CS 2731
Introduction to Natural Language Processing

Session 3: Machine learning intro, NLP tasks and applications

Michael Miller Yoder

September 3, 2025

University of Pittsburgh | School of Computing and Information

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”
- (If time allows) Coding activity: clickbait classification

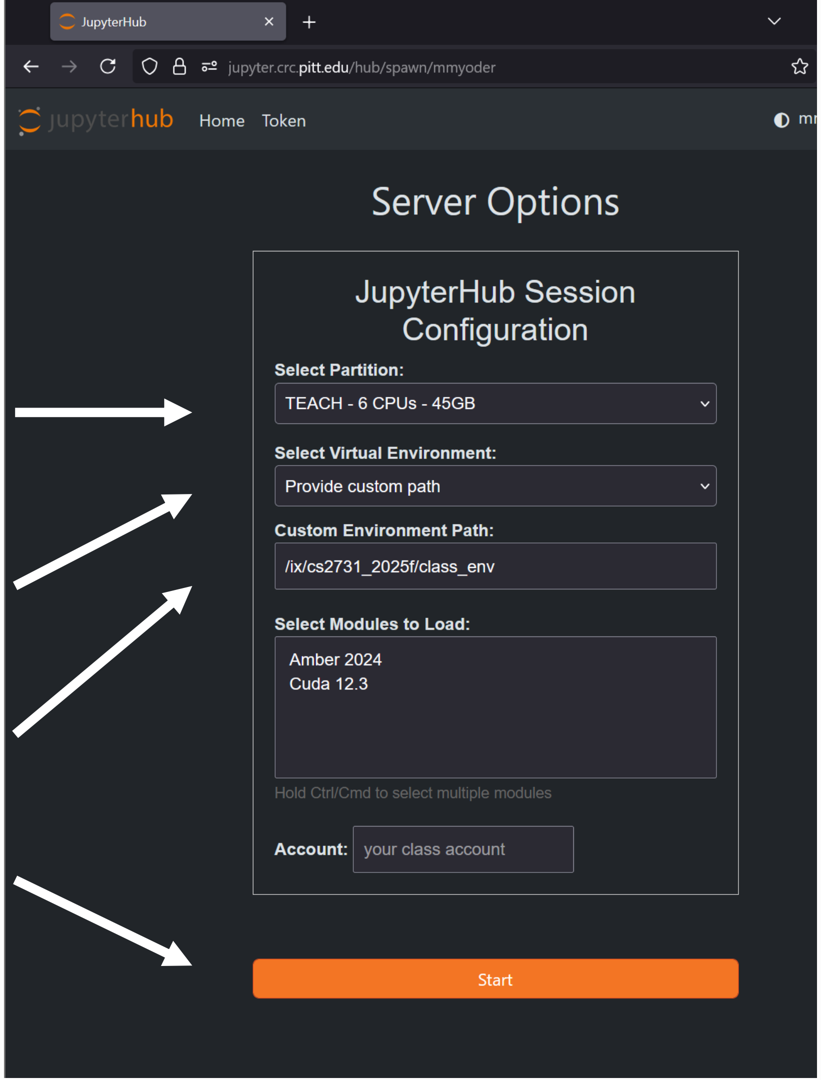
Course logistics

- 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 will release the project idea form tomorrow, Thu Sep 4. 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



Server Options

JupyterHub Session Configuration

Select Partition: TEACH - 6 CPUs - 45GB

Select Virtual Environment: Provide custom path

Custom Environment Path: /ix/cs2731_2025f/class_env

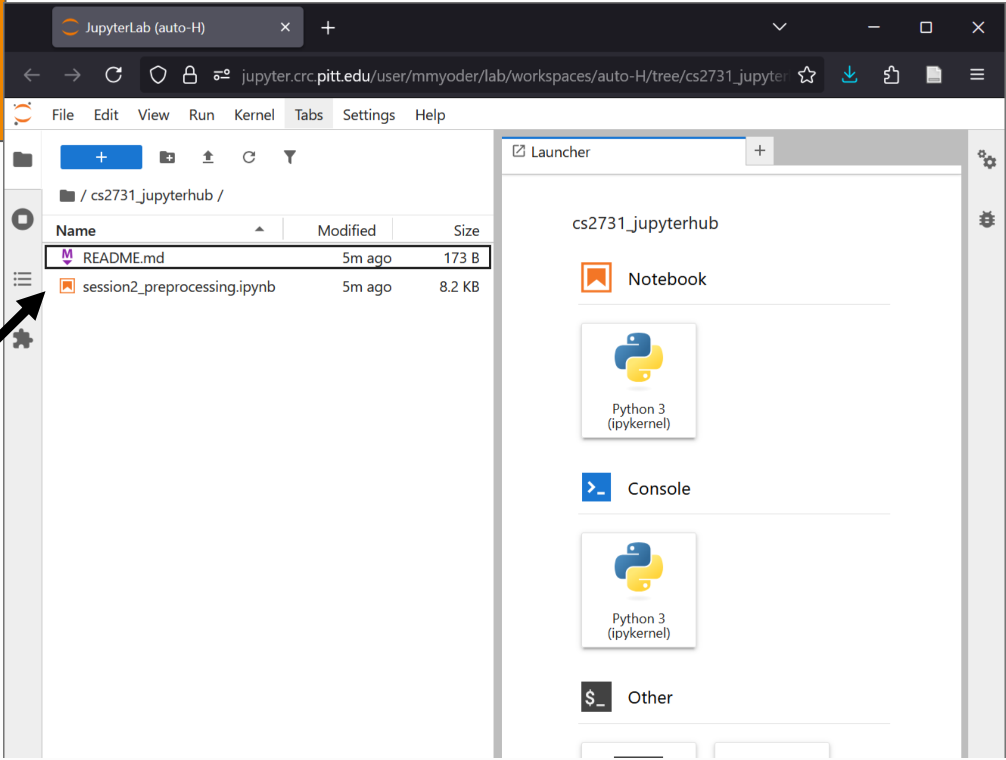
Select Modules to Load:
Amber 2024
Cuda 12.3

Hold Ctrl/Cmd to select multiple modules

Account: your class account

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



Name	Modified	Size
README.md	5m ago	173 B
session2_preprocessing.ipynb	5m ago	8.2 KB

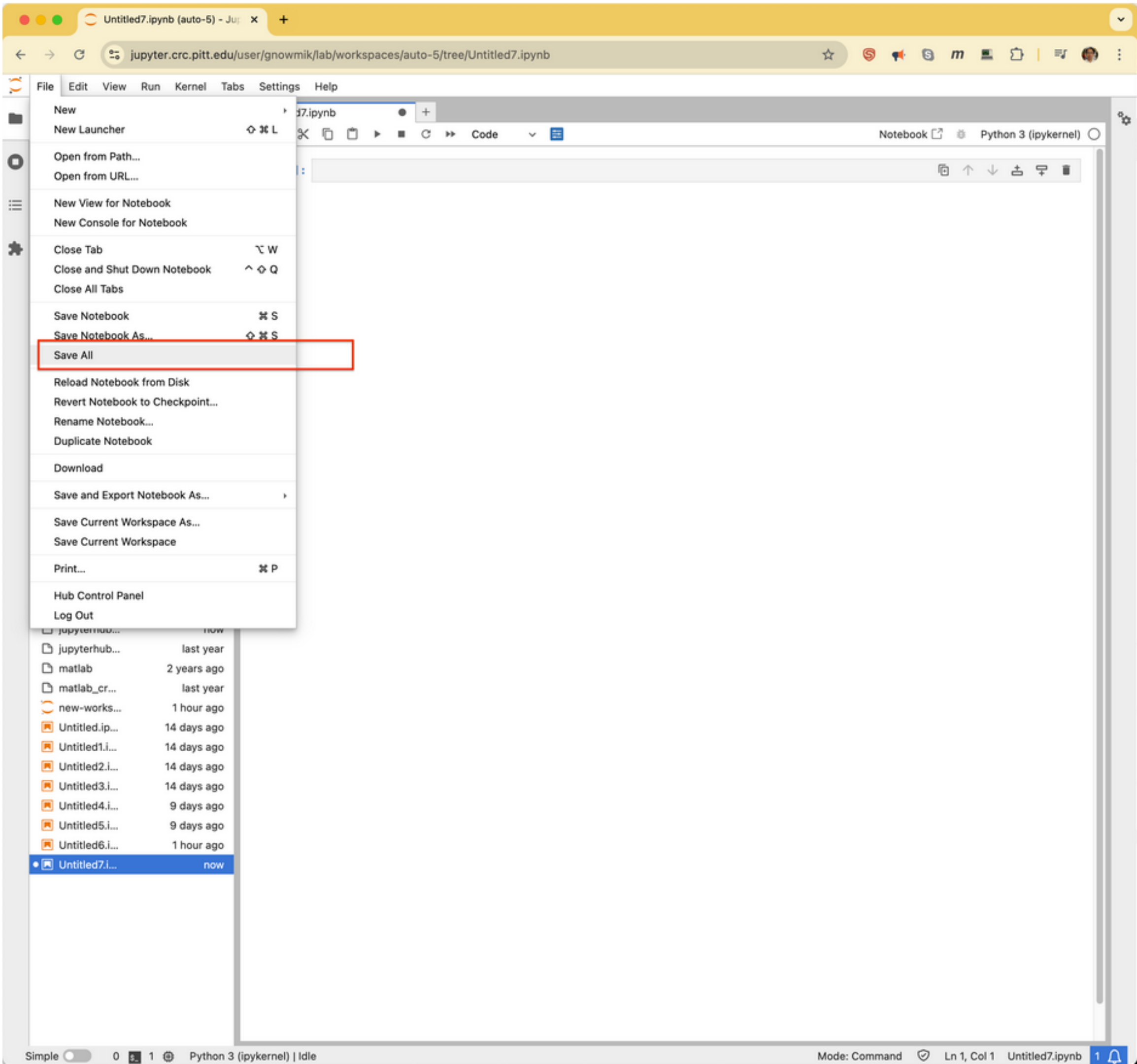
Jupyter Notebook basics

- 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

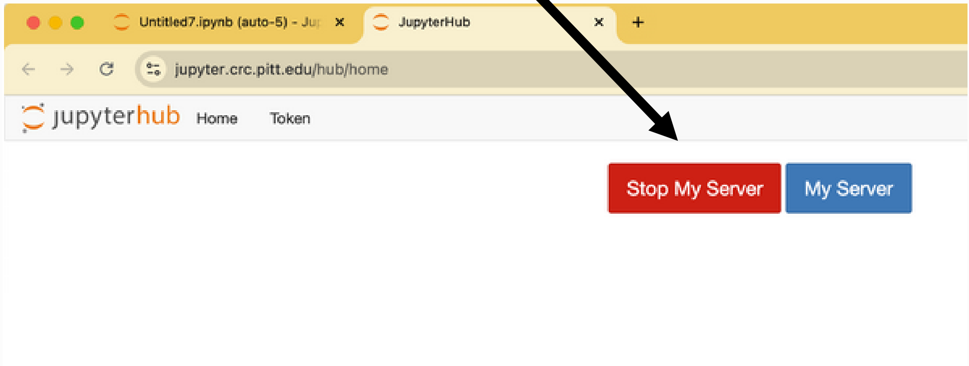
Saving your work

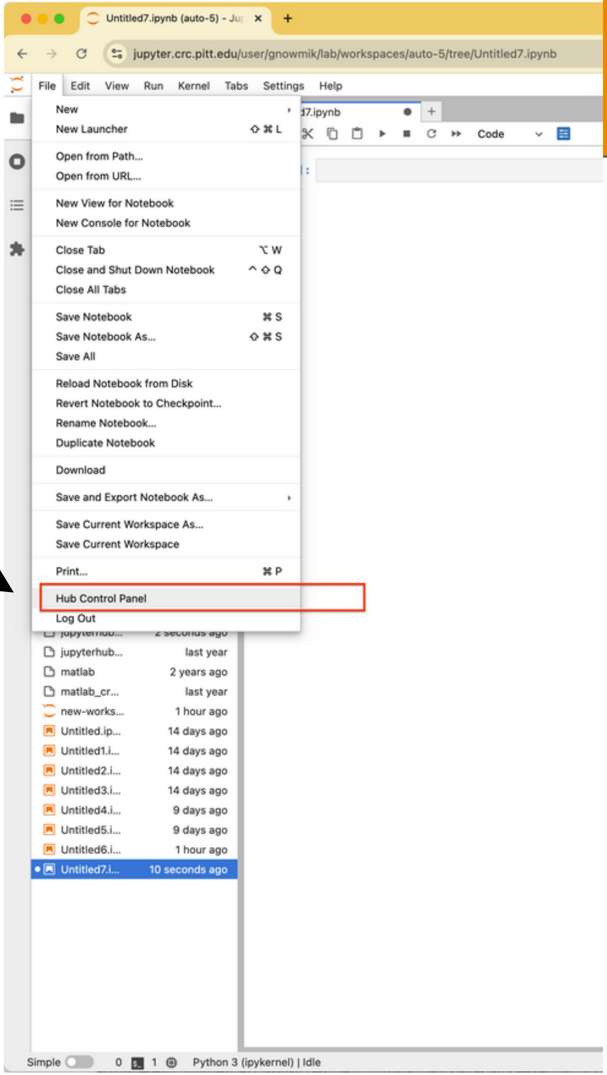


Ending your session

Be sure to save your work before ending the session

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




Untitled7.ipynb (auto-5) - Jup
jupyter.crc.pitt.edu/user/gnowmik/lab/workspaces/auto-5/tree/Untitled7.ipynb
File
Edit
View
Run
Kernel
Tabs
Settings
Help
New
New Launcher ⇧ ⌘ L
Open from Path...
Open from URL...
New View for Notebook
New Console for Notebook
Close Tab ⌥ W
Close and Shut Down Notebook ^ ⇧ Q
Close All Tabs
Save Notebook ⌘ S
Save Notebook As... ⇧ ⌘ S
Save All
Reload Notebook from Disk
Revert Notebook to Checkpoint...
Rename Notebook...
Duplicate Notebook
Download
Save and Export Notebook As...
Save Current Workspace As...
Save Current Workspace
Print... ⌘ P
Hub Control Panel
Log Out
d7.ipynb
Code
jupyterhub... last year
matlab 2 years ago
matlab_cr... last year
new-works... 1 hour ago
Untitled.ip... 14 days ago
Untitled1.i... 14 days ago
Untitled2.i... 14 days ago
Untitled3.i... 14 days ago
Untitled4.i... 9 days ago
Untitled5.i... 9 days ago
Untitled6.i... 1 hour ago
Untitled7.i... 10 seconds ago
Simple
0
1
Python 3 (ipykernel) | Idle

Intro to (supervised) machine learning

What is machine learning?

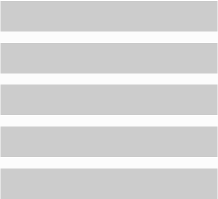
- A system that learns a function (maps from an input to an output) from examples/data
- Can predict things and perform tasks **without** explicit instructions
- Learns patterns from data with statistical algorithms

Machine learning models

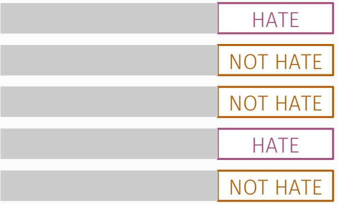
- Transform an input to an output with a “model”: a simplified mathematical/statistical version of reality
- Models have parameters **learned from patterns in data**
  - Usually encode how variables relate to each other

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

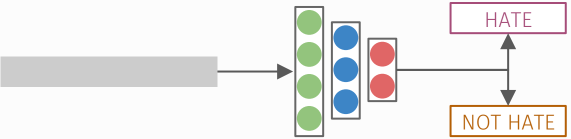
Supervised machine learning process



Data
(input text, *X*)

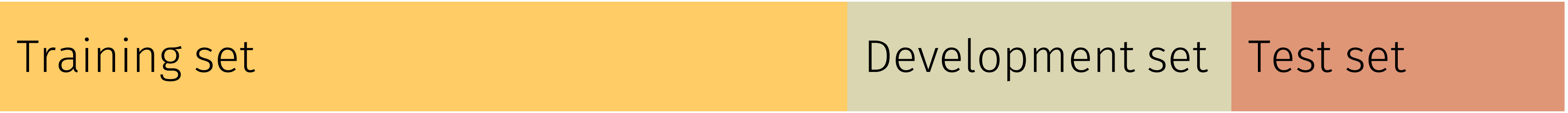


Annotate
labels (*Y*)



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

Training and test sets (and phases)



Training set

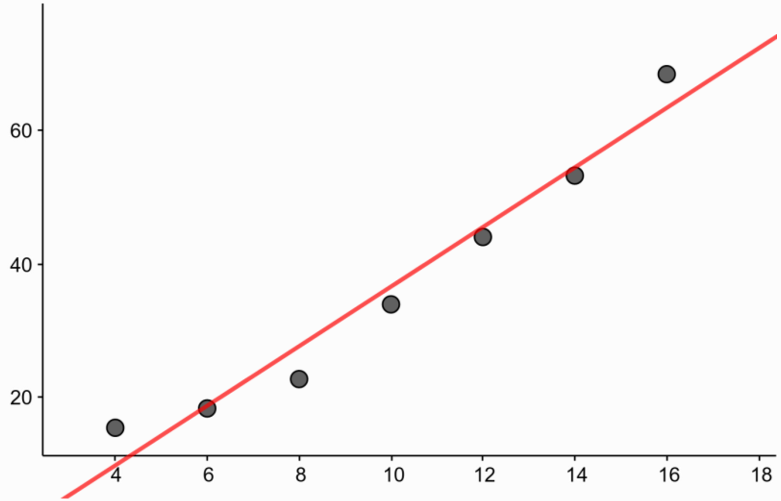
Development set

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

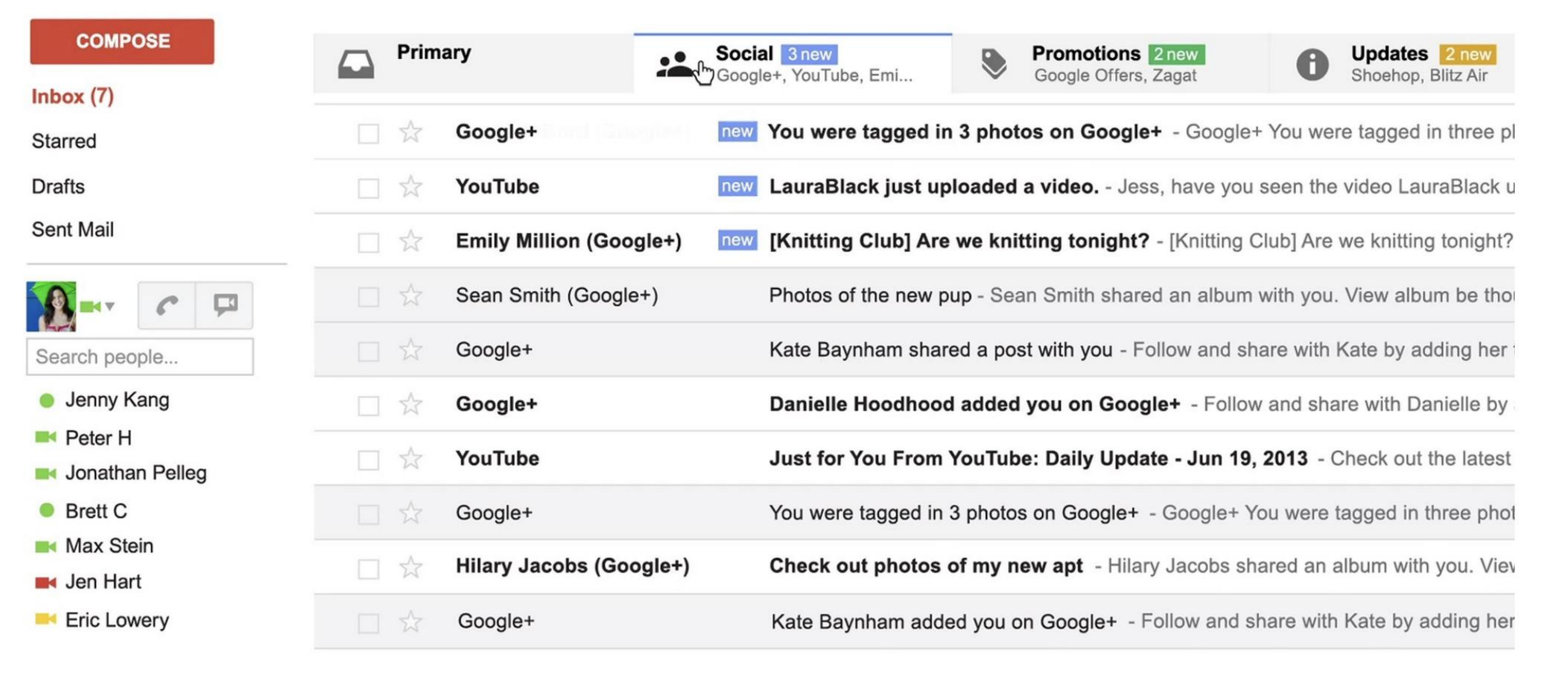
chatbots

information retrieval

summarization

question answering

NLP applications: email classification

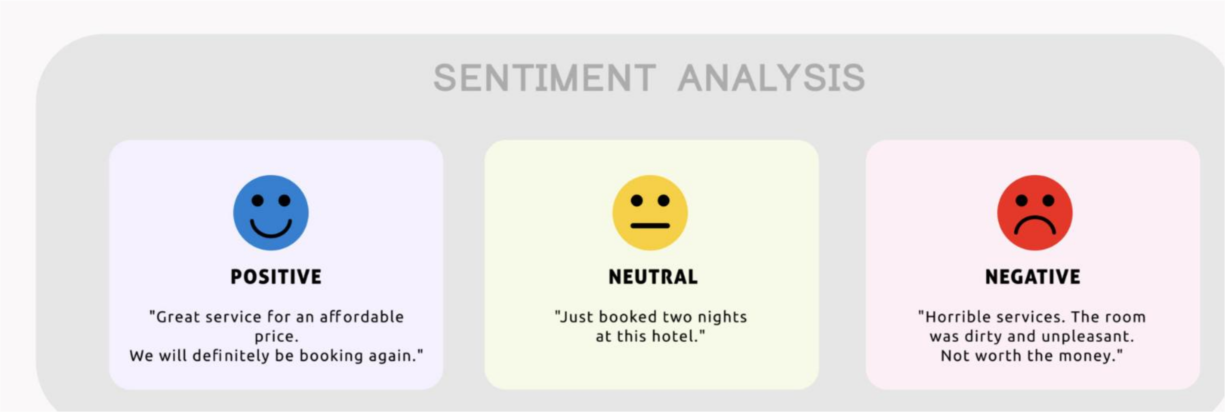


Google+	new	**You were tagged in 3 photos on Google+** - Google+ You were tagged in three pl
YouTube	new	**LauraBlack just uploaded a video.** - Jess, have you seen the video LauraBlack u
Emily Million (Google+)	new	**[Knitting Club] Are we knitting tonight?** - [Knitting Club] Are we knitting tonight?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

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

Slides Credit: Kevin Gimpel

NLP applications: sentiment analysis



SENTIMENT ANALYSIS

POSITIVE

"Great service for an affordable price.
We will definitely be booking again."

NEUTRAL

"Just booked two nights at this hotel."

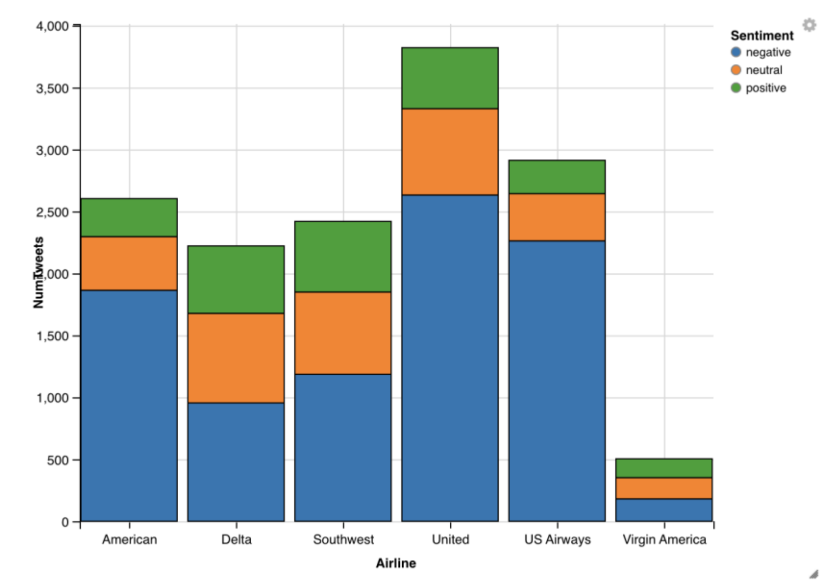
NEGATIVE

"Horrible services. The room was dirty and unpleasant.
Not worth the money."

Hotel review sentiment

Slides Credit: Kevin Gimpel

NLP applications: sentiment analysis

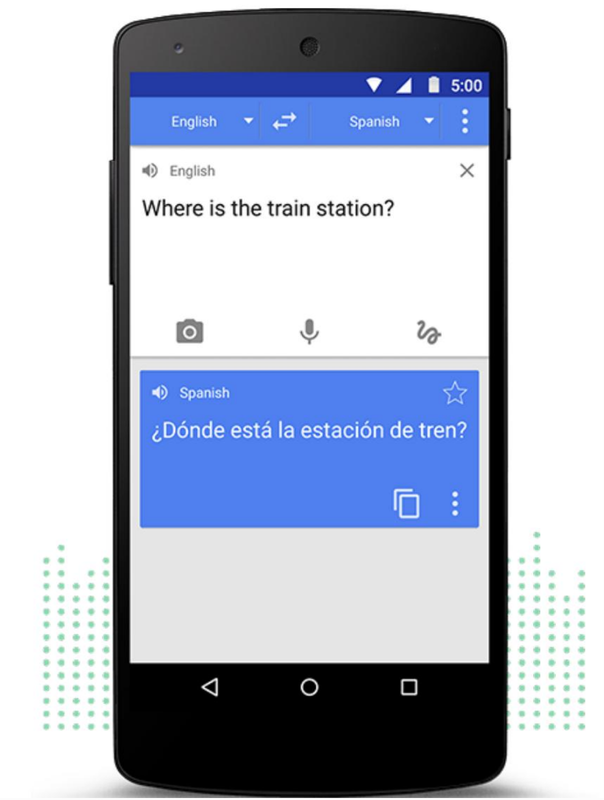


US Airline review sentiment

Slides Credit: Kevin Gimpel

https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment

NLP applications: machine translation



English

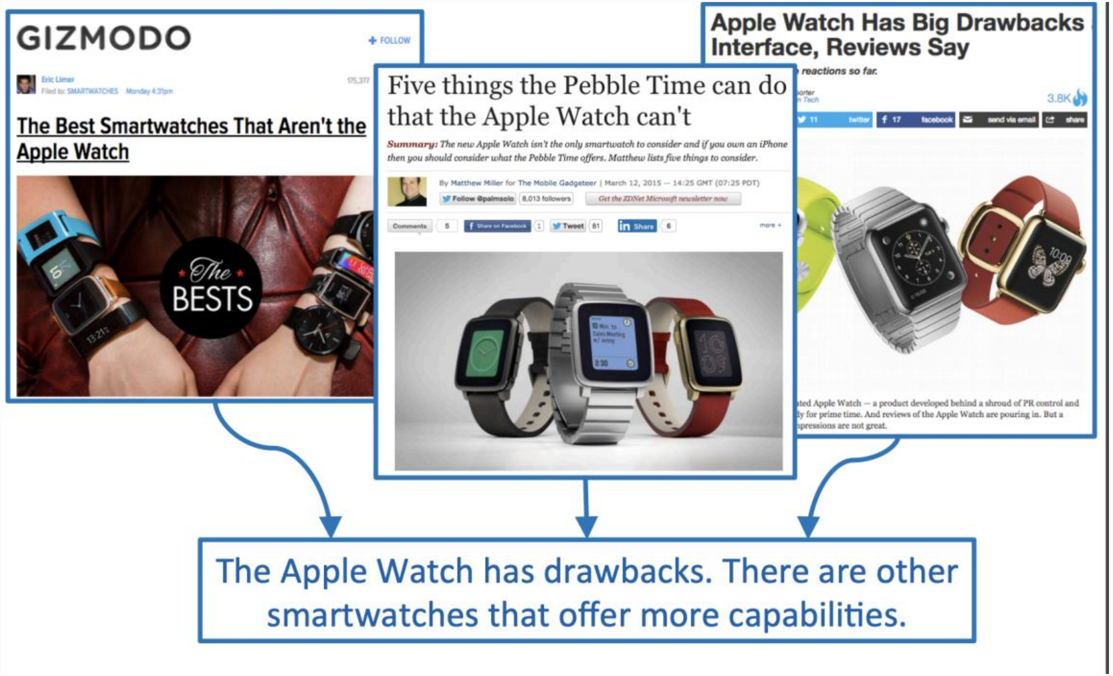
Where is the train station?

Spanish

¿Dónde está la estación de tren?

Slides Credit: Kevin Gimpel

NLP applications: summarization



The Best Smartwatches That Aren't the Apple Watch

Five things the Pebble Time can do that the Apple Watch can't

Summary: The new Apple Watch isn't the only smartwatch to consider and if you own an iPhone then you should consider what the Pebble Time offers. Matthew lists five things to consider.

By Matthew Miller for The Mobile Gadgeteer | March 12, 2015 -- 14:25 GMT (07:25 PDT)

Apple Watch Has Big Drawbacks Interface, Reviews Say

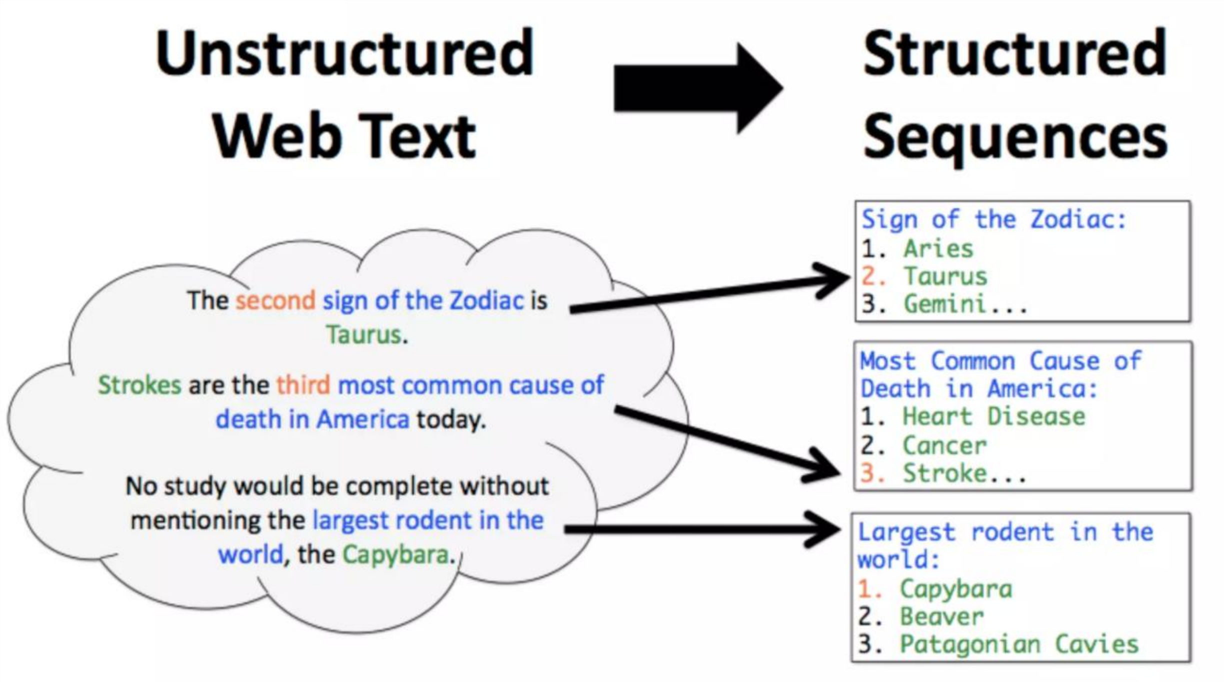
reactions so far.

ated Apple Watch -- a product developed behind a shroud of PR control and
ly for prime time. And reviews of the Apple Watch are pouring in. But a
pressions are not great.

The Apple Watch has drawbacks. There are other smartwatches that offer more capabilities.

Slides Credit: Kevin Gimpel

NLP applications: information extraction



Unstructured Web Text

Structured Sequences

The second sign of the Zodiac is Taurus.

Strokes are the third most common cause of death in America today.

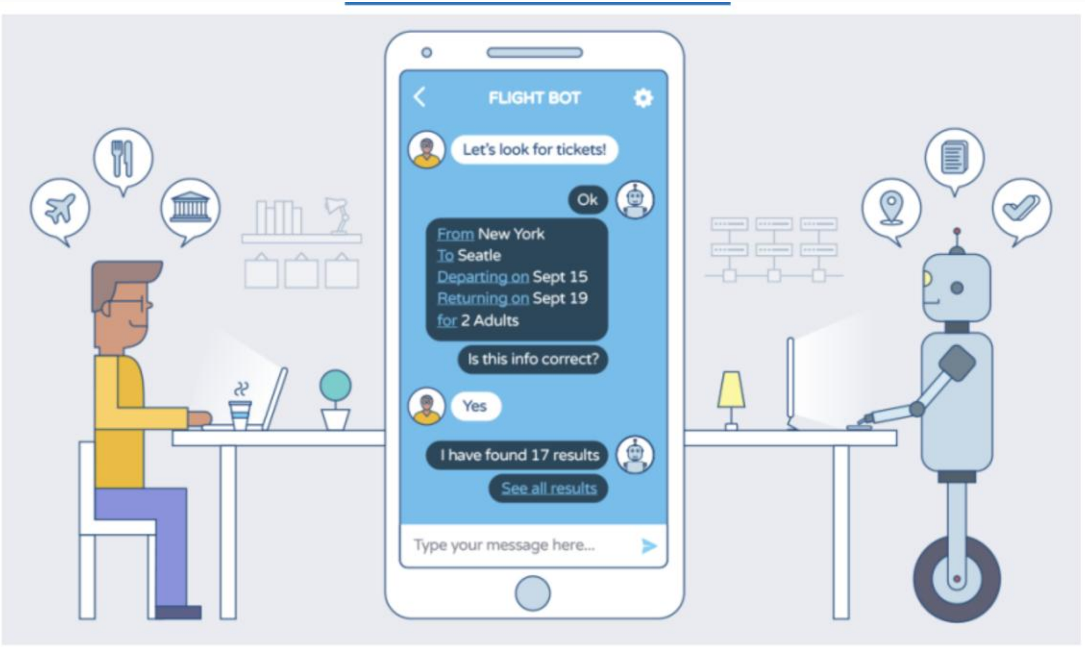
No study would be complete without mentioning the largest rodent in the world, the Capybara.

Sign of the Zodiac:
1. Aries
2. Taurus
3. Gemini...

Most Common Cause of Death in America:
1. Heart Disease
2. Cancer
3. Stroke...

Largest rodent in the world:
1. Capybara
2. Beaver
3. Patagonian Cavies

NLP applications: dialogue systems/chatbots



Let's look for tickets!

Ok

From New York
To Seatle
Departing on Sept 15
Returning on Sept 19
for 2 Adults

Is this info correct?

Yes

I have found 17 results

See all results

Slides Credit: Kevin Gimpel

NLP applications: question answering

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

"Alexa, how's my commute?

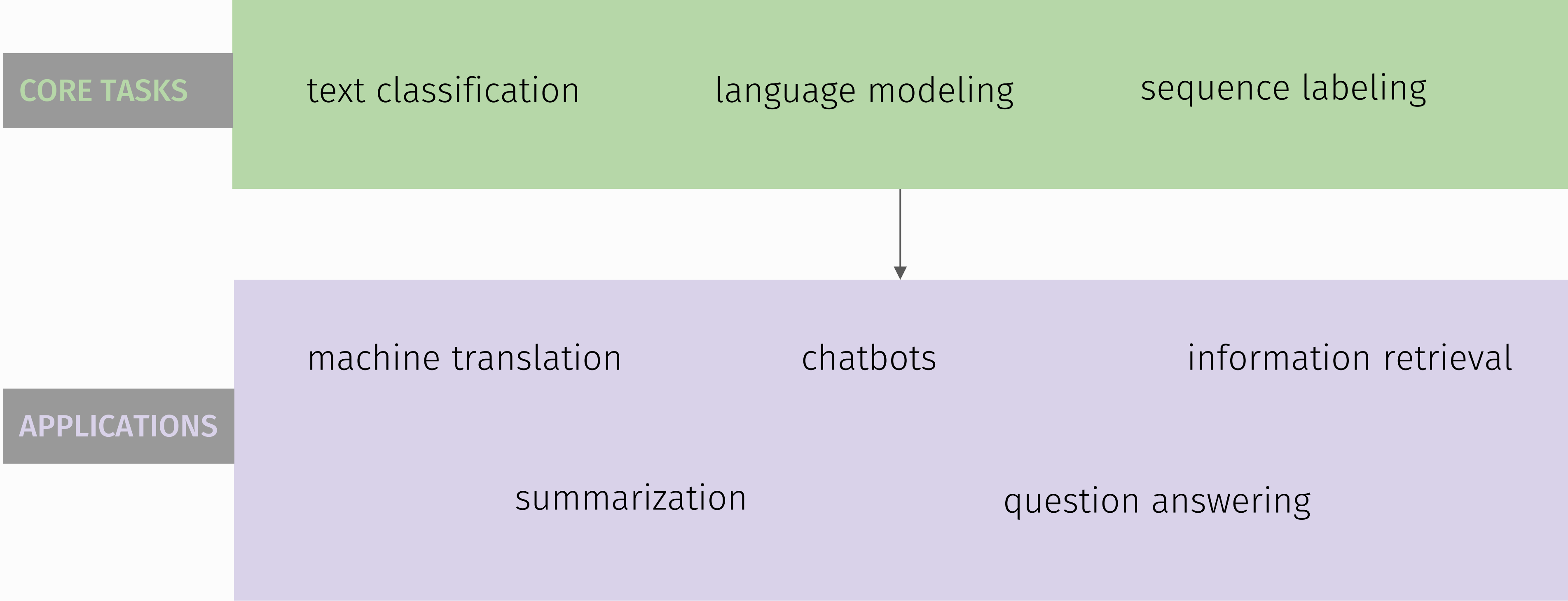
"Alexa, what's the weather?"

"Alexa, did the 49ers win?"

Slides Credit: Kevin Gimpel

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*