

CS 2731

Introduction to Natural Language Processing

Session 3: Machine learning intro, NLP tasks and applications

Michael Miller Yoder

September 3, 2025

Overview: Machine learning intro, NLP tasks and applications

- Coding activity: preprocessing Airbnb reviews
- Intro to machine learning
 - Definitions
 - Models and algorithms
 - Data: training, development, test
- NLP applications
- NLP “core tasks”

Course logistics

- I re-released [Homework 0](#) on getting set up on the CRCD JupyterHub with a custom class conda environment
 - Is **due this Fri Sep 5 at 11:59pm**
- I released the [project idea form](#). Is **due next Thu Sep 11**
 - You will be able to submit any project ideas that you're interested in: from the [example list](#) or any you have on your own
 - It's fine to incorporate your own research, there just needs to be an NLP component
 - You can submit multiple project ideas
- You will later choose from an anonymized list of project ideas on Project Match Day, Sep 17

Coding activity: Preprocessing Airbnb listings

Starting a Jupyter Notebook on the CRCD JupyterHub

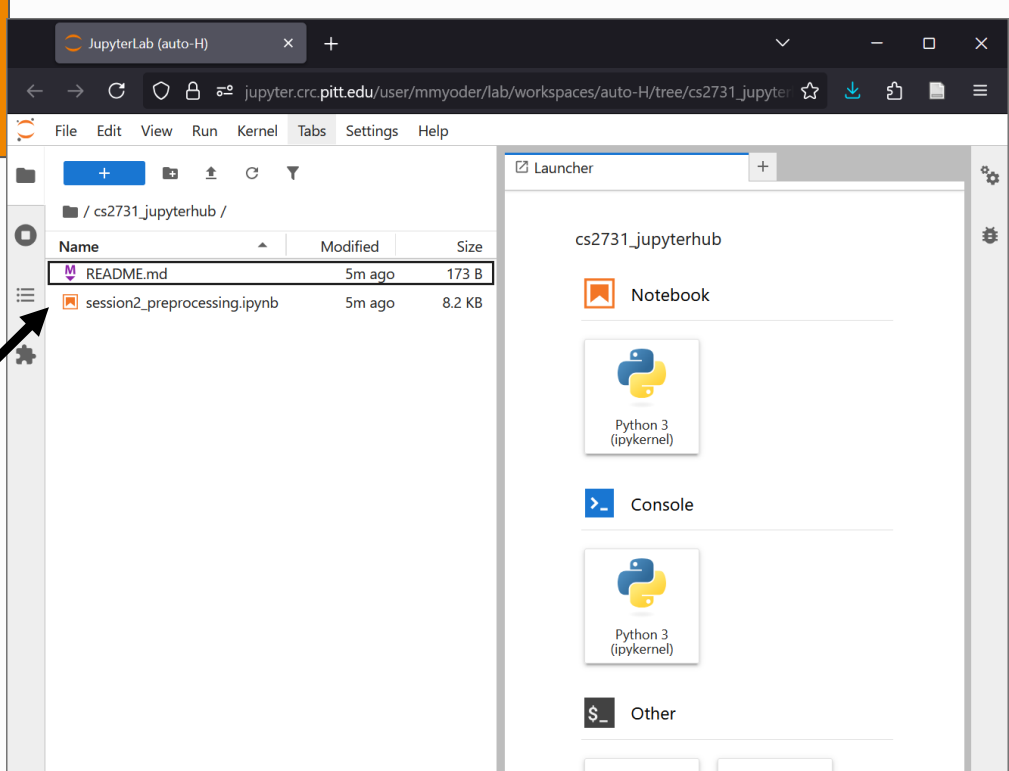
1. Go to this [nbgitpuller link](#) (also available on course website)
2. Partition: **TEACH – 6 CPUs – 45 GB**
We might use the GPU options later on in the course
3. Under **Select Virtual Environment**, select **Provide custom path**
4. Custom Environment Path:
`/ix/cs2731_2025f/class_env`
5. Click **Start**
6. Wait for the server to start up

The screenshot shows the JupyterHub 'Server Options' page. The browser address bar shows 'jupyter.crc.pitt.edu/hub/spawn/mmyoder'. The page has a dark theme. The 'JupyterHub Session Configuration' section contains the following fields:

- Select Partition:** A dropdown menu with 'TEACH - 6 CPUs - 45GB' selected. An arrow points to this dropdown from the left.
- Select Virtual Environment:** A dropdown menu with 'Provide custom path' selected. An arrow points to this dropdown from the left.
- Custom Environment Path:** A text input field containing '/ix/cs2731_2025f/class_env'. An arrow points to this field from the left.
- Select Modules to Load:** A text area containing 'Amber 2024' and 'Cuda 12.3'. Below it is a note: 'Hold Ctrl/Cmd to select multiple modules'. An arrow points to this section from the left.
- Account:** A text input field containing 'your class account'.
- Start:** A large orange button at the bottom right.

Open Jupyter notebook

1. This should pull a folder (cs2731_jupyterhub) into your JupyterLab
2. Double-click **session2_preprocessing.ipynb** on the left panel to open the notebook



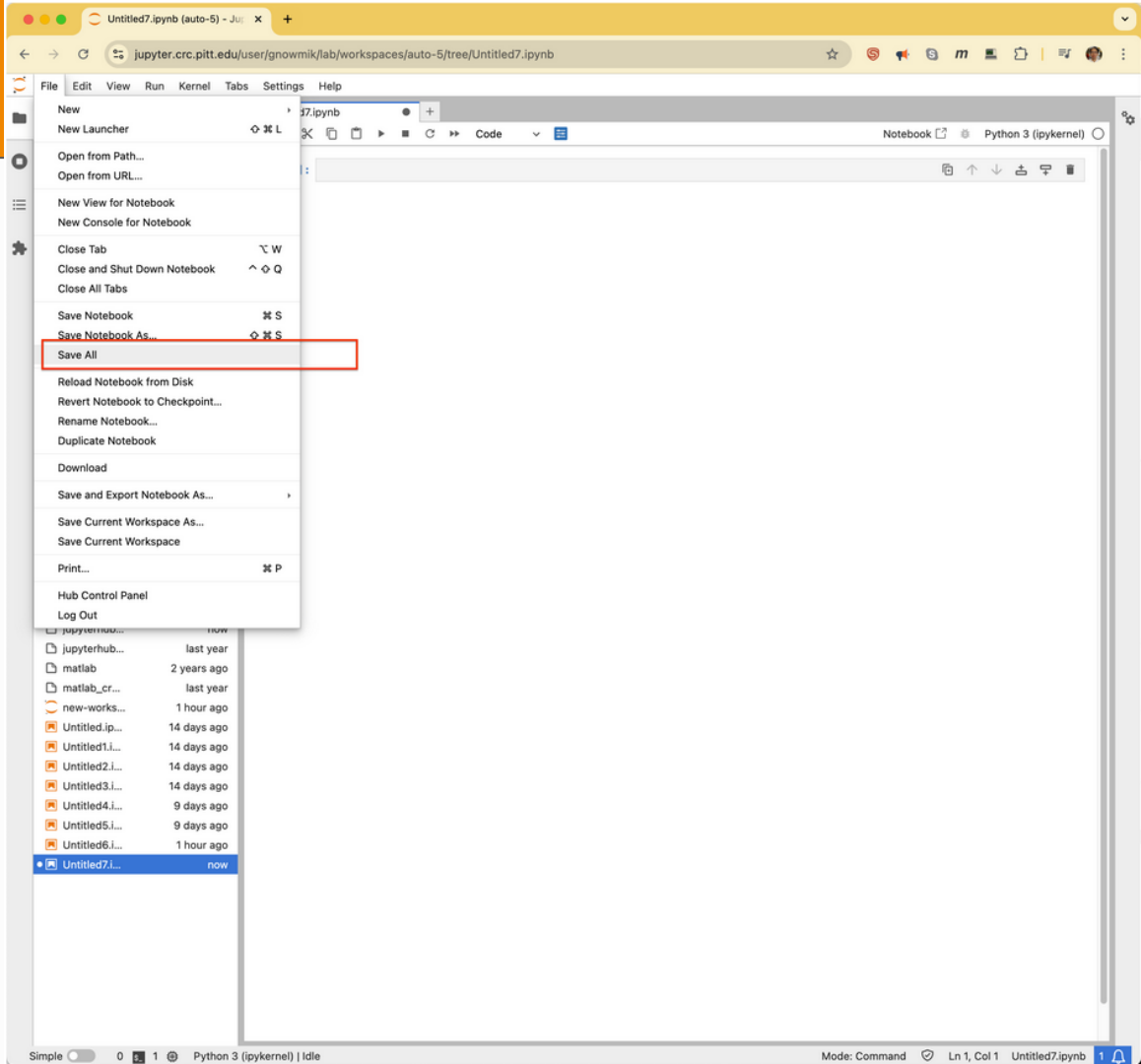
Jupyter Notebook basics

- Each block is called a “cell”
 - Has input and possibly output
 - Input can be Python code, Markdown or shell commands (after !)
- Modes
 - Command mode
 - Move, select, manipulate cells
 - Get into command mode by clicking anywhere outside of a cell
 - Edit mode
 - Edit content of a particular cell
- Running cells
 - Click “Run” button or do Ctrl+Enter (on Windows or Linux, Cmd+Enter on Mac) to run code or render Markdown
 - Any result will be shown in the output of the cell

Implementation

- Remove undesired text with regular expressions
- Lowercase
- Remove stopwords
- Tokenize with the NLTK package
- Stem the tokens with NLTK

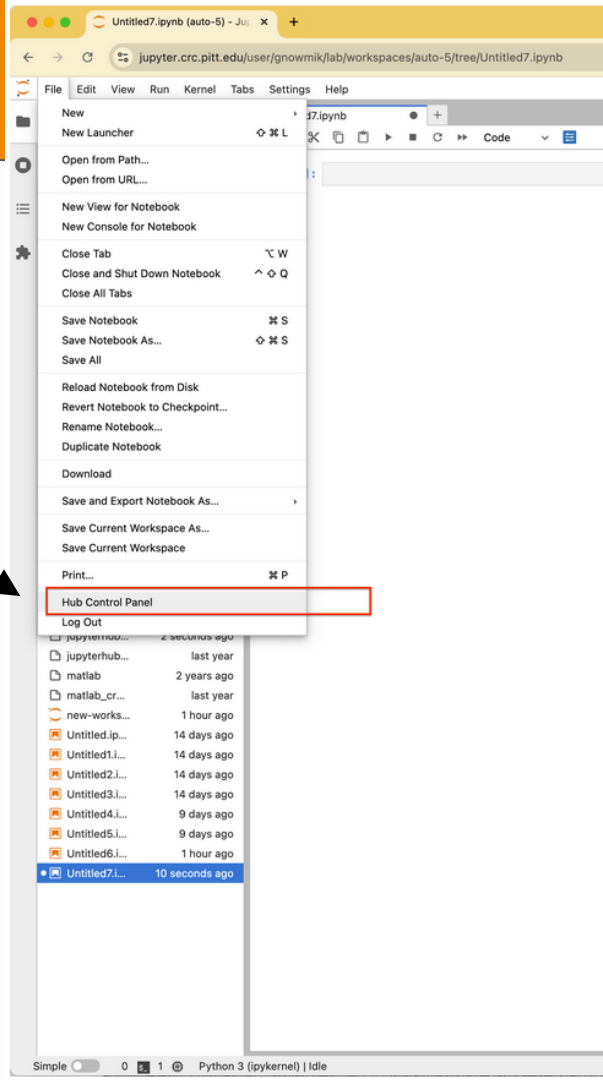
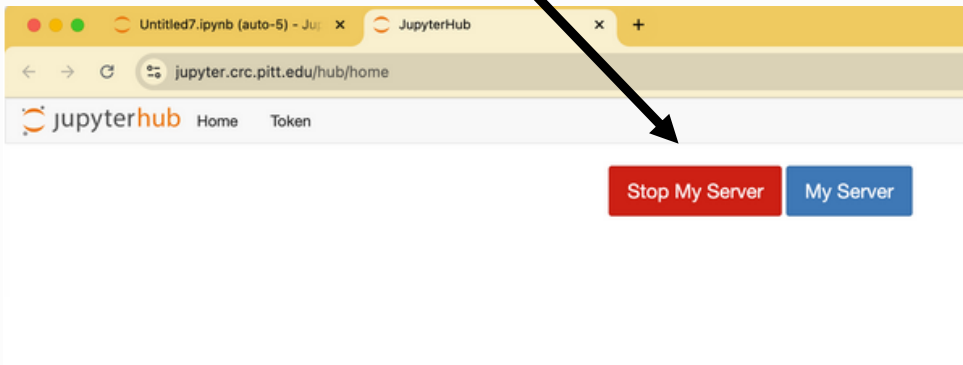
Saving your work



Ending your session

Be sure to save your work
before ending the session

1. Select **File > Hub Control Panel**
2. Click **Stop My Server**



• Intro to (supervised) machine learning

What is machine learning?

- A system that learns a function (maps from an input to an output) from examples/data
- Can predict things and perform tasks **without** explicit instructions
- Learns patterns from data with statistical algorithms

Machine learning models

- Transform an input to an output with a “model”: a simplified mathematical/statistical version of reality
- Models have parameters **learned from patterns in data**
 - Usually encode how variables relate to each other

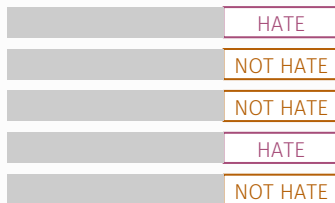


Queens Museum of Art, photo Chris Devers. <https://www.flickr.com/photos/cdevers/8063002401>

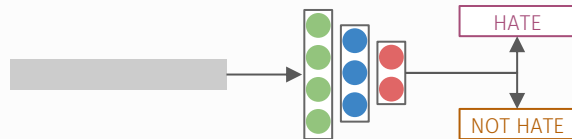
Supervised machine learning process



Data
(input text, X)



Annotate
labels (Y)



Train a model to
predict labels (Y)
from input text (X)

Training and test sets (and phases)

Training set

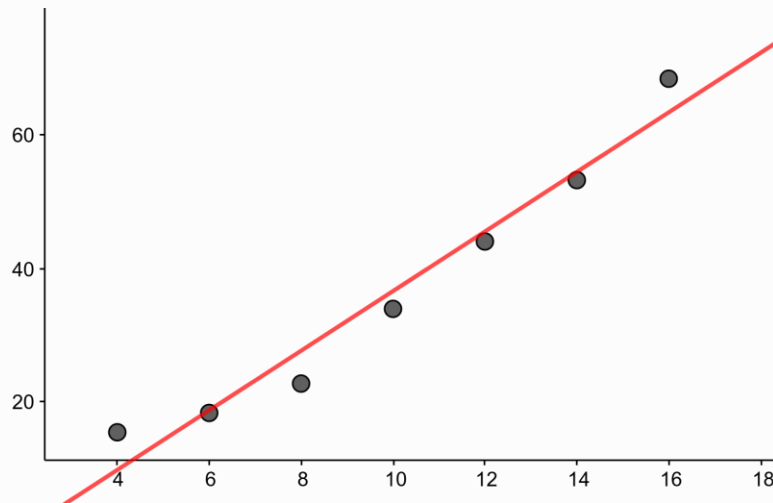
Development set

Test set

- Train parameters of the model on training set (training phase)
 - Sees examples of input and (assumed correct) output that it will mimic
- Development set to run tests of the model and choose hyperparameters
- Test time
 - Freeze parameters of the model
 - Predict input from an unseen set
 - Evaluate on correct answers and see how well the model performs
- **Don't look at the test set too much when developing/choosing models**

What can you do with machine learning models?

- Prediction: predict an output from an unseen input
 - That fits the pattern learned by looking at input it has seen before
- Interpretation
 - Examine the learned model weights to characterize the relationship between variables



$$y = 4x - 10$$

NLP applications

Core tasks and applications of NLP

APPLICATIONS

machine translation

chatbots

information retrieval

summarization

question answering

NLP applications: email classification

The screenshot displays a Gmail interface with a sidebar on the left and a main inbox area. The sidebar includes a 'COMPOSE' button, a list of folders (Inbox (7), Starred, Drafts, Sent Mail), and a 'Search people...' section with a list of contacts: Jenny Kang, Peter H, Jonathan Pelleg, Brett C, Max Stein, Jen Hart, and Eric Lowery. The main inbox area is divided into four tabs: Primary, Social (3 new), Promotions (2 new), and Updates (2 new). The 'Social' tab is selected, showing a list of emails from Google+, YouTube, and Emily Million. Each email entry includes a checkbox, a star icon, the sender, and the subject line. The emails are categorized by their source (Google+, YouTube) and their content (tagged in photos, video uploads, knitting club, new pup, shared albums, new apt, added on Google+).

COMPOSE

Inbox (7)

Starred

Drafts

Sent Mail

Search people...

- Jenny Kang
- Peter H
- Jonathan Pelleg
- Brett C
- Max Stein
- Jen Hart
- Eric Lowery

Primary

Social 3 new
Google+, YouTube, Emi...

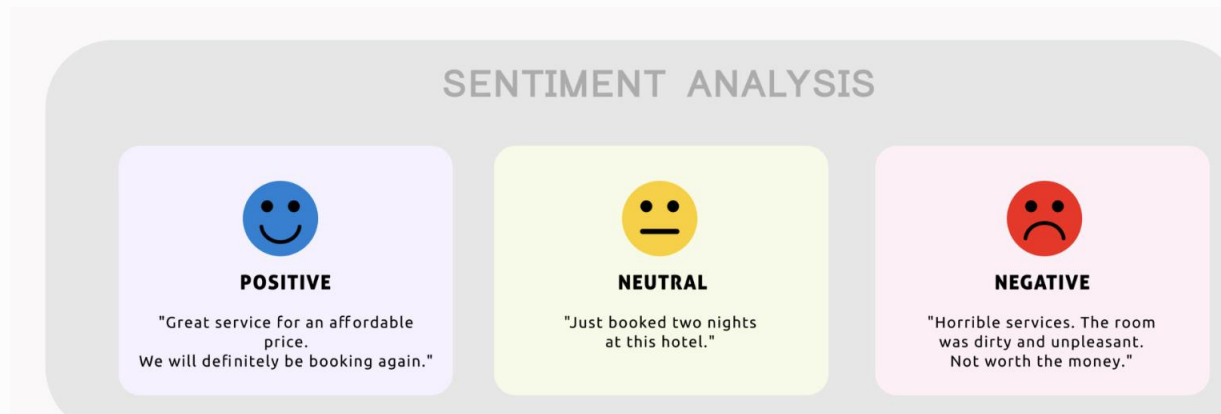
Promotions 2 new
Google Offers, Zagat

Updates 2 new
Shoehop, Blitz Air

Category	Sender	Subject
Google+	Google+	You were tagged in 3 photos on Google+ - Google+ You were tagged in three pl
YouTube	LauraBlack	LauraBlack just uploaded a video. - Jess, have you seen the video LauraBlack u
Emily Million (Google+)	[Knitting Club]	[Knitting Club] Are we knitting tonight? - [Knitting Club] Are we knitting tonight?
Sean Smith (Google+)	Sean Smith	Photos of the new pup - Sean Smith shared an album with you. View album be thoi
Google+	Kate Baynham	Kate Baynham shared a post with you - Follow and share with Kate by adding her
Google+	Danielle Hoodhood	Danielle Hoodhood added you on Google+ - Follow and share with Danielle by
YouTube	YouTube	Just for You From YouTube: Daily Update - Jun 19, 2013 - Check out the latest
Google+	Google+	You were tagged in 3 photos on Google+ - Google+ You were tagged in three phot
Hilary Jacobs (Google+)	Hilary Jacobs	Check out photos of my new apt - Hilary Jacobs shared an album with you. View
Google+	Kate Baynham	Kate Baynham added you on Google+ - Follow and share with Kate by adding her

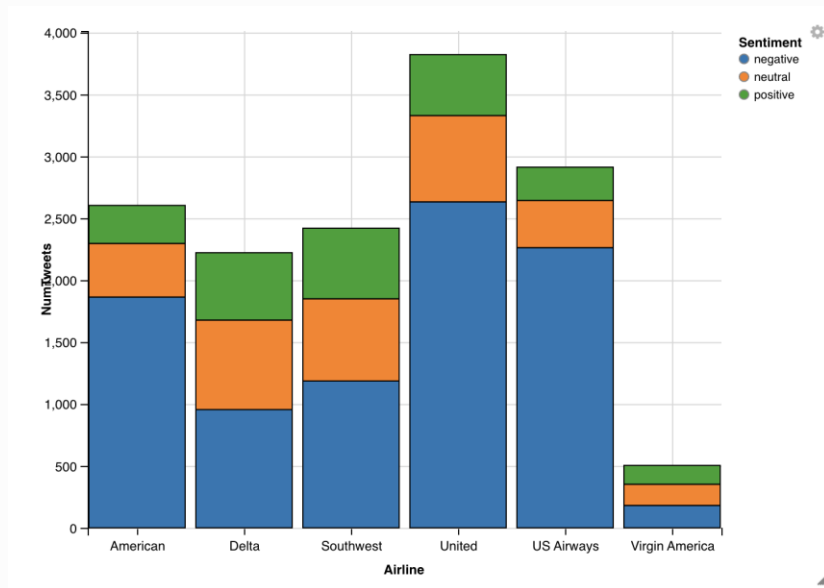
- Spam / Not spam
- Priority Level
- Category (primary / social / promotions / updates)

NLP applications: sentiment analysis



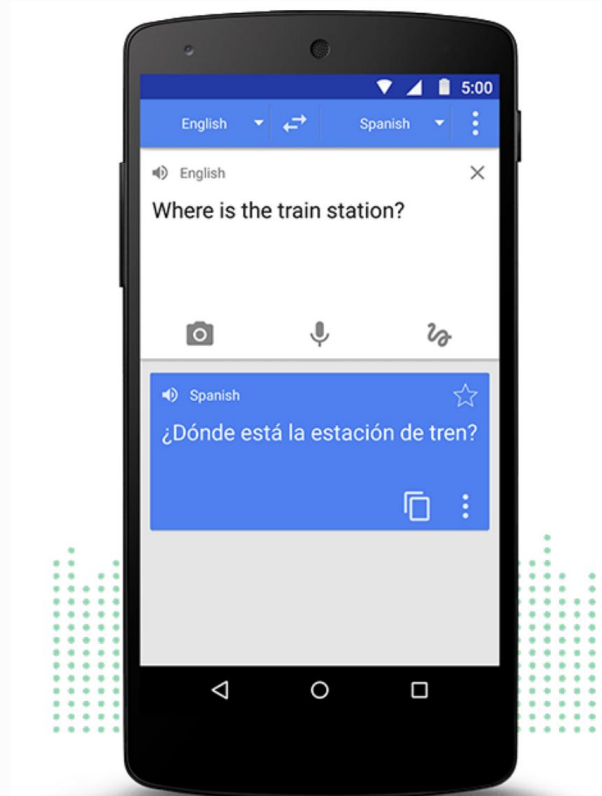
Hotel review sentiment

NLP applications: sentiment analysis



US Airline review sentiment

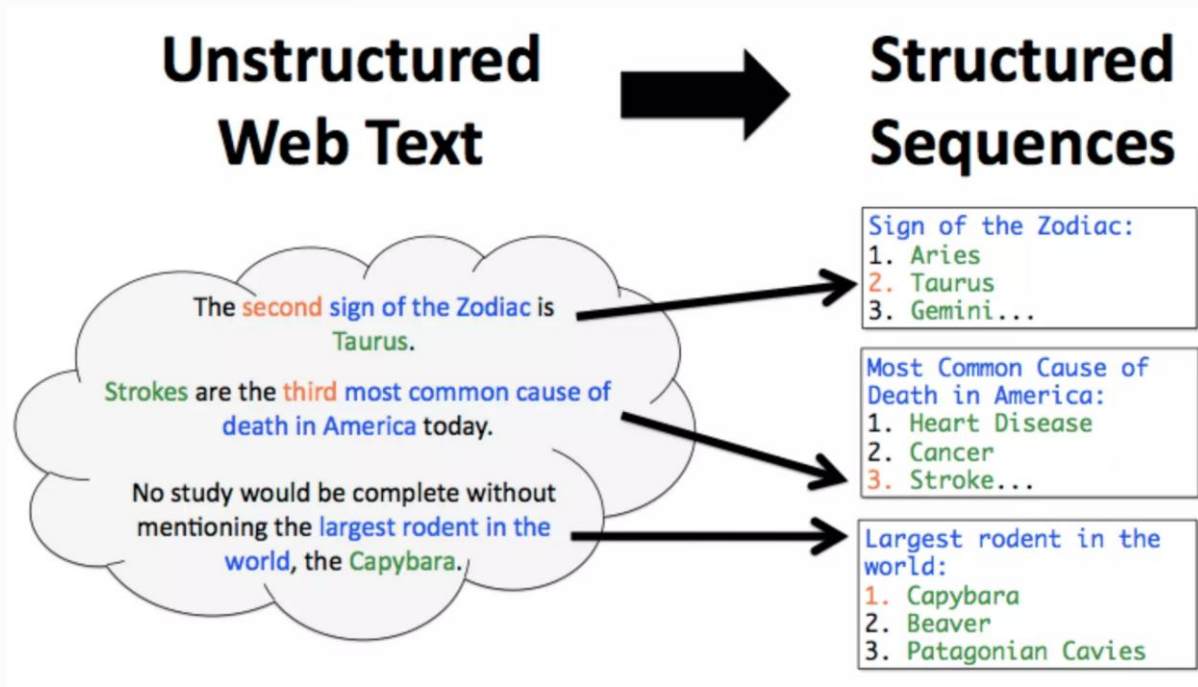
NLP applications: machine translation



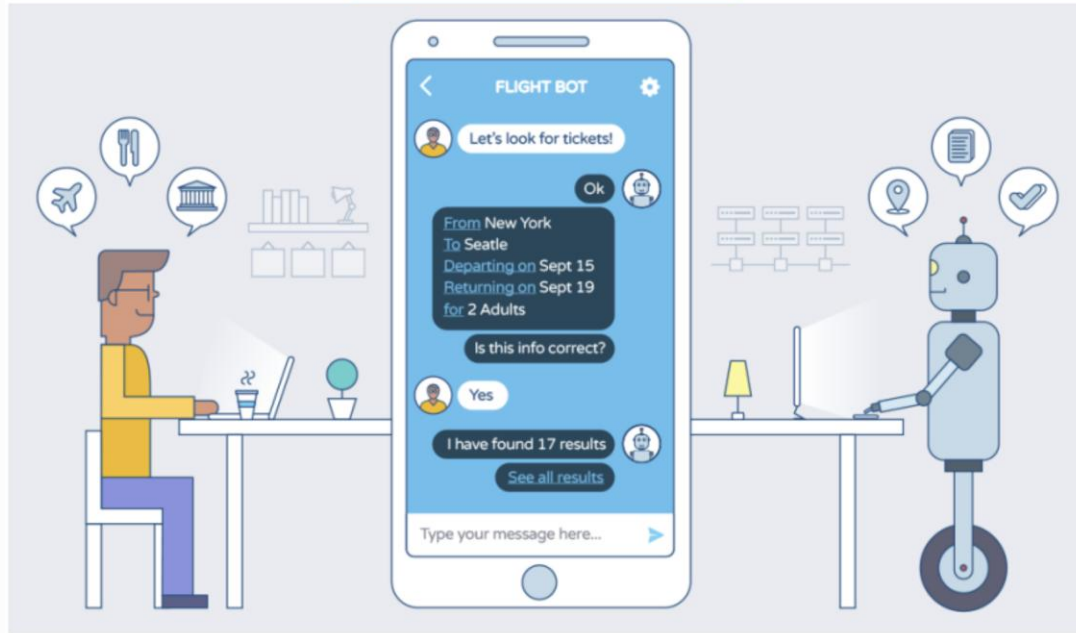
NLP applications: summarization



NLP applications: information extraction




NLP applications: dialogue systems/chatbots



NLP applications: question answering



 amazon alexa

"Alexa, who was President when Barack Obama was nine?"

"Alexa, how's my commute?"

"Alexa, what's the weather?"

"Alexa, did the 49ers win?"



• NLP core tasks

Core tasks and applications of NLP

CORE TASKS

text classification

language modeling

sequence labeling



APPLICATIONS

machine translation

chatbots

information retrieval

summarization

question answering

Text classification

- Input: a span of text
- Output: a label from a set of discrete options
- *Example:* sentiment analysis
 - *Text* -> {positive, neutral, negative}

Language modeling

- Input: a span of text, or no text at all
- Output: the next word
- *Example:* text generation for chatbots (ChatGPT)
 - *context text -> next word*

Sequence labeling

- Input: a span of text
- Output: a sequence of labels, one for each word (token)
- *Example: part-of-speech tagging*
 - *The book was brilliant -> DET NOUN VERB ADJ*