CS 1671/2071
Human Language Technologies

Session 4: Machine learning intro, NLP tasks and applications

Michael Miller Yoder

January 22, 2025

School of Computing and Information

Overview: Machine learning intro, NLP tasks and applications

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

Course logistics

- I released a new **optional, extra credit** homework assignment, Homework 0 on getting set up for installing Python packages on the CRCD JupyterHub
  - Is **due tonight, Wed Jan 22, at 11:59pm**
- Homework 1 is **due this Thu Jan 23 at 11:59pm**
- I plan on releasing example projects and a form to submit project ideas you may want to work on Fri Jan 24
  - Project idea submission form will be due next Thu Jan 30
- Check out Pitt's Tech4Good club!

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

Machine learning models and algorithms

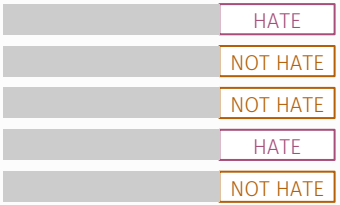
“All models are wrong. Some are useful.”

Machine learning algorithms

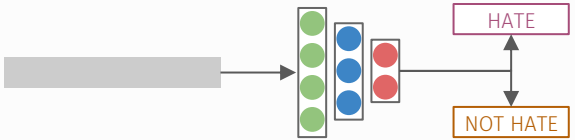
- Algorithms are systematic ways of doing things
- In machine learning, “algorithms” refers to systematic ways of estimating model parameters from data

Supervised machine learning process

Data
(input text, *X*)



Annotate
labels (*Y*)



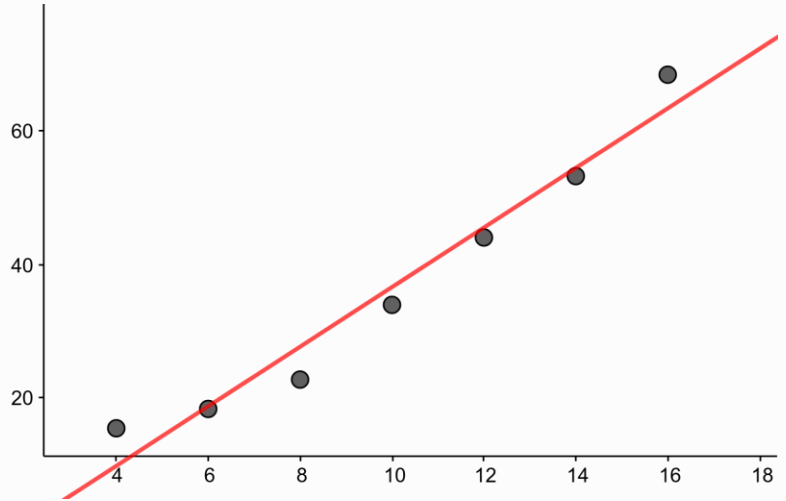
Train a model to
predict labels (*Y*)
from input text (*X*)

Training and test sets (and phases)

- Train parameters of the model on training set (training phase)
  - Sees examples of input and (assumed correct) output that it will mimic
- 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

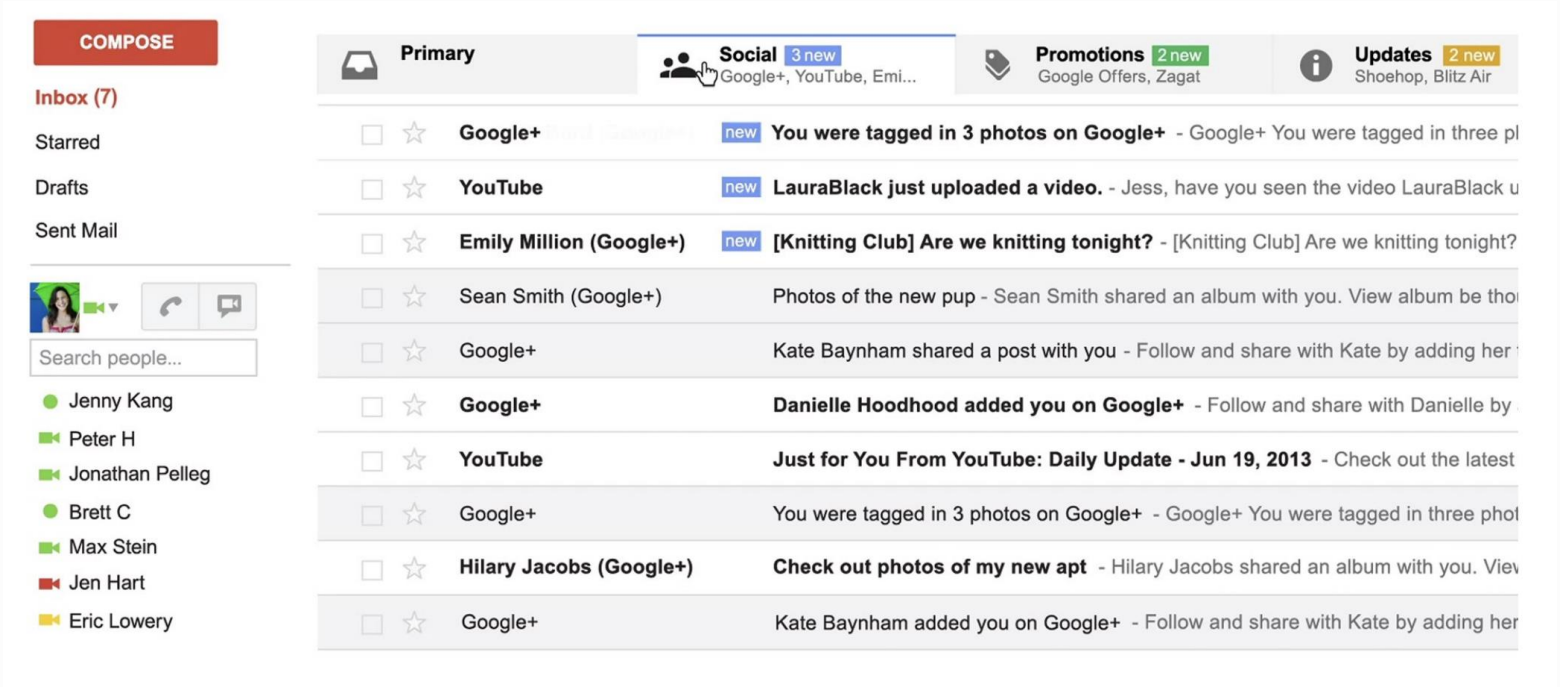
chatbots

information retrieval

summarization

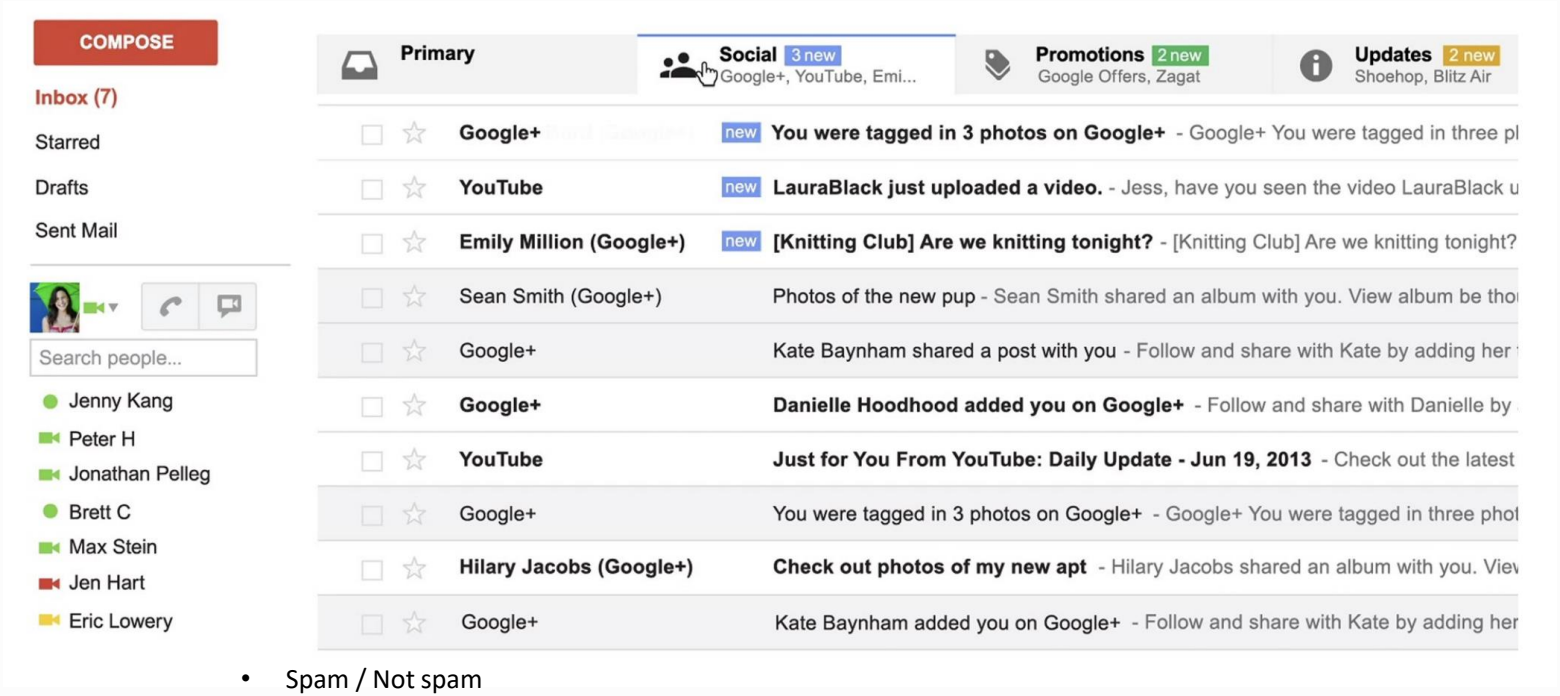
question answering

NLP applications: email classification



Slides Credit: Kevin Gimpel

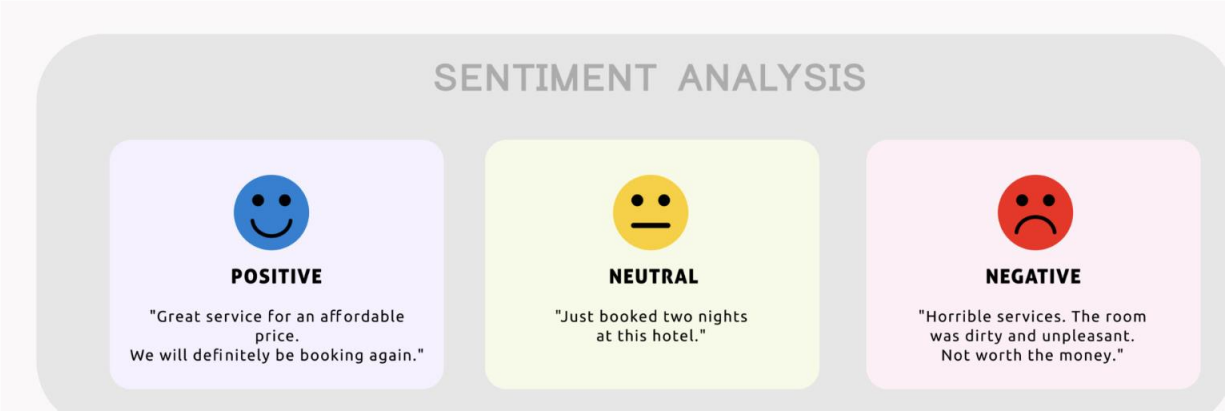
NLP applications: email classification



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

Slides Credit: Kevin Gimpel

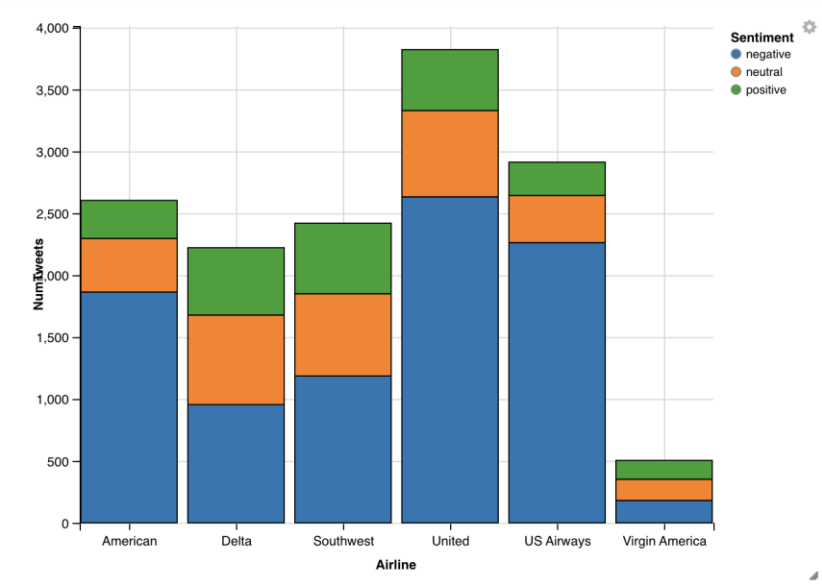
NLP applications: sentiment analysis



- 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

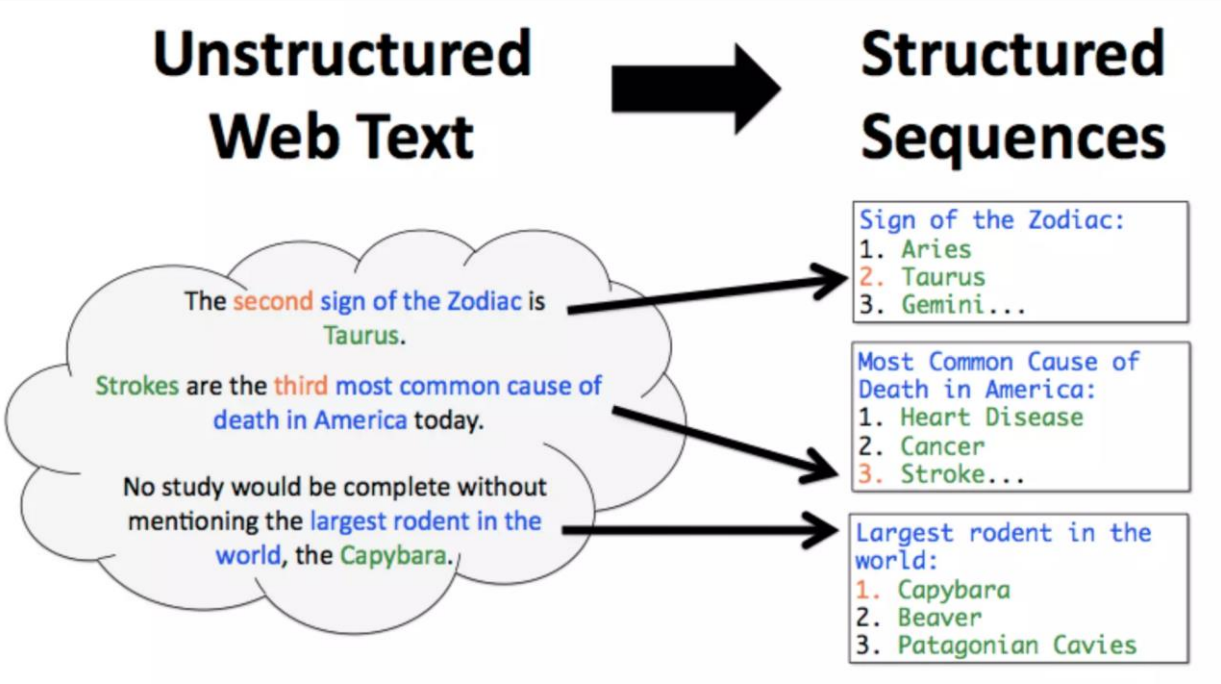
NLP applications: machine translation

Slides Credit: Kevin Gimpel

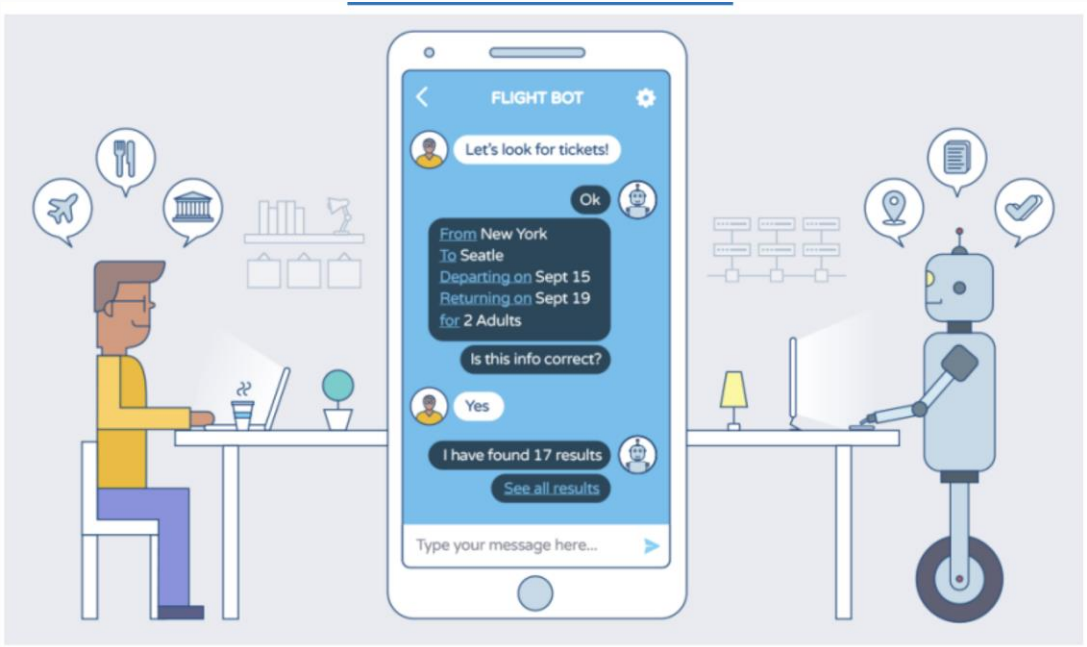
NLP applications: summarization

Slides Credit: Kevin Gimpel

NLP applications: information extraction



NLP applications: dialogue systems/chatbots



Slides Credit: Kevin Gimpel

NLP applications: question answering

amazon alexa

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

"Alexa, how's my commute?"

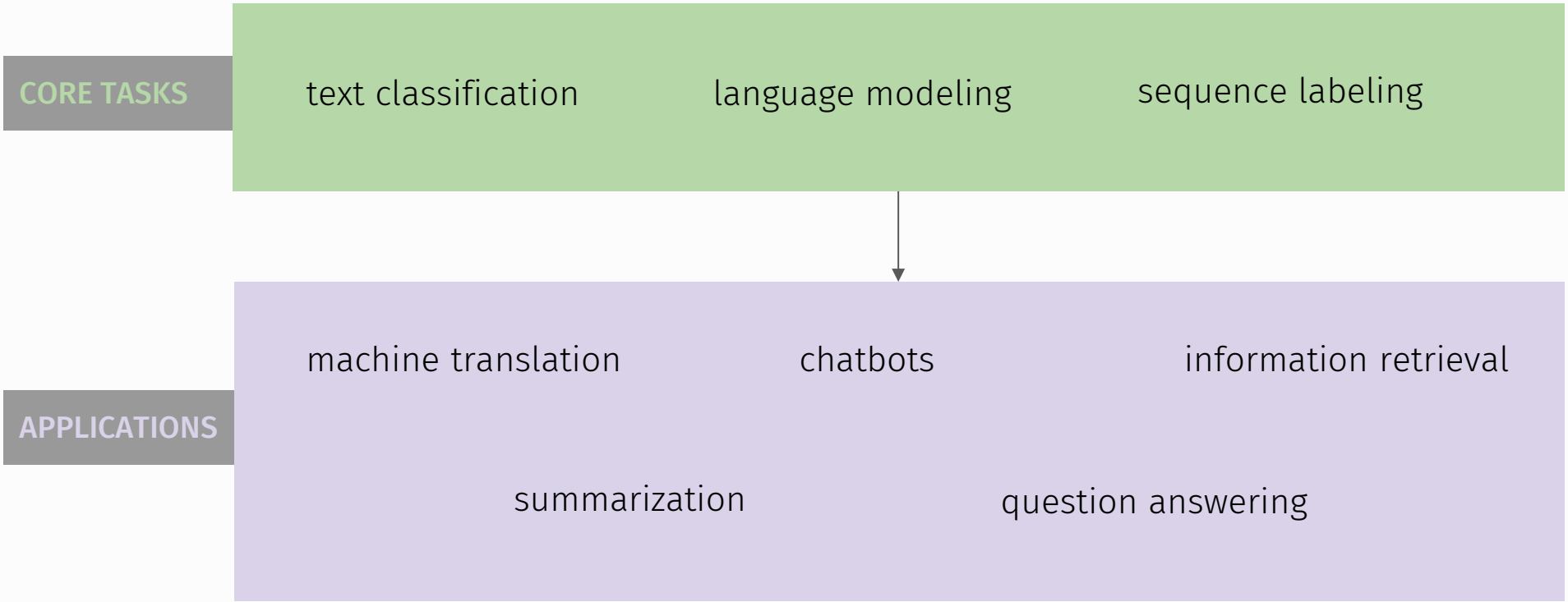
"Alexa, what's the weather?"

"Alexa, did the 49ers win?"

Slides Credit: Kevin Gimpel

NLP core tasks

Core tasks and applications of NLP



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*

Coding activity: clickbait classification

Clickbait classification on JupyterHub

- Click on this nbgitpuller link
- Open **session4_clickbait_classification.ipynb**