYTHON

- Python can be run as an interactive system
  - Type in expressions or small pieces of programs to try them out
- or as a command-line system.
  - Run stored python programs
- For both, it is recommended to use a Python development environment
  - IDLE is standard but really simple: especially good to edit Python programs in IDLE to keep track of the indentation for block structure
    - Or try Wing free version, PyCharm or iPython to get an IDE

NATURAL LANGUAGE TOOLKIT (NLTK)

- A suite of Python libraries for symbolic and statistical natural language programming
  - Developed at the University of Pennsylvania
- Developed to be a teaching tool and a platform for research NLP prototypes
  - Data types are packaged as classes
  - Goal of code is to be clear, rather than fastest performance
    - But increasingly production level software is made available through wrappers
- Latest version is compatible with Python 3.x
- Online book:
  http://www.nltk.org/book/
- Authors:
  Edward Loper, Ewan Kline
  and Steven Bird

USING NLTK IN NLP

- NL ToolKit provides libraries of many of the common NLP processes at various language levels
  - Leverage these libraries to process text
- Goal is to learn about and understand how NLP can be used to process text without programming all processes
  - However, some programming is required to
    - Call libraries
    - Process data
    - Customize NLP processes
  - Programming language is Python

INTRODUCTION TO NLTK

- NLTK provides:
  - Basic classes for representing data relevant to Natural Language Processing.
  - Standard interfaces for performing NLP tasks such as tokenization, tagging and parsing
  - Standard implementation of each task which can be combined to solve complex problems

SOME NLTK MODULES

- **corpora:** a package containing modules of example text
- **tokenize:** functions to separate text strings
- **probability**: for modeling frequency distributions and probabilistic systems
- **stem** – package of functions to stem words of text
- **wordnet** – interface to the WordNet lexical resource
- **chunk** – identify short non-nested phrases in text
- **etree**: for hierarchical structure over text
- **tag**: tagging each word with part-of-speech, sense, etc.
- **parse**: building trees over text
  - recursive descent, shift-reduce, probabilistic, etc.
- **cluster**: clustering algorithms
- **draw**: visualize NLP structures and processes
- **contrib:** various pieces of software from outside contributors

TUTORIALS FOR PYTHON AND NLTK

- Python

  many language constructs best documented in Python 2.x: https://docs.python.org/2/

  Python 3.x language reference, particularly for Unicode and string representations: https://docs.python.org/3/

- NLTK is a SourceForge project at: http://www.nltk.org

  documentation: http://www.nltk.org/documentation, including

  book: http://www.nltk.org/book/

  API: http://www.nltk.org/api/nltk.html