108-NLP @ NCTU

e-Lab 4

Boaz



Good morning!

Please make sure your microphone and camera are OFF

We'll take a break or two (4)(4)(1)

Groups: join a group (1~3 people) if you haven't done so

Lab 3, lessons learned

Incorrect probability

gamma p(w2|w1) + (1-gamma) p(w1|w2)

Incorrect N computation of cross entropy -1/N sum(log2(p_i))

unidecode()

Creating a vocabulary of INFREQUENT words

```
def oov_replace(sentence, unknown_tokens):
    for idx, token in enumerate(sentence):
        if token in unknown_tokens:
            sentence[idx] = "<UNK>"
        return sentence
```

How to use Counter()

count.update(sentence)

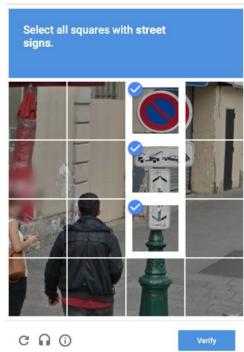
```
sentence = tokenize('the quick brown fox ...')
count = Counter()
for token in sentence:
    count[token] += 1
count = Counter()
```

Something different

Crowdsourcing (眾包)

Why?

- Create datasets
- Add labels to data
- Many other uses

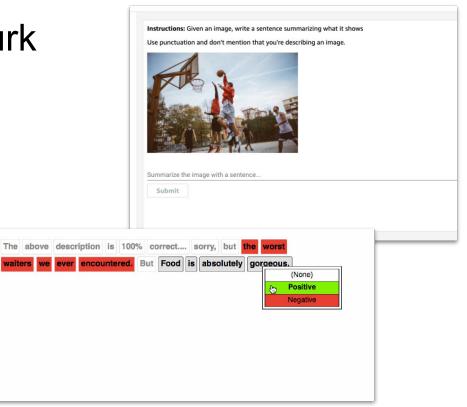




Amazon Mechanical Turk

Crowdsourcing website to hire remote workers to perform small, well-defined tasks that computers are unable to do.





Some MTurk terminology

Requesters are the crowdsourcing "employers"

Workers (or mturkers) are the (invisible) crowdsourcing "employees"

HITs (Human Intelligence Tasks) are the tasks

HIT Group: HITs with identical parameters

Assignment: a HIT assigned to a worker

Requesters

Get Results from Mechanical Turk Workers

Ask workers to complete HITs - Human Intelligence Tasks - and get results using Mechanical Turk, Register Now

As a Mechanical Turk Requester you:

- . Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- · Pay only when you're satisfied with the results



Workers

Make Money by working on HITs

HITs - Human Intelligence Tasks - are individual tasks that you work on. Find HITs now.

As a Mechanical Turk Worker you:

- Can work from home
- · Choose your own work hours
- Get paid for doing good work

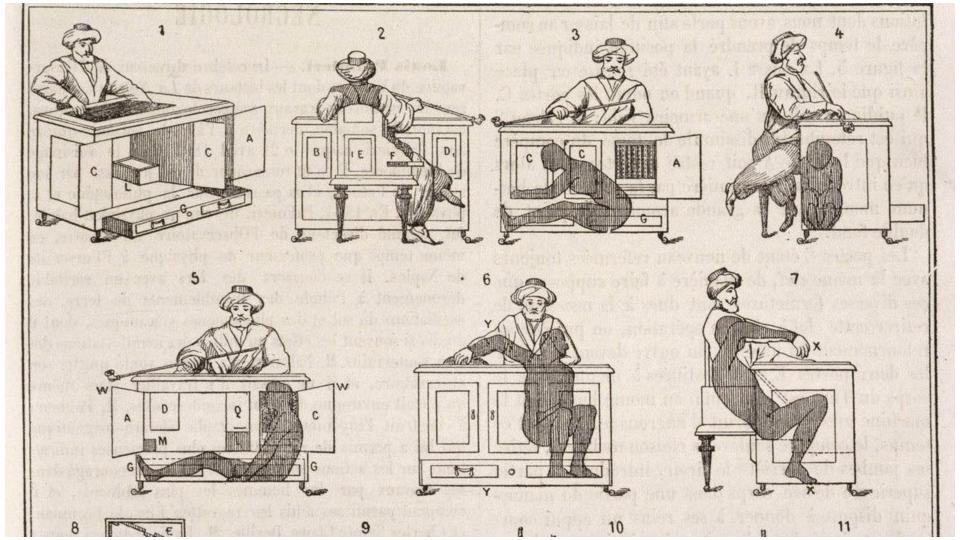


Why is it called **Mechanical Turk**?



Wikipedia Article





Who are the Workers?

Mostly (but not only) from:

- India (mostly use MTurk as main income)
- USA (mostly use MTurk as extra income)

Many MTurker forums/websites, e.g.:

https://turkopticon.ucsd.edu/

https://forum.turkerview.com/



Some VERY IMPORTANT Rules

Treat workers ethically and with respect

Pay QUICKLY and FAIRLY

Estimate the required time, pay at least MINIMUM WAGE

Let's take a look

https://WorkerSandbox.mturk.com/





HIT properties

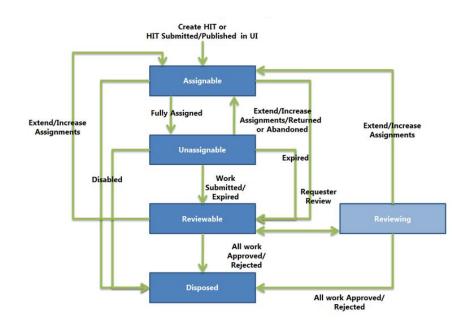
- ☐ Title
- Description
- ☐ Tags
- Assignments per HIT
- ☐ Reward per assignment (\$\$\$)
- ☐ Time allotted per assignment
- HIT expiration
- ☐ How long to pay workers
- Worker requirements

HIT States

- Assignable HIT can be accepted by a worker
- Unassignable HIT accepted, work in progress
- Reviewable HIT finished
- Disposed the HIT has been deleted

Optional:

 Reviewing — HIT currently being reviewed by Requester



Worker Requirements

- Geographical location (e.g., TW)
- Language spoken
- Minimum HIT approval rate
- Minimum number of HITs approved
- More...

```
Employment Industry - Media & Entertainment
Employment Industry - Retail, Wholesale & Distribution
Employment Industry - Software & IT Services
Employment Sector - Non-Profit
Employment Status - Full time (35+ hours per week)
Employment Status - Part time (1-34 hours per week)
Employment Status - Unemployed
Exercise - Every Day
Exercise - Four Plus Times a Week
Exercise - Not at All
Exercise - Once a Week
Exercise - Two to Three Times a Week
Facebook Account Holder
Financial Asset Owned - Certificate of Deposit (CD)
Financial Asset Owned - Checking Account
Financial Asset Owned - Common Share Stocks
Financial Asset Owned - Exchange-Traded Fund (ETF)
Financial Asset Owned - Money Market Account
Financial Asset Owned - Mutual Funds
```

MTurk Options for Requesters

- ✓ Graphical User Interface (GUI)
- ✓ Command Line Interface (CLI)
- ✓ Software Development Kit using the MTurk API

Boto

Python SDK (Software Development Kit)

for

AWS (Amazon Web Services)

https://boto.readthedocs.io/en/latest/ref/mturk.html



Let's create an experiment

Relation between LANGUAGE and COLOR

- ✓ Created an account on Amazon Web Services (AWS) to use the API
- ✓ Created an account on Amazon Mechanical Turk (MTurk) as a REQUESTER
- ✓ Linked the two accounts
- Created an external question https://bshmueli.github.io/108-nctu-mturk/pink.html

Behind the scenes

The Form Action

The form on the external website must post the result data back to Amazon Mechanical Turk using the following URL:

https://www.mturk.com/mturk/externalSubmit

Or, if you are using the Amazon Mechanical Turk sandbox, you should post the result data back to Mechanical Turk using the following sandbox URL:

https://workersandbox.mturk.com/mturk/externalSubmit

https://bshmueli.github.io/108-nctu-mturk/pink.html?assignmentId=123RVWYBAZW00EXAMPLE456RV WYBAZW00EXAMPLE&hitId=123RVWYBAZW00EXAMPLE&turkSubmitTo=https://www.mturk.com&workerId=AZ3 456EXAMPLE

```
<!doctype html>
<html lang="en">
<head>
   <meta content="text/html;charset=UTF-8" http-equiv="Content-Type" />
   <script src='https://s3.amazonaws.com/mturk-public/externalHIT v1.js' type='text/javascript'></script>
</head>
<body>
   <form id="mturk form" method="post" name="mturk form">
       <input type="hidden" id="assignmentId" value="" name="assignmentId"/>
       <h2>Which of the following best represents the color pink?</h2>
       <fieldset style="font-size: 150%">
            <input name="darknessVal" type="radio" value="pink" />
                                                                                                  <span style="</pre>
            <input name="darknessVal" type="radio" value="lightpink" />
                                                                                                  <span style="</pre>
            <|abel><input name="darknessVal" type="radio" value="hotpink" />
                                                                                           <span style="backgr"</pre>
            <|abel><input name="darknessVal" type="radio" value="deeppink" />
                                                                                           <span style="backgr"</pre>
            <|abel><input name="darknessVal" type="radio" value="palevioletred" />
                                                                                           <span style="backgr"</pre>
            <|abel><input name="darknessVal" type="radio" value="mediumvioletred" />
                                                                                           <span style="backgr"</pre>
       </fieldset>
       <br>
       <input type="submit">
   </form>
   <script language='Javascript'>turkSetAssignmentID();</script>
</body>
</html>
```

https://docs.aws.amazon.com/AWSMechTurk/latest/AWSMturkAPI/ApiReference ExternalQuestionArticle.html

Team Assignment:

Create an MTurk Experiment to

Label the emotion in a short text

Group Assignment: same teams as Final Project

Only one submission per team

Emotion Models

Discrete

- Happy
- Sad
- Angry
- Jealous
- Surprised
- ...

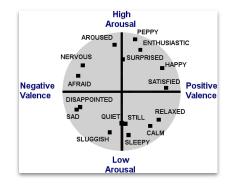


Emotion Models

Dimensional

2 or 3 (almost) independent dimensions

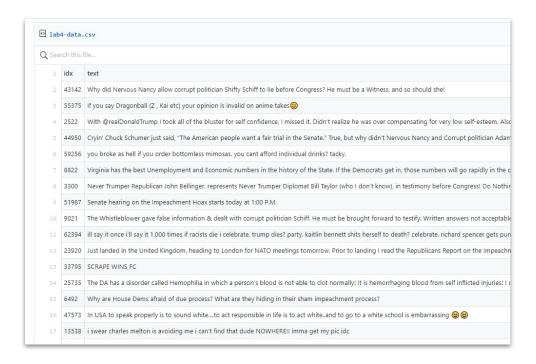
- Valence (Pleasure)
- Arousal
- Dominance



	Hostility	Guilt	Fear	Sadness	Shyness	Fatigue	Surprise	Calmness	Attentive	Self-assurance	Joviality
alence											
Mean	-3.73	-3.48	-3.28	-3.26	-2.65	-2.04	0.58	2.53	3.02	3.19	4.17
SD	0.78	1.05	0.88	1.15	0.98	1.35	0.93	1.07	1.10	0.97	0.58
Assignment	NEG	NEG	NEG	NEG	NEG	NEG	POS	POS	POS	POS	POS
rousal											
Mean	3.08	1.69	2.36	0.04	0.44	-1.00	2.62	-0.91	2.18	3.50	3.93
SD	1.12	1.76	1.15	2.16	1.76	2.18	1.28	2.06	1.56	1.03	0.77
Assignment	н	MD	н	LO	LO	LO	н	LO	HI	HI	HI
ominance											
Mean	1.81	-1.24	-0.93	-2.52	-3.10	-3.34	0.12	-1.73	2.03	3.38	2.80
SD	1.64	1.68	1.64	1.56	1.56	1.41	1.40	1.72	1.52	1.06	1.16
Assignment	н	LO	LO	LO	LO	LO	MD	LO	н	н	н

Assignment (1)

Data: https://bit.ly/nlp-amt-data



Assignment (2)

Data: https://bit.ly/nlp-amt-data

Each HIT is one tweet

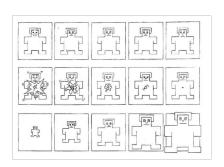
Display the tweet and a VAD form

Make sure worker reads tweet: enable SUBMIT only after *C* * (number of words)

Return:

- VAD scores (1~5, 1~5, 1~5)
- Time invested by worker

Scale	Valence rating	Arousal rating	Dominance rating
5	Pleasant	Excited	Dependent
4	Pleased	Wide-awake	Powerlessness
3	Neutral	Neutral	Neutral
2	Unsatisfied	Dull	Powerful
1	Unpleasant	Calm	Independent



Assignment (3)

- Worker Qualification
 - Residing in Taiwan or English speaking countries
 - Agree to work on offensive materials (tweets can be offensive!)
- Payment according to length
 - Calculate average number of tokens avgw
 - \$0.50 for short tweets (less than avgw)
 - \$1.00 for long tweets (more than avgw)
- 3 assignments per HIT
- HITs expire in 8 weeks
- Once started, worker has 30 minutes to complete each HIT
- Payment is 1 hour after completion

Assignment (4)

Set user name to NCTU_{Team_Name}

For example, NCTU_Team_Pudding



Host HITs on External Website (ExternalQuestion)

<u>Fast.io</u> is free, can serve HTML files from GDrive, Dropbox, Github, ...

Deliverable: Colab notebooks

3 parts:

- 1. Create and upload the HITs (using **Full Access** credentials)
- 2. Work as MTurker:)
- 3. Fetch and save the results to a CSV file (using **Read-only** credentials)
 - Computes average for each tweet
 - CSV columns: idx, avg_valence, avg_arousal, avg_dominance, time

See <u>Lab 4 Assignment</u> for detailed homework information

SOP for assignment submission

Open your Python notebook
Rename your notebook to lab4- <studentid>.ipynb</studentid>
"Restart and run all"
Make sure the output is correct
Download the Python file (File, Download .py)
Share the Google colab using "Get shareable link"; Copy link
Submit both link and file in e3:
Paste link into e3 (link starts with https://colab.research.google.com/)
Upload file into e3 (file name is lab4- <studentid>.py)</studentid>
Also attach credentials.csv file for user with Read-Only access

Appendix

Useful Links

Mechanical Turk

https://www.mturk.com/

https://requester.mturk.com/

Sandbox

https://requestersandbox.mturk.com/

https://workersandbox.mturk.com/

Amazon Web Services

https://aws.amazon.com/

Developer Reference

https://docs.aws.amazon.com/AWSMechTurk/latest/AWSMechanicalTurkRequester/



https://docs.aws.amazon.com/AWSMechTurk/latest/AWSMechanicalTurkReguester/amt-dg.pdf

API Reference

https://docs.aws.amazon.com/AWSMechTurk/latest/AWSMturkAPI/



https://docs.aws.amazon.com/AWSMechTurk/latest/AWSMturkAPI/amt-API.pdf

Python SDK (Boto 3)

https://boto.readthedocs.io/en/latest/ref/mturk.html

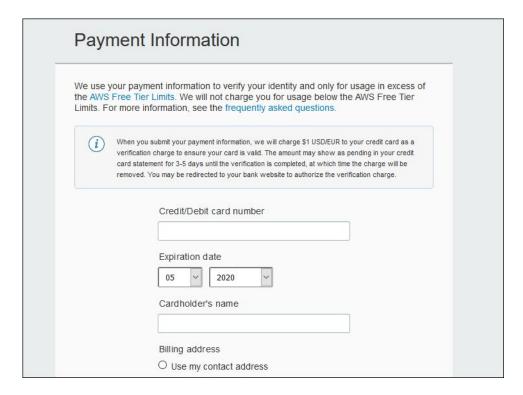
Setup

https://docs.aws.amazon.com/AWSMechTurk/latest/AWSMechanicalTurkGettingStartedGuide/SetUp.html

Follow Steps 1~5

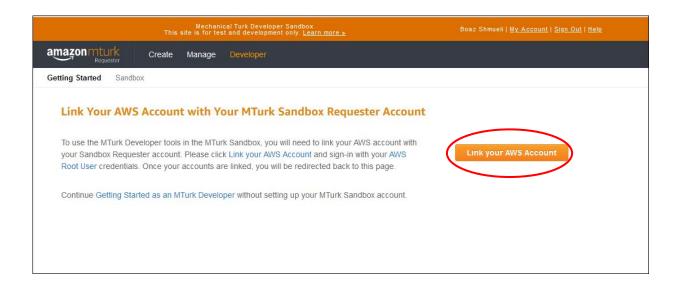
- . Step 1: Sign Up for an AWS Account
- Step 2: Create a Requester Account
- Step 3: Link Your AWS account to your MTurk Requester account
- Step 4: Create an IAM User
- Step 5: Set up the Developer Sandbox
- Step 6: Set up an AWS SDK
- Step 7: Prepay for Your HITs or Enable AWS Billing

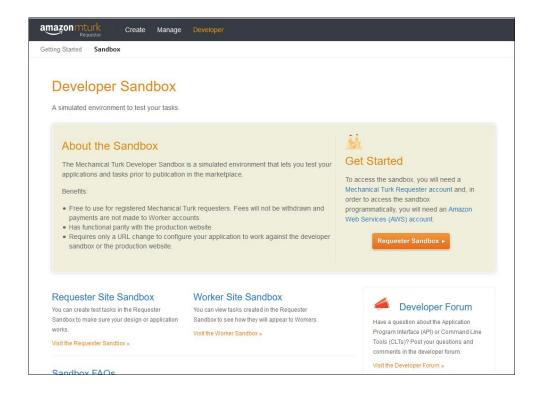
No need for Credit Card account (hopefully!)

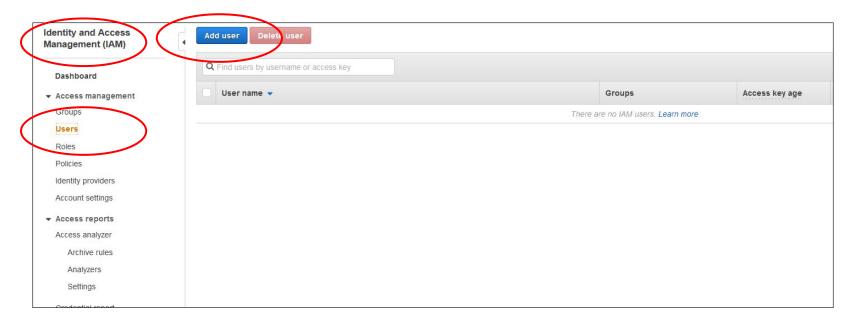


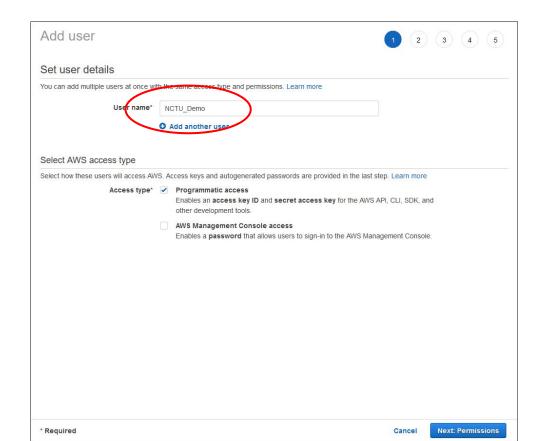
Create an Amazon Mechanical Turk Account Contact Information Please enter your contact information below. * Required Company name: * (See details) Your phone number may be used to verify your identity. Tell us how you plan to use MTurk: Address: * O Conduct surveys or market research O Moderate user-generated content, e.g. comments, images, or videos Annotate data sets to train machine learning models City: * O Collect data sets, e.g. gather the contact info for a list of businesses State, province or region: O Clean or categorize data for data analysis O My use for MTurk isn't listed above Zip or postal code: * How often do you expect to use MTurk? O I have a recurring need, e.g. processing data regularly Country: * O I have a one-time need, e.g. a specific project Don't see your country? United States I will occasionally use MTurk Phone number: * an account, you agree to Amazon Mechanical Turk's Participation Agreement. Please see Your phone number may be used to verify your identity. **Create Account**

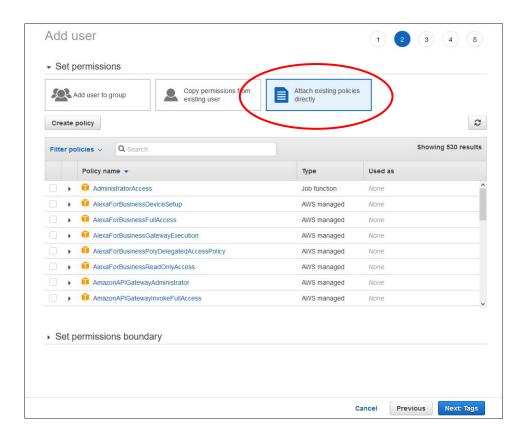
Link Sandbox with AWS account

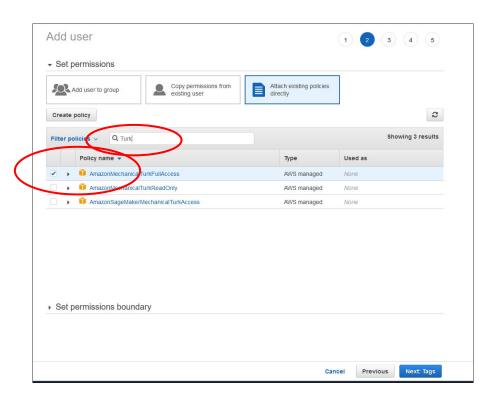


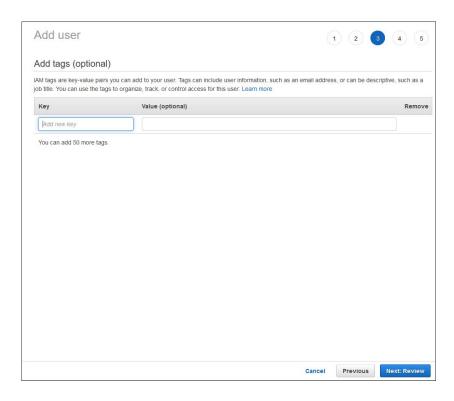


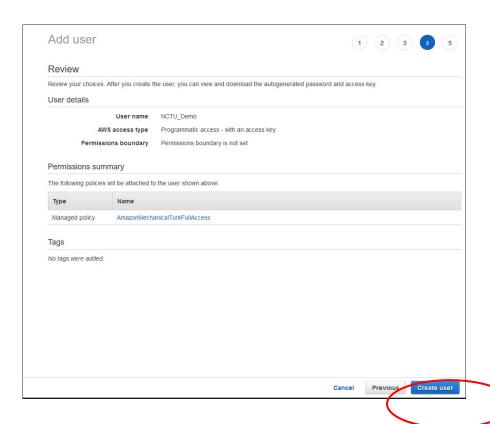


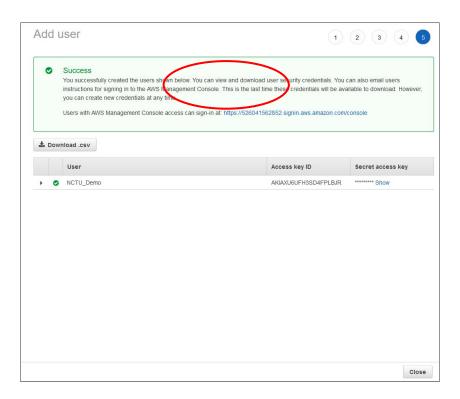












credentials.csv

```
credentials.csv - Notepad

File Edit Format View Help

User name, Password, Access key ID, Secret access key, Console login li

NCTU_Demo,, AKIAXU6UFH3SD4FPLBJR, K1K4s0dVwN4qz4JRTstB91UsM2fwVxtcy/6
```

Useful Links

Mechanical Turk

https://www.mturk.com/

https://requester.mturk.com/

Sandbox

https://requestersandbox.mturk.com/

https://workersandbox.mturk.com/

Amazon Web Services

https://aws.amazon.com/

Developer Reference

https://docs.aws.amazon.com/AWSMechTurk/latest/AWSMechanicalTurkRequester/



https://docs.aws.amazon.com/AWSMechTurk/latest/AWSMechanicalTurkRequester/amt-dg.pdf

API Reference

https://docs.aws.amazon.com/AWSMechTurk/latest/AWSMturkAPI/



https://docs.aws.amazon.com/AWSMechTurk/latest/AWSMturkAPI/amt-API.pdf

Python SDK (Boto 3)

https://boto.readthedocs.io/en/latest/ref/mturk.html