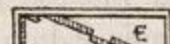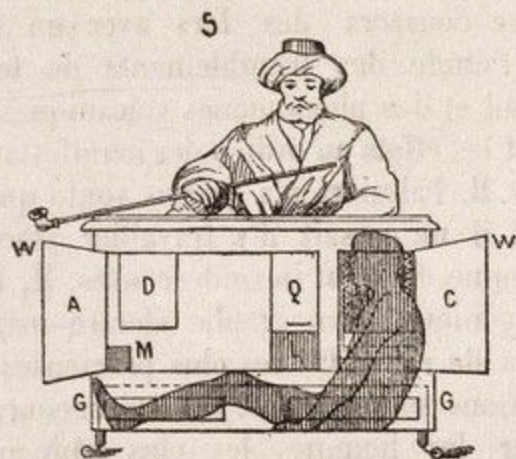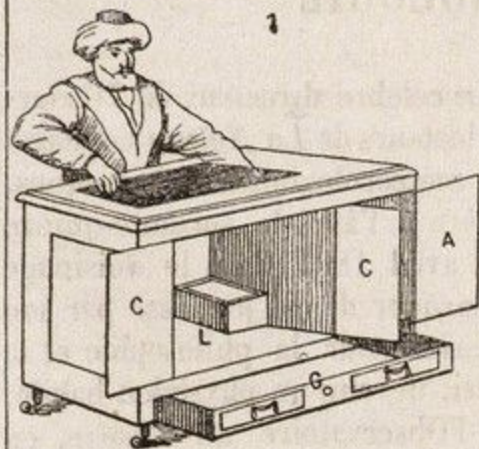# 108-NLP @ NCTU

e-Lab 5

Boaz

Good morning!

Please make sure your **microphone** and **camera** are **OFF**

We'll take a break or two 🕧🕝🕜🕕

Lab 4 (MTurk)

Step 1: Upload HITs

Step 2: MTurker
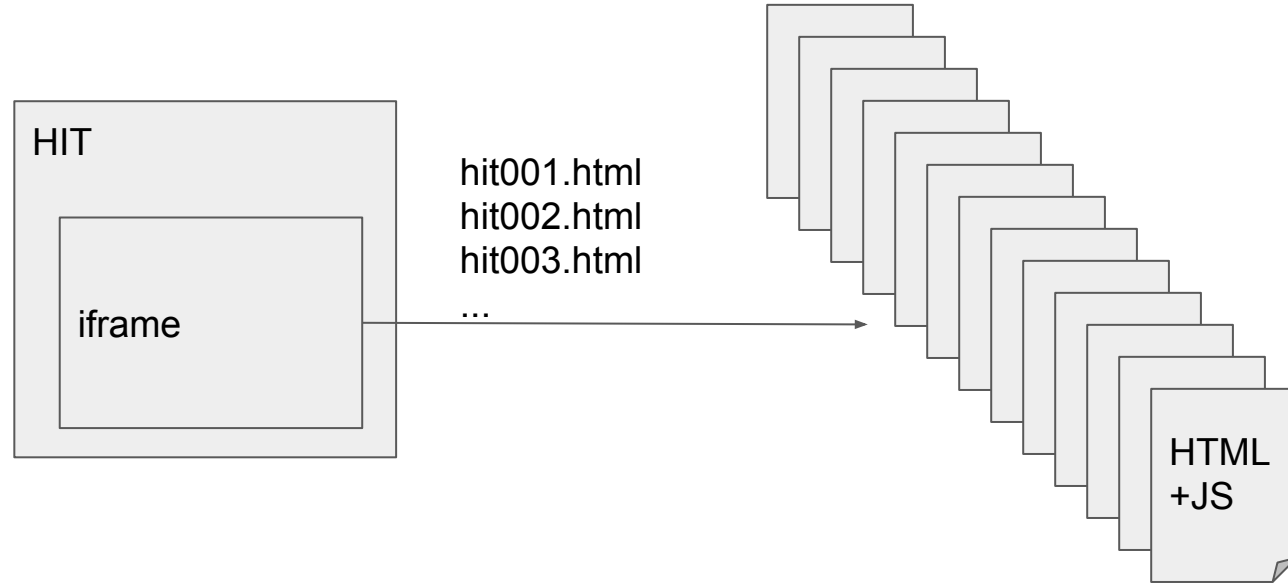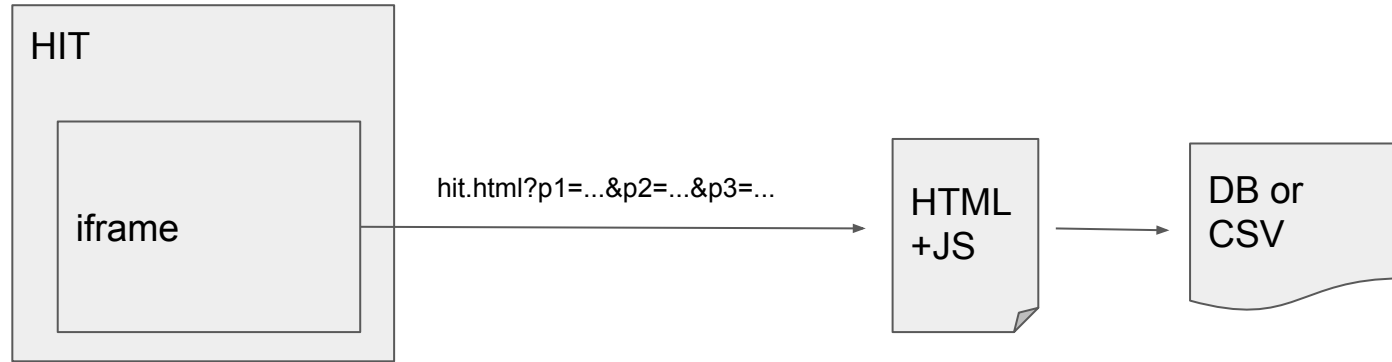
Step 3: Download assignments

ExternalQuestion

Why?

The most flexible

Multiple HTML Files

HIT

iframe

hit001.html
hit002.html
hit003.html

...

HTML
+JS

One HTML file

HIT

iframe

hit.html?p1=...&p2=...&p3=...

HTML +JS

DB or CSV

A world of languages
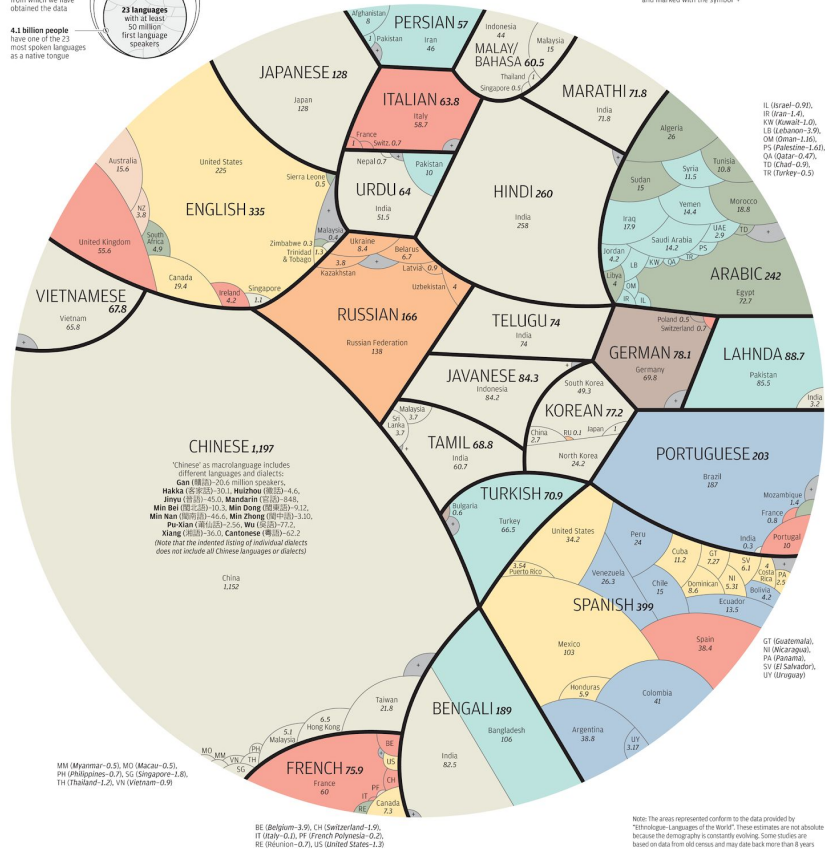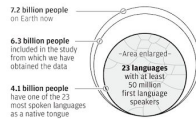
There are at least 7,102 known languages alive in the world today. Twenty-three of these languages are a mother tongue for more than 50 million people. The 23 languages make up the native tongue of 4.1 billion people. We represent each language within black borders and then provide the numbers of native speakers (in millions) by country. The colour of these countries shows how languages have taken root in many different regions

Regions in which these languages are present

Europe East
Asia Major
Europe West
North America
Asia Minor
Middle East
South America
Africa
Oceania

Countries whose figures in each language is too small to be represented have been put into a single group and marked with the symbol '+'

7.2 billion people on Earth now

6.3 billion people included in the study from which we have obtained the data

4.1 billion people have one of the 23 most spoken languages as a native tongue

~Area enlarged~
23 languages with at least 50 million first language speakers

PERSIAN 57
Afghanistan 8
Pakistan
Iran 46
Indonesia 44
Malaysia 15
MALAY/BAHASA 60.5
Thailand 0.5
Singapore 0.5
MARATHI 71.8
India 71.8

JAPANESE 128
Japan 128

ITALIAN 63.8
Italy 58.7
France
Switz. 0.7

IL (Israel–0.9), IR (Iran–1.4), KW (Kuwait–1.0), LB (Lebanon–3.9), OM (Oman–1.16), PS (Palestine–1.61), QA (Qatar–0.47), TD (Chad–0.9), TR (Turkey–0.5)

Algeria 26
Tunisia 10.8
Sudan 15
Syria 11.5
Morocco 16.8
Yemen 14.4
Iraq 17.9
Saudi Arabia 14.2
UAE 2.9
TD
Jordan 4.2
LB
KW QA OM PS
Libya 4
OM
IL

ARABIC 242
Egypt 72.7

Australia 15.6
NZ 3.8
United States 225
Sierra Leone 0.5
Nepal 0.7
Pakistan 10
URDU 64
India 51.5
HINDI 260
India 258

ENGLISH 335
South Africa 4.9
United Kingdom 55.6
Canada 19.4
Ireland 4.2
Singapore 1.1
Zimbabwe 0.3
Trinidad & Tobago 1.3
Malaysia 0.4
Ukraine 8.4
Belarus 6.7
Latvia 0.9
Kazakhstan 3.8
Uzbekistan 4

VIETNAMESE 67.8
Vietnam 65.8

RUSSIAN 166
Russian Federation 138

TELUGU 74
India 74

GERMAN 78.1
Germany 69.8
Poland 0.5
Switzerland 0.7

LAHNDA 88.7
Pakistan 85.5
India 3.2

JAVANESE 84.3
Indonesia 84.2

KOREAN 77.2
South Korea 49.3
Japan 1
North Korea 24.2

CHINESE 1,197
'Chinese' as macrolanguage includes different languages and dialects:
Gan (贛語)–20.6 million speakers, Hakka (客家語)–30.1, Huizhou (徽語)–4.6, Jinyu (晋語)–45.0, Mandarin (官話)–848, Min Bei (閩北語)–10.3, Min Dong (閩東語)–9.12, Min Nan (閩南語)–46.6, Min Zhong (閩中語)–3.10, Pu-Xian (莆仙語)–2.56, Wu (吳語)–77.2, Xiang (湘語)–36.0, Cantonese (粵語)–62.2
(Note that the indented listing of individual dialects does not include all Chinese languages or dialects)

China 1,152

TAMIL 68.8
India 60.7
Malaysia 3.7
Sri Lanka 3.7

TURKISH 70.9
Turkey 66.5
Bulgaria 0.6

PORTUGUESE 203
Brazil 187
Mozambique 1.4
France 0.8
Portugal 10

United States 34.2
Peru 24
Cuba 11.2
GT 7.27
SV 6.1
NI 5.3?
Costa Rica 4.1
Bolivia 4.2
PA
Venezuela 26.3
Dominican 8.6
Ecuador 13.5
India 0.3

SPANISH 399
Mexico 103
Chile 15
Colombia 41
Spain 38.4
Argentina 38.8
3.54 Puerto Rico
Honduras 5.9
UY 3.17

GT (Guatemala), NI (Nicaragua), PA (Panama), SV (El Salvador), UY (Uruguay)

Taiwan 21.8
6.5 Hong Kong
5.1 Malaysia

BENGALI 189
Bangladesh 106
India 82.5

MM (Myanmar–0.5), MO (Macau–0.5), PH (Philippines–0.7), SG (Singapore–1.8), TH (Thailand–1.2), VN (Vietnam–0.9)

MM MO PH VN TH SG

FRENCH 75.9
France 60
BE
US
PF
IT
CH
Canada 7.3
RE

BE (Belgium–3.9), CH (Switzerland–1.9), IT (Italy–0.1), PF (French Polynesia–0.2), RE (Réunion–0.7), US (United States–1.3)

Note: The areas represented conform to the data provided by "Ethnologue–Languages of the World". These estimates are not absolute because the demography is constantly evolving. Some studies are based on data from old census and may date back more than 8 years

8

Qualification

Qualifications Required

✔ Adult Content Qualification is 1

✔

Qualifications Required

✔ Offensive tolerance is 100

✔ Location is one of: TW, US

Some User Interface Examples

Please rate the arousal of the tweet:

○ Calm

○ Dull

● Neutral

○ Wide-awake

○ Excited

Please rate the dominance of the tweet:

○ Independent

⚠ Please select one of these options.

○ Neutral

○ Powerlessness

○ Dependent

[Submit]

Report this HIT ▾ | Why Report ▾

11

Tweet content

Adam Schiff will be compelled to testify should the Democrats decide, despite the fact that my presidential conversations were totally appropri

Your emotion ratings

🙂 Valence rating*

- ◯ Pleasant
- ◯ Pleased
- ◯ Neutral
- ● Unsatisfied
- ◯ Unpleasant

🔥 Arousal rating*

- ◯ Excited
- ● Wide-awake
- ◯ Neutral
- ◯ Dull
- ◯ Calm

👥 Dominance rating*

- ◯ Dependent
- ◯ Powerlessness
- ◯ Neutral
- ● Powerful
- ◯ Independent

Report this HIT ▾ | Why Report ▾

12

if you say Dragonball (Z , Kai etc) your opinion is invalid on anime takes😊

Valence rating

○ 5 ○ 4 ○ 3 ○ 2 ○ 1

Arousal rating

○ 5 ○ 4 ○ 3 ○ 2 ○ 1

Dominance rating

○ 5 ○ 4 ○ 3 ○ 2 ○ 1

Submit

How do you feel about this tweet?

Like all of Banco staff went on lunch the same time man .. wtf

Valence rating

Select your feeling about valence:

Pleasant ⌄

Arousal rating

Select your feeling about arousal:

Excited ⌄

Dominance rating

Select your feeling about dominance:

Dependent ⌄

○Level 3: Neutral

○Level 2: Unsatisfied

○Level 1: Unpleasant

Arousal rate

○Level 5: Excited

○Level 4: Wide-awake

○Level 3: Neutral

○Level 2: Dull

○Level 1: Calm

Dominance rate

○Level 5: Dependent

○Level 4: Powerlessness

○Level 3: Neutral

○Level 2: Powerful

○Level 1: Independent

16 seconds remaining

while doing so. She lost Congress once, she will do it again!

Choose the emotion that matches your feeling.

How much valence do you feel in this tweet? (Valence rating)

○ Pleasant   ○ Pleased   ○ Neural   ○ Unsatisfied   ○ Unpleasant

How much Arousal do you feel in this tweet? (Arousal rating)

○ Excited   ○ Wide-awake   ○ Neural   ○ Dull   ○ Calm

How much dominance do you feel in this tweet? (Dominance rating)

○ Dependent   ○ Powerlessness   ○ Neural   ○ Powerful   ○ Independent

Time left: 5 seconds

React on this long Tweet !!! (HIT Details)   ☐ Auto-accept next HIT   Request

I believe in sending flowers, and paying for dinner, opening doors plans, planning road trips, and PDA and saying I love you and alw fucking old fashioned. ☠

Please read the tweet above before you answer the question below !!

How much pleasure you feel about the tweet ?
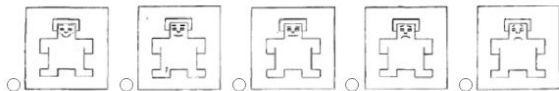
How much arousal you feel about the tweet ?

How much dominance you feel about the tweet ?
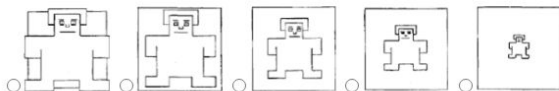
7 Seconds more to load the submit button

Tweet

Donald J Trump Simpson
@FackDonaldTrump

The Never Trumper Republicans, though on respirators with not many left, are in certain ways worse and m
than the Do Nothing Democrats. Watch out for them, they are human scum!

Q: After reading the above tweet, please choose the best-match emotion respectively.

Valence rating

○ 5 Pleasant

○ 4 Pleased

○ 3 Neutral

○ 2 Unsatisfied

○ 1 Unpleasant

Arousal rating

Please read the following tweet:

When the impeachment-sham inevitably joins comey-gate, collusion-delusion, mueller-investigation-nothing-burger, and fake-dossier in history's trash heap, what hoax will the radical left, deep state, and left leaning media push ne

After re

How mu

○ 5 (Pleas

How mu

Please read the following

When the impeachment-shan
burger, and fake-dossier in hist
push next?

More Information about Emotion labelling ✕

Reading this tweet, label a emotion that best describes it. The question design is based on Valence-Arousal-Dominance (VAD) model. It will give a comprehensive score after considering three independent dimensional emotion, which are valence (pleasant), arousal (excited), and dominance (powerful).

Close

After reading this tweet, p ions. ?

How much "pleasant" does this tweet imply ?

○ 5 (Pleasant)   ○ 4 (Pleased)   ○ 3 (Neutral)   ○ 2 (Unsatisfied)   ○ 1 (Unpleasant)

1/3
The best thing that could happen to Democrats, happened today!
"Hell yea, we're coming for your AR-15s and AK-47s", Beto

the ten senators he's closest to, and if each told the senators how much he fears for the country
untrammeled and unhinged, at the helm...Trump would be gone."
"7909","1/3
The best thing that could happen to Democrats, happened today!
""Hell yea, we're coming for your AR-15s and AK-47s"", Beto O'Rourke packed it in!

Like many, I wish we could!
However, their are too many people that are proud of their killing ability!

These people saw no war!"
"109","
"24310"
is disg

1/3 The best thing that could happen to Democrats, happened today! "Hell yea, we're coming for your AR-15s and AK-47s", Beto O'Rourke packed it in! Like many, I wish we could! However, their are too many people that are proud of their killing ability! These people saw no war!

Work Duration

AcceptTime → SubmitTime

AcceptTime → → SubmitTime

Multiple Experiments

If running multiple experiments, store HIT ids in a DB after creating them
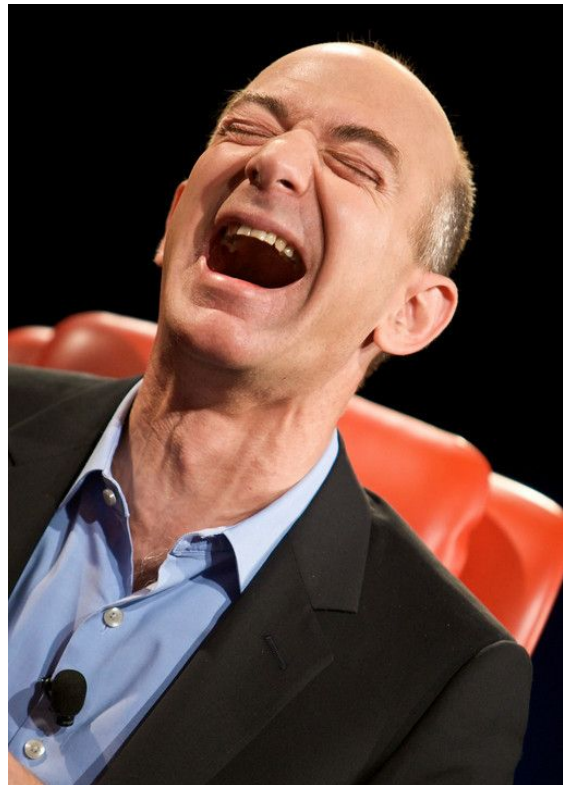
How much do you pay?

Long 200 x $1.00 x 3 = $600

Short 177 x $0.50 x 3 ~ $270

Subtotal $600 + $270 = $870

Amazon Fee = 20% x $870 ~ $170 ⟶

Jeff Bezos BEST laughs compilation EVER !!!

Remember

First: start with the Sandbox

- Process/verify data

Then: SMALL live experiment ($)

- Process/verify data

Finally: BIG live experiment ($$$)

- Process/verify data

Questions?

Lab 5: Project Prep

EmotionGIF 2020

Twitter demo

- https://twitter.com

EmotionGIF 2020

- https://sites.google.com/view/emotiongif-2020/

Why is this useful?

DEMO

Colab Notebook

https://colab.research.google.com/drive/1QEFndCG_6LNi3vIPCDDeA7mQLT2kM0FZ

Assignments

For Lab 5 we will be using the following **secret NCTU-only** website:

https://competitions.codalab.org/competitions/25115?secret_key=e3d89637-74fe-41e4-acb5-c734d91edebe

- Register on Codalab (all)
- Create a team (one person)
- Add members to team
- Download data

Do NOT share or post the data

Part 1: Data Exploration

Write code that will compute and print the following training data statistics:

- Number of samples that have N categories, for 1 <= N <= 6
- Category distribution (how many times does each category appear)
- 10 most common pairs of co-occurring categories
- Optional: any other statistics that you think are useful

Part 2: Prediction

Write code that will create the following 2 prediction files.

1. **Majority Prediction (Baseline!)**
   - ○ Always predict the same 6 majority labels
2. **Naive Bayes (expect bad performance -- for PRACTICE only!)**
   - ● **Use 90% for training, use 10% for testing**
   - ● Only consider first GIF category in each sample (for simplification)
   - ● Compute sklearn TF-IDF vectorizer (TfidfVectorizer), **default settings**
   - ● Use sklearn Multinomial Naive Bayes classifier (MultinomialNB), default **settings**
   - ● For each sample, pick the 6 highest-probability categories
   - ● **Check: is the performance on your test data similar to the performance on the evaluation data?**

Upload the predictions to the **secret NCTU-only** website.

For more information about the prediction files to be uploaded, see Phase 1 and Submission Format.

Deliverables

✓ Colab Notebook with Part 1 & Part 2
✓ Two submissions to **secret** NCTU website

SOP for assignment submission

☐ Open your Python notebook

☐ Rename your notebook to lab5-<TeamName>.ipynb

☐ "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 lab5-<TeamName>.py)

How to proceed

- Other vectorizers
- Other classifiers
- Consider the data imbalance
- Consider all categories, not only the first one
- Using ideas we learned in class/labs
- Check carefully -- we didn't use all the data fields in the file
- Use techniques from Machine Learning/Deep Learning
- Lots of online info
- Think out of the box :)