

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



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

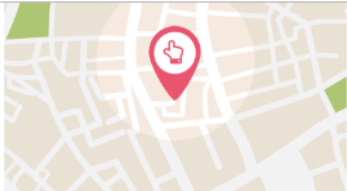
il y a plus d'un an

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 :

- Thumbs up icon
- Checklist icon
- First aid kit icon
- No smoking icon
- Maths icon (4+5=9)
- Broom icon
- Cooking icon

[Voir le profil complet](#)

South Shore Cotton Candy 4 Drawer...



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

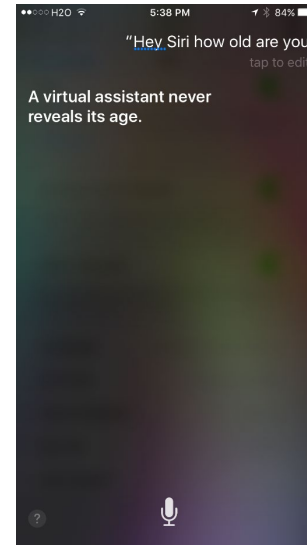
Weblog Analytics

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

- The odds are quite good, but the fact is irrelevant to the deletion discussion. [BMK \(talk\)](#) 00:59, 1 February 2016 (UTC)
 - Agreed, simply thought it was a factor for editors to consider. Either way, Majerus-Collins is clearly simply a local LARP-er with no national or international significance as Wikipedia, I believe, requires. [Dcpoliticaljunkie \(talk\)](#) 17:48, 1 February 2016 (UTC)
- Note: This debate has been included in the [list of Politicians-related deletion discussions](#). [Dcpoliticaljunkie \(talk\)](#) 22:24, 31 January 2016 (UTC)
- Note: This debate has been included in the [list of United States of America-related deletion discussions](#). [Dcpoliticaljunkie \(talk\)](#) 22:24, 31 January 2016 (UTC)
- **Delete** Fails [WP:GNG](#) and [WP:NPOLITICIAN](#). [Joseph2302 \(talk\)](#) 22:43, 31 January 2016 (UTC)
 - **Delete** Fails [WP:GNG](#) and [WP:NPOLITICIAN](#). Pretty cut and dry. [dimaspivak \(talk\)](#) 22:44, 31 January 2016 (UTC)
 - **Keep** [Annekeye \(talk\)](#)
 - **Note to closing admin:** [Annekeye \(talk · contribs\)](#) is the creator of the page that is the subject of this AfD.

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

CMV: I think it's a selfish motive to purposely try to have children.
1 year ago by [Caitybeck](#)

I want children one day. Part of me wants to have my own children but I can't justify birthing my own kids when it's such an extremely selfish motive. Sure, once you've had the kids it's selfish because of how much you have to give up for them. But the initial desire to birth them in the first place is selfish. I want my own children because I want to carry on my own genes. I want to have a little human that resembles me. As a woman, I want to experience the feeling of a baby inside of me. These reasons are the main reasons are people choose to birth their own kids. Here's a few reasons why I find it selfish.

- People who choose to birth children out there are selfish parents. Over 150 million orphans in the world. Don't mention foster children.
- There are people who are trying to birth their own when there are so many helpless children without a loving home.
- There are over 800 million people starving in the world. People are dying from hunger and you're trying to bring another mouth that needs feeding into the world. [+184 words]

[Omega037](#) 74Δ 7 points 1 year ago

Economically and socially, not having children when you have the means to raise them is a far more selfish act. First, let's focus the discussion on the US, since that is where I assume you live. After all, most of those 150 million orphans cannot be easily or legally adopted into the US, and those 800 million starving people is not because we lack food (we have a major surplus), it is because of the political and economic situation in their countries. Anyways, in the US, there were only 101,666 children legally up for adoption in 2012, and of them 52,039 children were adopted. [Source](#) That contrasts the ~3.9 million babies born each year. [Source](#) That contrasts the ~3 million babies born each year. [Source](#) In other words, the majority of children in the US who can be adopted are adopted, and even if all of them were, it would not cover even 25% of the births that happen each year. [Source](#) It is also important to note that even with all those births, the US doesn't meet its replacement rate. That means that our population would be declining if not for massive amount of immigrants we take in. [*168 words]

CMV: A Trump vs Hillary election would be the worst choice we've had in a presidential election (at least)

13 hours ago by [SayNoToStim](#)

So as of today, it looks like the presidential race is shaping up to be Donald Trump vs Hillary Clinton. If this is the worst combination of choices we've had in 75 years, if not more. The second worst, in my opinion, was Bush vs Gore. Both parties had some sort of merit behind them (at the time)

In my eyes Hillary is the most corrupt presidential candidate we've seen since Nixon, and even then, taking consideration, the general public wasn't aware of how corrupt [REDACTED]

I would start listing why Trump is a bad idea but I don't r

Both parties seem wildly disconnected to the general public, and quite frankly scare me when it comes to technology, and quite frankly scare me when it comes to

Hello, users of CMV! This is a footnote from your moderator. Please remember to [read through our rules](#). If you see a comment that is inappropriate, please [report it](#). Speaking of which, [downvotes don't change views](#). Please use our [popular topics wiki](#) first. Any questions or comments, please contact the moderators.

341 comments share

[top 200 comments](#) [show all 341](#)

sorted by: **best** ▼

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

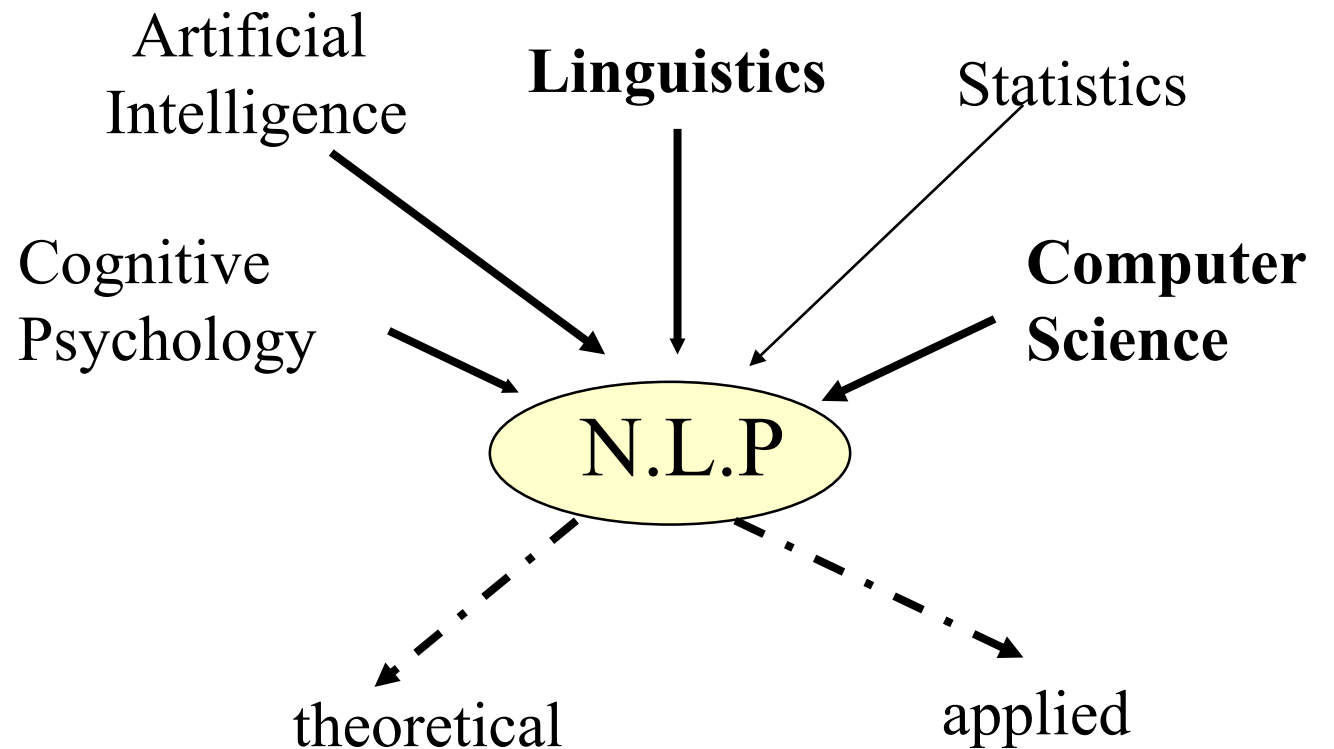
- Find at least 5 meanings of this sentence:
 - **I made her duck**
 - I cooked waterfowl for her benefit (to eat)
 - I cooked waterfowl belonging to her
 - I created the (plaster?) duck she owns
 - I caused her to quickly lower her head or body
 - I waved my magic wand and turned her into undifferentiated 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]



Natural Language Processing

Language Analysis* Language Generation

*Main emphasis in this course

CATEGORIES OF KNOWLEDGE

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

COURSE OUTLINE

- Lectures
 - Apr. 1 – away for a conference trip
 - Option 1: Ph.D. student Ayse Dalgali will deliver the lecture and hold the lab session
 - Option 2: lecture slides uploaded to the course web site and students will study them at their own pace
 - Ph.D. student Huan Zhuang will hold office hours in the two weeks before the assignment due date to help you on the homework

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	1 st : Feb. 17 2 nd : Mar. 31 3 rd : Apr. 21
NLP Application Investigations	Group report: May 3 Group poster: Apr. 28
Exam	Mar. 18

GROUP PROJECT

You will work in groups of four for this investigation (please discuss with the instructor if there is an issue on forming such a group). Your group will choose a general area or a specific topic to explore, and write a report about it (20 - 25 pages long including figures, tables, and references, single column, Times New Roman, 12 points). You will also use an openly accessible representative technique in the area or topic to analyze a dataset or perform an important task in the area.

GROUP PROJECT

A list of the areas and topics that you may be interested in exploring. You may suggest another area of interest to you.

- Machine Translation
- Information Retrieval/Search Engines
- Human Computer Interfaces / Dialogues, including “chatbots”
- Summarization
- Language Generation
- Question/Answering Systems
- Image to Text
- If your group is interested in working on one of my NLP projects, please let me know --- this requires more work with the goal of co-authoring at least a short conference paper with the instructor in the end; your group may need to meet with the instructor more to work on this project

GROUP PROJECT

Meet the instructor:

- **Feb. 11 – Feb. 15** (by Feb. 4 to schedule this meeting): In this meeting, you should be prepared to talk about one or two topics that your group is interested in exploring and any progress made in this project.
- **Mar. 18 – Mar. 22** (by Mar. 8 to schedule this meeting): In this meeting, you should be prepared to talk about the progress with respect to the topic that your group decides to explore.

Form the group here:

Morning Session: <https://docs.google.com/document/d/1NWNQzP5-7OufZtV5puW5lhClYceabpixAVADl7Rcy4/edit>

Afternoon Session:

<https://docs.google.com/document/d/1HKW5087LzRf3neTnVeohB9z4VGkTmh8Lx40CzW3wEU8/edit>

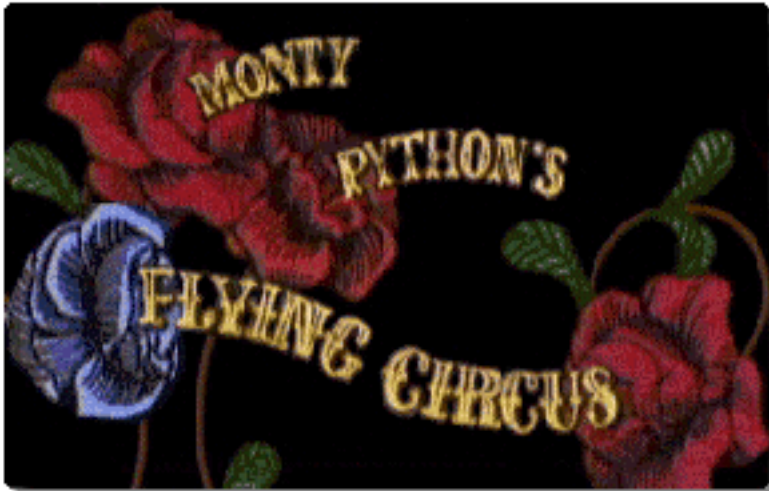


NATURAL LANGUAGE TOOLKIT (NLTK) AND PYTHON



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>

2019-01-13

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>