What is NLP
NLP In a Nutshell

takes as input text in human language and process it in a way that suggests an intelligent process was involved.
What is NLP

Text in human language

"Meaning"
"insights"
What is NLP

Text in human language → Text in another language
What is NLP

structured data in some format

Text in human language
How do we do NLP?

• 1950 -- ~1990s   ---> Write many rules
• 1990s -- ~2000s   ---> Corpus based statistics
• 2000s -- ~2014    ---> Supervised machine learning
• 2014 -- today     ---> "deep learning"
How do we do NLP?

- 1950 -- ~1990s  ---> Write many rules
- 1990s -- ~2000s ---> Corpus based statistics
- 2000s -- ~2014  ---> Supervised machine learning
- 2014 -- today   ---> "deep learning"

2021+  ---> write rules, aided by machine learning.
How do we do NLP?

- 1950 -- ~1990s  ---> Write many rules
- 1990s -- ~2000s ---> Corpus based statistics
- 2000s -- ~2014  ---> Supervised machine learning
- 2014 -- today   ---> "deep learning"

2021+  --->  write rules, aided by machine learning.
How do we do NLP?

- 1950 -- ~1990s  ---> Write many rules
- 1990s -- ~2000s ---> Corpus based statistics
- 2000s -- ~2014  ---> Supervised machine learning
- 2014 -- today    ---> "deep learning"  89-687

2021+ --> write rules, aided by machine learning.
This Course

• Natural Language Processing (NLP):
  • Algorithms to **process**, **analyze** and **understand** texts in natural language.
  • Understanding **Structure**
  • Understanding **Meaning**
  • Before you solve a problem, you need to understand it.
  • What is natural language? what are the building blocks?
What

• How to think about working with text data.
• Understand the challenges.
• Understand the linguistic concepts.
• Understand the ML concepts.
• Understand the algorithms.
• Understand where algorithms fail, and how to assess the quality of algorithms.
• Understand what is missing.
What Not

- We assume you know basic statistics / ML.
- We will only briefly mention neural nets / deep learning. (see 89-687, today 17:00-19:00)
- In many cases we will show the lower-level algorithms, without much details on how their outputs are used---unless you ask us to.
Discussion / Communication


- Discussion about assignments will be done through a Piazza forum. Don't email us with questions, everything should go through the piazza: https://piazza.com/biu.ac.il/fall2019/89680/

- Communications with me regarding course, also through piazza (private messages), not via email.
External Materials

• Ido presented some books. They are great.

• But no books covers the entire class.

• We will post links to related courses/materials on the course website.

• Each class will be accompanied by a list of related papers. You are encouraged to read them.
Levels of Language Analysis

- Phonetics (how sounds are made: lips, spit, throat)
- Phonology (how sounds can combine)
- Morphology (how words are built)
- Syntax (how words are combined)
- Semantics (the meaning of words/phrases)
- Pragmatics (the true meaning of words/phrases)
- Discourse (structure/meaning across sentences)

as Reut mentioned:
as Reut mentioned:

Levels of Language Analysis

- Phonetics (how sounds are made: lips, spit, throat)
- Phonology (how sounds can combine)
- Morphology (how words are built)
- Syntax (how words are combined)
- Semantics (the meaning of words/phrases)
- Pragmatics (the true meaning of words/phrases)
- Discourse (structure/meaning across sentences)
as Reut mentioned:

Levels of Language Analysis

- Phonetics (how sounds are made: lips, spit, throat)
- Phonology (how sounds can combine)
- Morphology (how words are built)
- Syntax (how words are combined)
- Semantics (the meaning of words/phrases)
- Pragmatics (the true meaning of words/phrases)
- Discourse (structure/meaning across sentences)
Levels of Language Analysis

- Phonetics (how sounds are made: lips, spit, throat)
- Phonology (how sounds can combine)
- Morphology (how words are built)
- Syntax (how words are combined)
- Semantics (the meaning of words/phrases)
- Pragmatics (the true meaning of words/phrases)
- Discourse (structure/meaning across sentences)
Levels of Language Analysis

- Phonetics (how sounds are made: lips, spit, throat)
- Phonology (how sounds can combine)
- Morphology (how words are built)
- Syntax (how words are combined)
- Semantics (the meaning of words/phrases)
- Pragmatics (the true meaning of words/phrases)
- Discourse (structure/meaning across sentences)

As Reut mentioned:
Levels of Language Analysis

- **Phonetics** (how sounds are made: lips, spit, throat)
- **Phonology** (how sounds can combine)
- **Morphology** (how words are built)
- **Syntax** (how words are combined)
- **Semantics** (the meaning of words/phrases)
- **Pragmatics** (the true meaning of words/phrases)
- **Discourse** (structure/meaning across sentences)

---

as Reut mentioned:
Levels of Language Analysis

- Phonetics (how sounds are made: lips, spit, throat)
- Phonology (how sounds can combine)
- Morphology (how words are built)
- Syntax (how words are combined)
- Semantics (the meaning of words/phrases)
- Pragmatics (the true meaning of words/phrases)
- Discourse (structure/meaning across sentences)

As Reut mentioned:
Levels of Language Analysis

- Phonetics (how sounds are made: lips, spit, throat)
- Phonology (how sounds can combine)
- Morphology (how words are built)
- Syntax (how words are combined)
- Semantics (the meaning of words/phrases)
- Pragmatics (the true meaning of words/phrases)
- Discourse (structure/meaning across sentences)

There is ambiguity on all levels. (can you think of examples?)

as Reut mentioned:
the basic units of text processing
What are we processing?

Document Collection / Cluster

Document

Section

Paragraph

Sentence

Word
What are we processing?

Document Collection / Cluster

- Document
  - Section
    - Paragraph
    - Sentence
      - Word

by date
by topic
by web domain
...

NLP
What are we processing?

- Document Collection / Cluster
- Document
- Section
- Subsection
- Paragraph
- Sentence
- Word
What are we processing?

- Document Collection / Cluster
- Document
- Section
- Subsection
- Paragraph
- Sentence
- Subsection
- Phrase...
What are we processing?

Document Collection / Cluster
Document
Section
Paragraph
Sentence
Phrase...
Word
Characters, morphemes
What are we processing?

- Document Collection / Cluster
  - Document
    - Section
    - Paragraph
    - Sentence
    - Word
What is a document?
What is a document?

British lawmakers clash ahead of Brexit vote 05:57

Story highlights

Britons will vote Thursday on whether their country should remain in the European Union

Henry Wismayer: Urge to leave has caught the wind of native disaffection

Editor's Note: Henry Wismayer is a writer and commentator based in London. You can follow him @henrywismayer. The views expressed are his own.

(CNN) — For any external observer curious to understand the issues shaping the debate over Britain's membership of the European Union, last Wednesday's edition of The Sun, the country's most popular daily newspaper, produced a revealing metaphor.

A day after the Rupert Murdoch-owned tabloid declared its support for the "Leave" campaign, its front-page went full-hysteria. A "biblical swarm" of supermoths has invaded Britain from mainland Europe, it exclaimed without a trace of irony, "and is set to annihilate our cabbages."

The prominence given to these ghastly foreign moths, accompanied on Page 2 with a helpful diagram depicting a battalion of Lepidoptera hovering over the English Channel, tells you all you need to know about Britain's EU referendum debate, where fear and frustration have come to overshadow the biggest political decision in a generation.
What is a document?

British lawmakers clash ahead of Brexit vote

Story highlights

Britons will vote Thursday on whether their country should remain in the European Union

Henry Wismayer: Urged to leave has caught the wind of native disaffection

Editor's Note: Henry Wismayer is a writer and commentator based in London. You can follow him @henrywismayer. The views expressed are his own.

(CNN) — For any external observer curious to understand the issues shaping the debate over Britain’s membership of the European Union, last Wednesday’s edition of The Sun, the country’s most popular daily newspaper, produced a revealing metaphor.

A day after the Rupert Murdoch-owned tabloid declared its support for the "Leave" campaign, its front-page went full-hysteria. A "biblical swarm" of super-moths has invaded Britain from mainland Europe, it exclaimed without a trace of irony, "and is set to annihilate our cabbages."

The prominence given to these ghastly foreign moths, accompanied on Page 2 with a helpful diagram depicting a battalion of Lepidoptera hovering over the English Channel, tells you all you need to know about Britain’s EU referendum debate, where fear and frustration have come to overshadow the biggest political decision in a generation.
What is a document?

and maybe each paragraph is a document?

British lawmakers clash ahead of Brexit vote 05:57

Story highlights

Britons will vote Thursday on whether their country should remain in the European Union

Henry Wismayer: Urge to leave has caught the wind of native disaffection

Editor's Note: Henry Wismayer is a writer and commentator based in London. You can follow him @henrywismayer. The views expressed are his own.

(CNN) — For any external observer curious to understand the issues shaping the debate over Britain's membership of the European Union, last Wednesday's edition of The Sun, the country's most popular daily newspaper, produced a revealing metaphor.

A day after the Rupert Murdoch-owned tabloid declared its support for the "Leave" campaign, its front-page went full-hysteria. A "biblical swarm" of super-moths has invaded Britain from mainland Europe, it exclaimed without a trace of irony, "and is set to annihilate our cabbages."

The prominence given to these ghastly foreign moths, accompanied on Page 2 with a helpful diagram depicting a battalion of Lepidoptera hovering over the English Channel, tells you all you need to know about Britain's EU referendum debate, where fear and frustration have come to overshadow the biggest political decision in a generation.
What are we processing?

- Document Collection / Cluster
- Document
- Section
- Paragraph
- Sentence
- Word
What is a sentence?

sentences = text.split(".")?
What is a sentence?

(we'll get back to this)
What are we processing?

Document Collection / Cluster

Document

Section

Paragraph

Sentence

Word
What is a word?
What is a word?

sequence of characters

basic unit of meaning?

white-space tokenized?
What is a word?

doesn't
John's
unlucky
“as if you are among those whom we were not able to civilize (=cause to become civilized)”

uygar: civilized
_laş: become
_tır: cause somebody to do something
_ama: not able
_dık: past participle
_lar: plural
_imız: 1st person plural possessive (our)
_dan: among (ablative case)
_miş: past
_sınız: 2nd person plural (you)
_casına: as if (forms an adverb from a verb)
What is a word?

שלום
What is a word?
What is a word?
What is a word?
What is a word?

they told  count!  his book

(ספר של هو)

סיפרו

(He is the book)
What is a word?

טיפוח

they told   count!   his book

(טיפוח של חוה)

he told it   he₁ cut his₂ hair
What is a word?
What is a word?

שלום
What is a word?

בצלם
בצל של הם
בצל של הם
בצלם
What is a word?

Words aren't just defined by blanks.

Problem 1: Compounding

“ice cream”, “website”, “web site”, “New York-based”

Problem 2: Other writing systems have no blanks

Chinese:

我 开始 写 小说  =  我  开始     写  小说

Problem 3: Clitics

English:

“doesn’t”, “I’m”

Italian:

“dirglielo” = dir + gli(e) + lo
tell + him + it
whitespace is not enough
What is a word?
What is a word?

- ice cream
- web site
- New York
- New York-Based
What is a word?

gave up

made sense

took a picture

took apart

took the toy apart
Case Study: Sentence Boundary Detection.
What is a sentence?

sentences = text.split("."),
A **graphic adventure game** is a form of **adventure game**. They are distinct from **text adventures**. Whereas a player must actively observe using commands such as "look" in a text-based adventure, graphic adventures revolutionized gameplay by making use of visual human perception. Eventually, the **text parser** interface associated with older **interactive fiction** games was phased out in favor of a **point-and-click interface**, i.e., a game where the player interacts with the game environment and objects using an on-screen cursor. In many of these games, the mouse pointer is **context sensitive** in that it applies different actions to different objects.
What is a sentence?

sentences = text.split(". ")?
What is a sentence?

Georg "Mr. George" Svendsen (19 March 1894 – 1966) was a Norwegian journalist and crime novelist. He was born in Eidanger, and started his journalistic career in Bratsberg-Demokraten before moving on to Demokraten where he was a subeditor. In 1921 he was hired in Fremtiden and replaced in Demokraten by Evald O. Solbakken. In 1931 he was hired in Arbeiderbladet. Under the pen name "Mr. George" he became known for his humorous articles in the newspaper. At his death he was also called "the last of the three great criminal and police reporters in Oslo", together with Fridtjof Knutsen and Axel Kielland. He was also known for practising journalism as a trade in itself, and not as a part of a political career. He retired in 1964, and died in 1966.

He released the criminal novels Mannen med ljåen (1942), Ridderne av øksen (1945) and Den hvite streken (1946), and translated the book S.S. Murder by Quentin Patrick as Mord ombord in 1945. He released several historical books: Rørleggernes Fagforenings historie gjennem 50 år (1934), Telefonmontørenes Forening Oslo, gjennem 50 år (1939), Norsk nærings- og nydelsesmiddelarbeiderforbund: 25-års beretning (1948), De tause vitner: av rettskjemiker Ch. Bruffs memoarer (1949, with Fridtjof Knudsen) and Elektriske montørers fagforening gjennom 50 år (1949).
Sentence Boundary Detection

• How would you solve this?

(discuss)
Sentence Boundary Detection

• Worth reading:

Sentence Boundary Detection and the Problem with the U.S.

Dan Gillick
Computer Science Division
University of California, Berkeley
dgillick@cs.berkeley.edu
What is a sentence?

It was a bad time. (it is always a bad time.)
What is a sentence?

In a quiet voice, he said "this will not work. I am quitting", and the he left the room.
How often do you find yourself on the internet looking at the same boring pages? You know there is something out there but you don’t know where to look.

Trust me, how bad could it be?

Click above for a new random website. Links will open in a new window.

**Disclaimer:** The websites you are forwarded to may contain naughty stuff. There is no known "porn" but sites change hands often and sometimes they change ethical direction. If you see something that looks like porn or is otherwise unacceptable let us know and we'll decide if it should be removed or if you're just a crank who likes to complain.

Don't leave the house with your fly open. Make sure you've got up-to-date virus protection.

**Colophon & Warning & Contact**

Want to add your own links? After reading the warning type the url in the box below.

If you see something that shouldn't be here let us know

random website dot com 2001-2011

"I'd like to quit thinking of the present, like right now, as some minor insignificant preamble to something else."
How often do you find yourself on the internet looking at the same boring pages? You know there is something out there but you don't know where to look.

Trust me, how bad could it be?

Click above for a new random website.
Links will open in a new window.

Disclaimer: The websites you are forwarded to may contain naughty stuff. There is no known "porn" but sites change hands often and sometimes they change ethical direction. If you see something that looks like porn or is otherwise unacceptable let us know and we'll decide if it should be removed or if you're just a crank who likes to complain.

Don't leave the house with your fly open. Make sure you've got up-to-date virus protection.

Colophon & Warning & Contact

Want to add your own links? After reading the warning type the url in the box below.

If you see something that shouldn't be here let us know.

random website dot com 2001-2011

"I'd like to quit thinking of the present, like right now, as some minor insignificant preamble to something else."
How often do you find yourself on the internet looking at the same boring pages? You know there is something out there but you don't know where to look.

Trust me, how bad could it be?

Click above for a new random website. Links will open in a new window.

Disclaimer: The websites you are forwarded to may contain naughty stuff. There is no known "porn" but sites change hands often and sometimes they change ethical direction. If you see something that looks like porn or is otherwise unacceptable let us know and we'll decide if it should be removed or if you're just a crank who likes to complain.

Don't leave the house with your fly open. Make sure you've got up-to-date virus protection.

Colophon & Warning & Contact

Want to add your own links? After reading the warning type the url in the box below.

If you see something that shouldn't be here let us know.

random website dot com 2001-2011

"I'd like to quit thinking of the present, like right now, as some minor insignificant preamble to something else."
How often do you find yourself on the internet looking at the same boring pages? You know there is something out there but you don't know where to look.

Trust me, how bad could it be?

Click above for a new random website Links will open in a new window

**Disclaimer:** The websites you are forwarded to may contain naughty stuff. There is no known "porn" but sites change hands often and sometimes they change ethical direction. If you see something that looks like porn or is otherwise unacceptable let us know and we'll decide if it should be removed or if you're just a crank who likes to complain.

Don't leave the house with your fly open. Make sure you've got up-to-date virus protection.

Colophon & Warning & Contact

Want to add your own links? After reading the warning type the url in the box below

If you see something that shouldn't be here let us know

random website dot com 2001-2011

"I'd like to quit thinking of the present, like right now, as some minor insignificant preamble to something else."
Sentence Boundary Detection

• Perhaps the most basic task.

• Non-trivial.

• Need to consider corpus, features, annotation procedure, biases.

• Need to think of your use-cases, choose your sentence definition, methods, and trade-offs.