

COMP 4446 / 5046

Lecture 8: Data – Annotation and Crowdsourcing

Jonathan K. Kummerfeld

Semester 1, 2025



THE UNIVERSITY OF
SYDNEY

menti.com
7824 5802



[Roundly-condemned
headlinese initiative
shuttered indefinitely.]

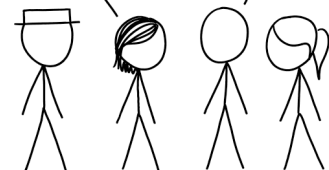
Headline Words

MAYBE ROB SHOULDN'T HOST THE PARTY. HE
HAS CATS AND SOME OF US ARE ALLERGIC.

WOW, MAJOR SNUB FOR
WIDELY-TOUTED TOP SPOT
AS LAVISH GALA BID NIXED.

WHY ARE YOU TALKING
SO WEIRD? PLEASE STOP.

ILL-ADVISED SCHEME MULLED
AS TENSION MOUNTS AMID
GROWING BACKLASH.



MY PROJECT TO SPEAK ONLY IN WEIRD
HEADLINE WORDS DIDN'T LAST LONG.

Source: <https://xkcd.com/2584/>

1



COMP 4446 / 5046
Lecture 8

Assignments

Data Sources

Annotation
Crowdsourcing
Workshop Preview



menti.com 7824 5802

Assignment 3

- Due tomorrow
- Remember, you have 5 slip days total for the semester.
You can check how many you have used in Canvas

Assignment 4

- Out tomorrow
- Groups of 2-3
- Very different in style from previous assignments!
- You can use slip days, but they are charged to all group members. All of your group members must have sufficient days. Late submissions get 0.

2



COMP 4446 / 5046
Lecture 8

Data Sources

Annotation

Crowdsourcing

Workshop Preview



menti.com 7824 5802

Data Sources

3



COMP 4446 / 5046
Lecture 8

Data Sources

Annotation

Crowdsourcing

Workshop Preview



menti.com 7824 5802

Where does the data to train these models come from?

Explore with an example:



FineWeb

<https://huggingface.co/datasets/HuggingFaceFW/fineweb>

4



Data Sources
Annotation
Crowdsourcing
Workshop Preview



menti.com 7824 5802

Step 1: Download data



250 billion pages
17 years

language ↕	% ▼
eng	46.4328
deu	5.8355
rus	5.5709
jpn	4.7475
fra	4.6351
spa	4.6199
zho	3.7971
ita	2.7243
<unknown>	2.7090
nld	2.1566
pol	1.7413
por	1.1169
ces	1.0611
vie	1.0570

<https://huggingface.co/datasets/HuggingFaceFW/fineweb>

5



Data Sources
Annotation
Crowdsourcing
Workshop Preview



menti.com 7824 5802

Step 2: Filter the data

Other work does
this with a model-
based detector, but
that has risks

1. Url Filtering, removing obvious and NSFW websites, using both block-list as well as subwords detection
2. Trafilatura text extraction from Crawl's warc files
3. FastText LanguageFilter, removing any document with en language score lower than **0.65**
4. Quality filtering
 1. Gopher Repetition / Quality
 2. C4 Quality filters except terminal_punct rule
 3. FineWeb custom filters, consisting of heuristics for removing list-like documents, documents with repeated lines and documents with likely wrong line formatting.
5. MinHash deduplication with each crawl deduplicated individually (5-grams, 14x8 hash functions)
6. PII Formatting to anonymize email and public IP addresses

6



Data Sources

Annotation

Crowdsourcing

Workshop Preview



menti.com 7824 5802

Innovation in FineWeb - choosing filters by training

Vary the filters

Train models (1.8b parameters each!)

Measure performance

7



Data Sources

Annotation

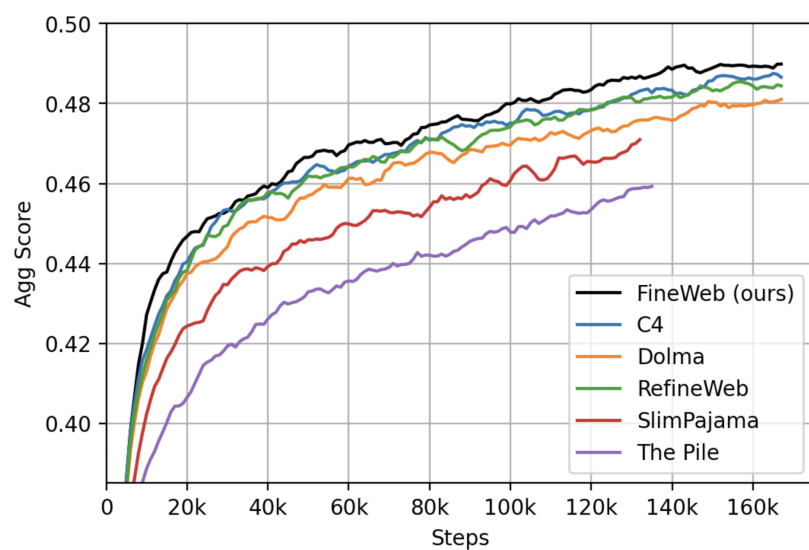
Crowdsourcing

Workshop Preview



menti.com 7824 5802

Innovation in FineWeb - choosing filters by training



8



Data Sources

Annotation

Crowdsourcing

Workshop Preview



menti.com 7824 5802

LLM developers seek more + better data

BERT - Books scraped from the internet (16 Gb)

RoBERTa - Add more data! (160 Gb)

'High quality':

- Wikipedia
- News
- Scientific papers

Vs.

Scale:

- The web

"Textbooks Are All You Need", 2023



Data Sources

Annotation

Crowdsourcing

Workshop Preview



menti.com 7824 5802

What about other tasks?

Named Entity Recognition

Coreference Resolution

Syntactic Parsing

Sentiment Analysis

Translation

....

Annotated data
isn't available for
free to download!
...mostly

Some translated data is available:

Canadian Hansards
European Parliament
United Nations



Data Sources

Annotation

Crowdsourcing

Workshop Preview



menti.com 7824 5802

Datasets can have massive impact

Penn Treebank - syntactic annotation, 11,000 citations

OntoNotes - coreference, 2,000 citations (across two papers)

SQuAD - question answering, 11,000 citations

Stanford Sentiment Treebank - 11,000 citations

For context, all of my research over the last 18 years has 3,800 citations



Data Sources

Annotation

Crowdsourcing

Workshop Preview



menti.com 7824 5802

My most cited paper is a dataset

☐	TITLE	CITED BY	YEAR
☐	An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction S Larson, A Mahendran, JJ Peper, C Clarke, A Lee, P Hill, JK Kummerfeld, ... EMNLP (Short Paper)	743	2019
☐	Improving Text-to-SQL Evaluation Methodology C Finegan-Dollak, JK Kummerfeld, L Zhang, K Ramanathan, ... ACL	381	2018
☐	Spatiotemporal Hierarchy of Relaxation Events, Dynamical Heterogeneities, and Structural Reorganization in a Supercooled Liquid R Candelier, A Widmer-Cooper, JK Kummerfeld, O Dauchot, G Biroli, ... Physical review letters 105 (13), 135702	221	2010
☐	A Mechanistic Understanding of Alignment Algorithms: A Case Study on DPO and Toxicity A Lee, X Bai, I Pres, M Wattenberg, JK Kummerfeld, R Mihalcea ICML	197	2024
☐	Factors Influencing the Surprising Instability of Word Embeddings L Wendlandt, JK Kummerfeld, R Mihalcea NAACL	178 *	2018
☐	A Large-Scale Corpus for Conversation Disentanglement JK Kummerfeld, SR Gouravajhala, J Peper, V Athreya, C Gunasekara, ... ACL	147 *	2019
☐	Tools for Automated Analysis of Cybercriminal Markets RS Portnoff, S Afroz, G Durrett, JK Kummerfeld, T Berg-Kirkpatrick, ... WWW	129	2017
☐	Parser Showdown at the Wall Street Corral: An Empirical Investigation of Error Types in Parser Output JK Kummerfeld, D Hall, JR Curran, D Klein EMNLP	118	2012
☐	Overview of the seventh dialog system technology challenge: Dstc7 LF D'Haro, K Yoshino, C Hori, TK Marks, L Polymenakos, JK Kummerfeld, ... Computer Speech & Language 62, 101068	108 *	2020
☐	Supporting Sensemaking of Large Language Model Outputs at Scale JK Kummerfeld, SR Gouravajhala, J Peper, V Athreya, C Gunasekara, ... ACL	92	2024



Data Sources

Annotation

Crowdsourcing

Workshop Preview



menti.com 7824 5802

Evaluate on multiple tasks at once - avoid overfitting to one

GLUE - General Language Understanding Evaluation

The Corpus of Linguistic Acceptability

The Stanford Sentiment Treebank

Microsoft Research Paraphrase Corpus

Semantic Textual Similarity Benchmark

Quora Question Pairs

MultiNLI

Question NLI

Recognizing Textual Entailment

Winograd NLI

One gripe - people cite a benchmark suite, but not the datasets that went into it



Data Sources

Annotation

Crowdsourcing

Workshop Preview



menti.com 7824 5802

Other risks - Shortcuts / Validity

Is there a shortcut - a way to get the right answer without doing the task?

"Clever Hans"



"If the eighth day of the month comes on a Tuesday, what is the date of the following Friday?"

Hans would answer by tapping his hoof eleven times

But only if it could see the questioner and the questioner knew the answer

Wikipedia



Data Sources

Annotation

Crowdsourcing

Workshop Preview



menti.com 7824 5802

Other risks - Shortcuts / Validity

What does Macduff do to Macbeth?

What becomes of Macbeth?

What violent act does Macduff perform upon Macbeth?

The same answer in every case, but
some are easier than others

From the UT Austin NLP class



Data Sources

Annotation

Crowdsourcing

Workshop Preview



menti.com 7824 5802

Other risks - Shortcuts / Validity

“On **October 19, 1512**, **Luther** was awarded his doctorate of theology and, on **October 21, 1512**, was received into the senate of the theological faculty of the **University of Wittenberg**. He spent the rest of his career in this position at the **University of Wittenberg**.”

Where ...

Who ...

When ...

Only one option in many cases!

From the UT Austin NLP class



Data Sources

Annotation
Crowdsourcing
Workshop Preview



menti.com 7824 5802

Other risks - Shortcuts / Validity

One hint is whether a model can do well without the question and/or context?

For example, for the Stanford Natural Language Inference task, looking just at the hypothesis:

Text: "A black race car starts up in front of a crowd of people."

Hypothesis: "A man is driving down a lonely road."

Correct answer: Contradiction

	Hyp-only model	Majority class	
SNLI	69.17	33.82	+35
MNLI-1	55.52	35.45	+20
MNLI-2	55.18	35.22	+19

From the UT Austin NLP class



Data Sources

Annotation
Crowdsourcing
Workshop Preview



menti.com 7824 5802

Other risks - Shortcuts / Validity

What to do?

- Get text from real world use
- Make hard cases
 - Wrong answers that have many matching words with the input or true answer
 - Make minimal edits to the true answer that make it wrong

From the UT Austin NLP class



Data Sources

Annotation

Crowdsourcing

Workshop Preview



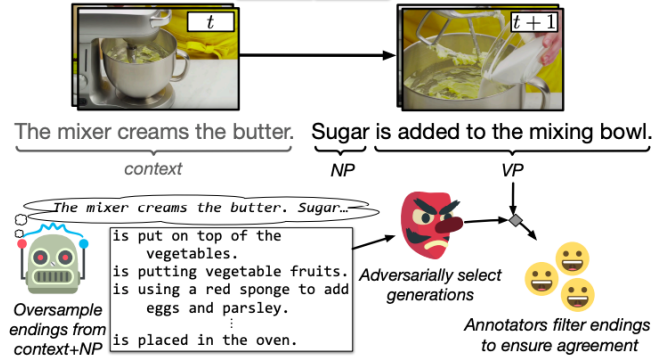
menti.com 7824 5802

Other risks - Does our task have statistical power?

A girl is going across a set of monkey bars. She

- a) jumps up across the monkey bars.
- b) struggles onto the monkey bars to grab her head.
- c) **gets to the end and stands on a wooden plank.**
- d) jumps up and does a back flip.

Using video captions from **ACTIVITYNET** **LSMDC** (the videos are never used)



From the UT Austin NLP class



Data Sources

Annotation

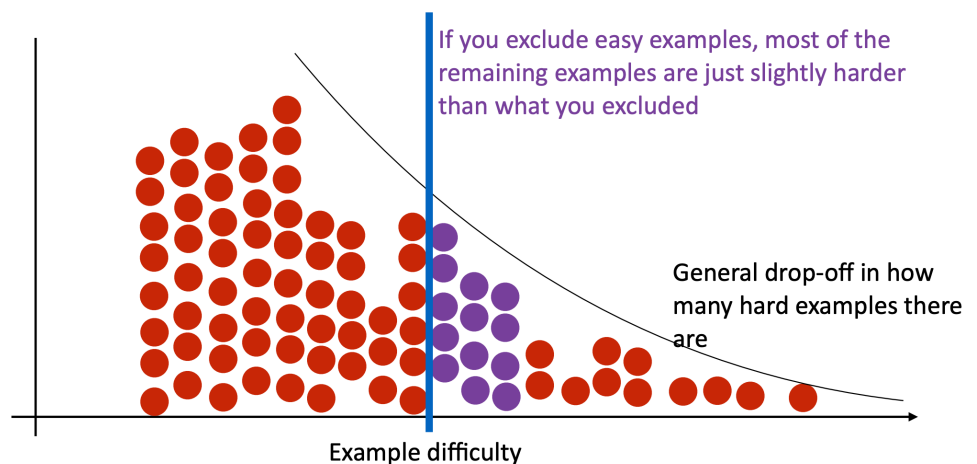
Crowdsourcing

Workshop Preview



menti.com 7824 5802

Other risks - Does our task have statistical power?



Models very quickly achieved human-level accuracy on the task

From the UT Austin NLP class



Data Sources

Annotation

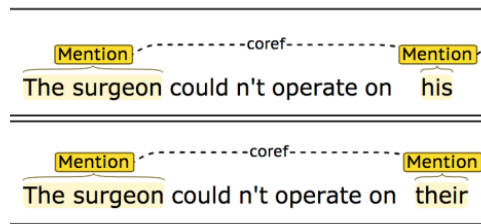
Crowdsourcing

Workshop Preview



menti.com 7824 5802

Other risks - Social bias



Data Sources

Annotation

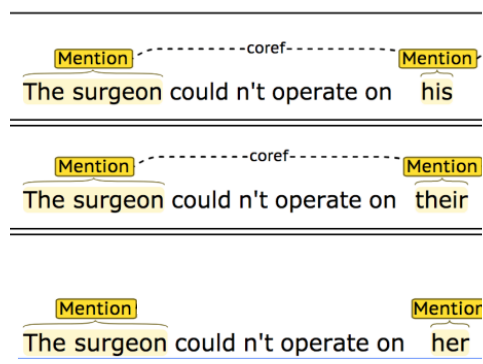
Crowdsourcing

Workshop Preview



menti.com 7824 5802

Other risks - Social bias





Data Sources

Annotation

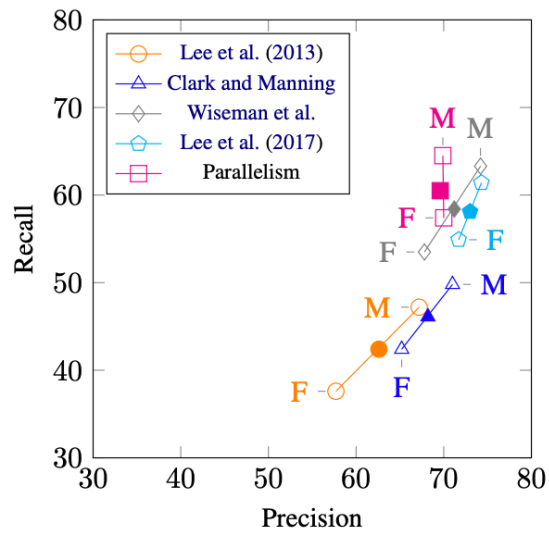
Crowdsourcing

Workshop Preview



menti.com 7824 5802

Other risks - Social bias



Webster et al., (2018)



Data Sources

Annotation

Crowdsourcing

Workshop Preview



menti.com 7824 5802

Other risks - Social bias

OntoNotes

- 80% of gendered pronouns are male
- Male mentions are twice as likely to contain a job title

Zhao et al. (2018)



Data Sources

Annotation

Crowdsourcing

Workshop Preview



menti.com 7824 5802

Other risks - Is this a task we want to create?

One research group introduced this task:

Given a photo of a face,
predict if the person is a criminal !!!

1. Major ethical concerns here
2. This task doesn't make sense - we have no reason to believe it is possible to do better than chance



Data Sources

Annotation

Crowdsourcing

Workshop Preview



menti.com 7824 5802

What makes a good dataset?

- Validity
Models that perform well on the task will perform well in the real world
- Reliability
Data is accurately labeled
- Statistical power
Enough examples, including enough hard ones, to provide a meaningful comparison
- Testable
- Publicly licensed
- Avoids social bias



Data Sources

Annotation

Crowdsourcing

Workshop Preview



menti.com 7824 5802

Recap: Data Sources

Data collection: This involves finding and filtering data in various ways. Choices can have a big impact on model performance and style.

Data annotation: Once we have raw data we need to assign labels to it. Even if we are using a few-shot approach, we want some labeled data for evaluation. This can be slow, expensive, and difficult. More in the next section!

Key goals in dataset creation:

- Validity
- Reliability
- Statistical power
- Testable
- Publicly licensed
- Avoids social bias

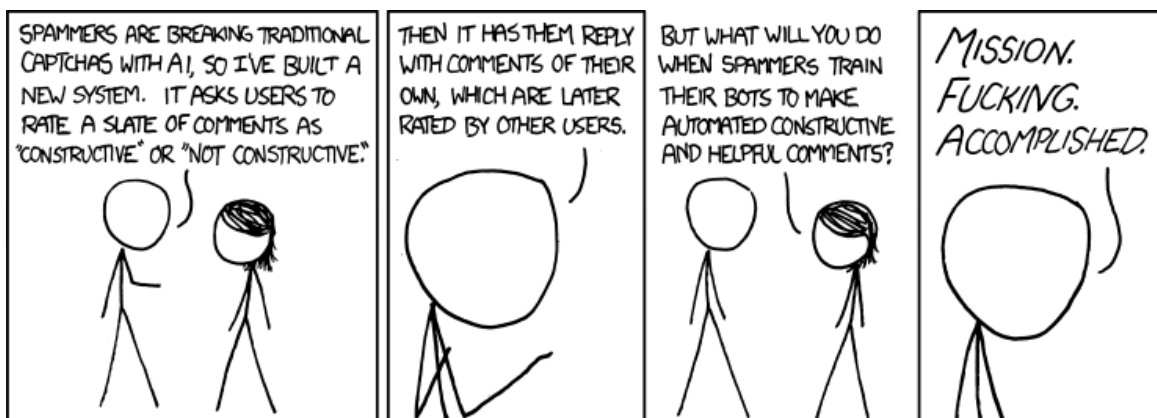
27

3 minute Break - stretch and visit Menti

menti.com
7824 5802



Constructive



[“And what about all the people who won’t be able to join the community because they’re terrible at making helpful and constructive co-- ... oh.”]

Source: <https://xkcd.com/810/>

28



Data Sources

Annotation

Crowdsourcing

Workshop Preview



menti.com 7824 5802

Annotation

29



Data Sources

Annotation

Crowdsourcing

Workshop Preview



menti.com 7824 5802

What to annotate?

Real world data - may only rarely contain the phenomena we care about or may have biases

Created data - may have unintended patterns or biases

Spider (Yu et al., 2018)

NL: *List the emails of the professionals who live in the state of Hawaii or the state of Wisconsin.*

SQL: `select email_address from
professionals where state =
'Hawaii' or state = 'Wisconsin';`

Suhr et al. (2020)

30



Data Sources
Annotation
Crowdsourcing
Workshop Preview



menti.com 7824 5802

Data preprocessing

All the steps we saw for FineWeb

31



Data Sources
Annotation
Crowdsourcing
Workshop Preview



menti.com 7824 5802

Tools for annotation

Benefits:

- Consistency
- Support

Annotate
from
scratch

Chocolate has been eaten for

Edit auto-
generated
labels

NN VBZ VBN VBN IN
Chocolate has been eaten for

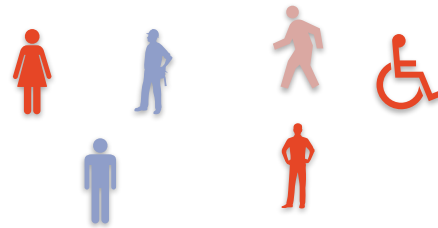
Saves time, but could increase error if
annotators start accepting everything

32



Annotation guidelines

Risk? Each annotator handles cases differently



Instructions

- Simple
- Clear
- Concise
- Consistent
- Generalisable

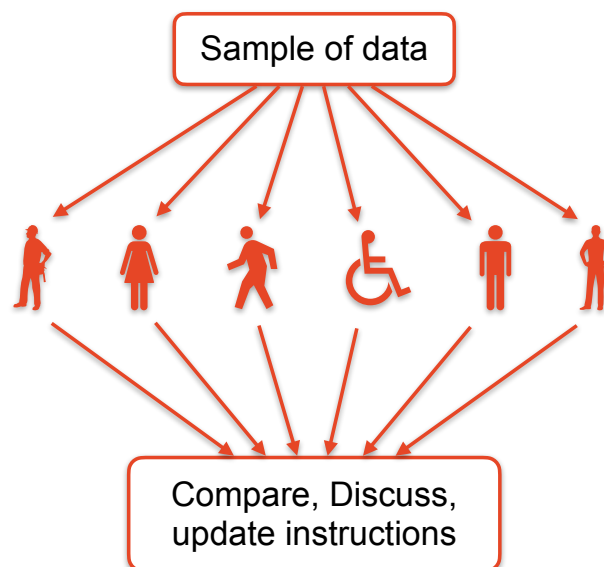
Examples

- Common cases
- Rare cases

Refine over time

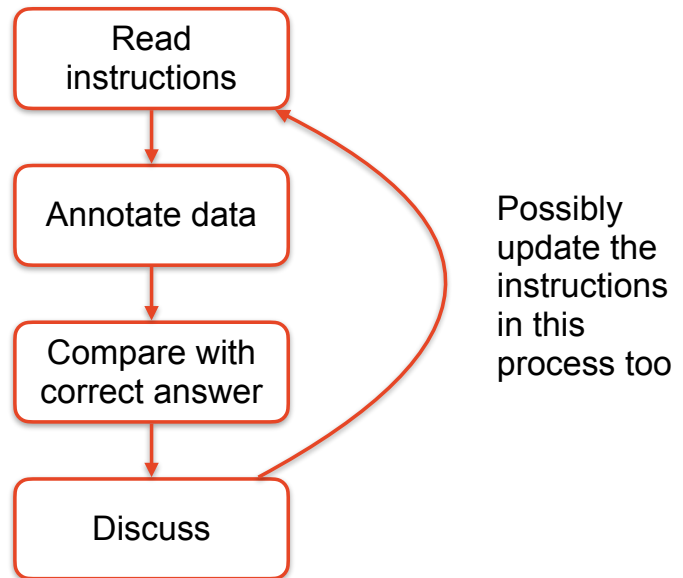


Pilot annotation





Train annotators

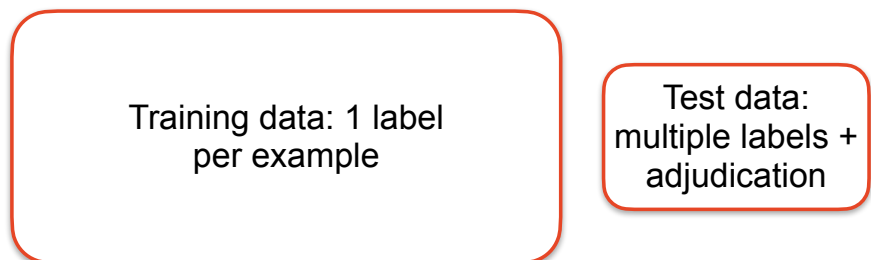


35



Full annotation

How many people annotate each example? Tradeoff between the amount of data and annotation quality.



36



Data Sources

Annotation

Crowdsourcing
Workshop Preview



menti.com 7824 5802

Full annotation

Attention checks - insert out of place content to check the person is focused

“... Ignore the rest of this paragraph, don't annotate anything ...”

Consistency checks - have some fraction of examples be done by other people and raise an error if there is too much inconsistency



Data Sources

Annotation

Crowdsourcing
Workshop Preview



menti.com 7824 5802

Adjudication

Filter to cases where there is a disagreement

Look at them together and determine a label

```
1. um2 - researchgate (Python)
[19:58] <wafflejack> corba, well depends on where on the local disk you're
trying to save, if it's your home folder should be fine
[19:58] <corba> MMM, i tried with NTP
[19:58] <wafflejack> corba, someone from 2010 https://ubuntuforums.org/st
[19:59] <MMM> oh wait... heard you were trying to get to windows right?
[19:59] <corba> wafflejack, nope, its a separate disk
[19:59] <corba> wafflejack, ill look into that thanks
[19:59] <MMM> did you write your share into /etc/samba/smb.conf?
[19:59] <corba> MMM, i didn't write the share, shouldnt i only have to do
that if i was trying to share from xubuntu?
[20:01] <corba> i also tried with gigolo
[20:01] <MMM> oh jeeze: I keep crossing what you got it mixed up again and was right the first
Windows from xubuntu... only xubuntu from wind
[20:02] <corba> no prob :)
=== nicomach2s is now known as nicomachus
[20:04] <corba> ive done it a thousand times in cinnamon but this is an
old computer and i have to use xfce or lxde, i dont know if i can run
mate...
[20:04] <wafflejack> shared folders, might be for adding
[20:05] <corba> not c... g it form the
```

Blue & Cyan: Options

Green: Current line

Red: Other lines with disagreements



Measuring agreement

We have some labels!

But were we consistent?

Note:
consistent != good
consistent != useful



Measuring agreement - Cohen's Kappa

		Annotator 1	
		Label A	Label B
Annotator 2	Label A	7	4
	Label B	8	81

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$
$$= \frac{0.88 - p_e}{1 - p_e}$$

$$p_o = P(A1 = A2 = A) + P(A1 = A2 = B)$$
$$= 0.07 + 0.81$$
$$= 0.88$$



Data Sources
Annotation
Crowdsourcing
Workshop Preview



menti.com 7824 5802

Measuring agreement - Cohen's Kappa

		Annotator 1	
		Label A	Label B
Annotator 2	Label A	7	4
	Label B	8	81

$$\begin{aligned}\kappa &= \frac{p_o - p_e}{1 - p_e} \\ &= \frac{0.88 - p_e}{1 - p_e} \\ &= \frac{0.88 - 0.773}{1 - 0.773} \\ &= 0.471\end{aligned}$$

$$\begin{aligned}p_e &= P(A1 = A)P(A2 = A) + P(A1 = B)P(A2 = B) \\ &= 0.15 * 0.11 + 0.85 * 0.89 \\ &= 0.773\end{aligned}$$

Berkeley NLP course

41



Data Sources
Annotation
Crowdsourcing
Workshop Preview



menti.com 7824 5802

Measuring agreement - Cohen's Kappa

		Annotator 1	
		Label A	Label B
Annotator 2	Label A	1	1
	Label B	1	97

$$\begin{aligned}\kappa &= \frac{p_o - p_e}{1 - p_e} \\ &= \frac{0.98 - p_e}{1 - p_e}\end{aligned}$$

$$\begin{aligned}p_o &= P(A1 = A2 = A) + P(A1 = A2 = B) \\ &= 0.01 + 0.97 \\ &= 0.98\end{aligned}$$

42



Measuring agreement - Cohen's Kappa

		Annotator 1	
		Label A	Label B
Annotator 2	Label A	1	1
	Label B	1	97

$$\begin{aligned}\kappa &= \frac{p_o - p_e}{1 - p_e} \\ &= \frac{0.98 - 0.961}{1 - 0.961} \\ &= 0.487\end{aligned}$$

$$\begin{aligned}p_e &= P(A1 = A)P(A2 = A) + P(A1 = B)P(A2 = B) \\ &= 0.02 * 0.02 + 0.98 * 0.98 \\ &= 0.961\end{aligned}$$



Measuring agreement - Cohen's Kappa

		Annotator 1	
		Label A	Label B
Annotator 2	Label A	49	1
	Label B	1	49

$$\begin{aligned}\kappa &= \frac{p_o - p_e}{1 - p_e} \\ &= \frac{0.98 - 0.5}{1 - 0.5} \\ &= 0.96\end{aligned}$$

$$\begin{aligned}p_o &= P(A1 = A2 = A) + P(A1 = A2 = B) \\ &= 0.49 + 0.49 \\ &= 0.98\end{aligned}$$

$$\begin{aligned}p_e &= P(A1 = A)P(A2 = A) + P(A1 = B)P(A2 = B) \\ &= 0.5 * 0.5 + 0.5 * 0.5 \\ &= 0.5\end{aligned}$$



Measuring agreement - Cohen's Kappa

Annotator 2	Annotator 1	
	Label A	Label B
	Label A	Label B
	25	25
	25	25

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

$$= \frac{0.5 - 0.5}{1 - 0.5}$$

$$= 0$$

$$p_o = P(A1 = A2 = A) + P(A1 = A2 = B)$$

$$= 0.25 + 0.25$$

$$= 0.5$$

$$p_e = P(A1 = A)P(A2 = A) + P(A1 = B)P(A2 = B)$$

$$= 0.5 * 0.5 + 0.5 * 0.5$$

$$= 0.5$$



Measuring agreement - Cohen's Kappa

What is a good value? One very rough guide:

0.80-1.00	Very good agreement
0.60-0.80	Good agreement
0.40-0.60	Moderate agreement
0.20-0.40	Fair agreement
< 0.20	Poor agreement



Measuring agreement - Fleiss' Kappa

Similar idea, but for more than 2 annotators

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

Number of annotators who
labeled item i with label j

n_{ij}

Average
over items

Overlap in
labels

Overall agreement

$$P_o = \frac{1}{N} \sum_{i=1}^N \left(\frac{1}{n(n-1)} \sum_{j=1}^K n_{ij}(n_{ij} - 1) \right)$$

Average
over classes

Chance agreement

$$P_e = \sum_{j=1}^K \left(\frac{1}{N} \sum_{i=1}^N n_{ij} \right)^2$$

Squared proportion
assigned to this class



Measuring agreement - Krippendorff's Alpha

More general approach that supports other data types
(e.g., ranks, ratios, intervals)



Data Sources
Annotation
Crowdsourcing
Workshop Preview



menti.com 7824 5802

What do you do if agreement is low?

Investigate & Iterate!

Unclear
guidelines?

Careless
annotators?

Ambiguous
cases?

Subjective
cases?



Data Sources
Annotation
Crowdsourcing
Workshop Preview



menti.com 7824 5802

Releasing data - Hosting

GitHub

HuggingFace Datasets

Your own website



Data Sources

Annotation

Crowdsourcing

Workshop Preview



menti.com 7824 5802

Releasing data - Licenses

Do you have consent to release the data?

Have you anonymised the data where appropriate?

Do you have the right to share the data?

Under what license?

51



Data Sources

Annotation

Crowdsourcing

Workshop Preview



menti.com 7824 5802

Annotation is human subjects research

Universities - Seek ethics approval

Companies - See guidelines

52



Data Sources

Annotation

Crowdsourcing

Workshop Preview



menti.com 7824 5802

Recap: Annotation

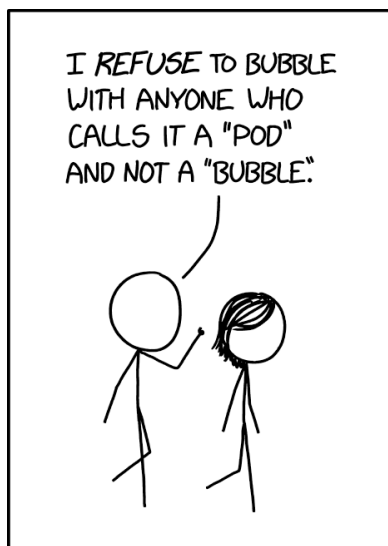
Key steps in annotation:

- Collect data (previous section of this lecture)
- Develop annotation guidelines
- Train annotators
- Label data
- Check disagreements
- Choose license and release data

Measuring disagreement: A range of metrics exist to understand how much annotators agree. For classification tasks, Cohen's Kappa is simple and widely used. However, a 'good' value is not well defined. For non-classification tasks, agreement is harder to measure and we generally don't have statistically well-motivated methods.

3 minute Break - stretch and visit Menti

menti.com
7824 5802



THIS IS PROBABLY MY OPINION THAT WOULD HAVE SOUNDED THE MOST INCOHERENT TO ME A YEAR AGO.

Pods vs Bubbles

[“Canada’s travel restrictions on the US are 99% about keeping out COVID and 1% about keeping out people who say ‘pod.’”]

Source: <https://xkcd.com/2339/>



Data Sources
Annotation
Crowdsourcing
Workshop Preview



menti.com 7824 5802

Crowdsourcing

55



Data Sources
Annotation
Crowdsourcing
Workshop Preview



menti.com 7824 5802



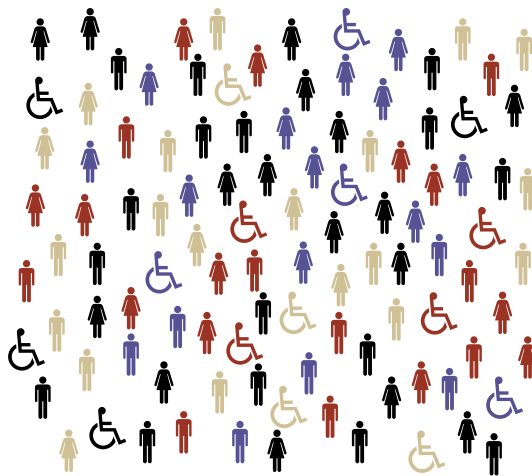
56



Data Sources
Annotation
Crowdsourcing
Workshop Preview



menti.com 7824 5802



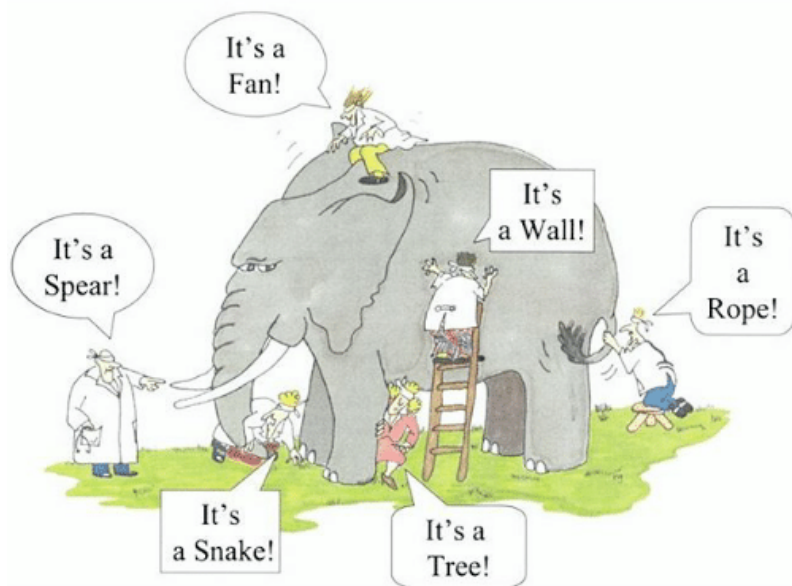
57



Data Sources
Annotation
Crowdsourcing
Workshop Preview



menti.com 7824 5802



58



COMP 4446 / 5046
Lecture 8

Data Sources
Annotation
Crowdsourcing
Workshop Preview



menti.com 7824 5802

800 Guesses at the 1906 country fair in Plymouth

Median: 1207 lb

Mean: 1197 lb

True: 1198 lb



PEC Large Livestock Scales, Amazon

59



COMP 4446 / 5046
Lecture 8

Data Sources
Annotation
Crowdsourcing
Workshop Preview



menti.com 7824 5802

Microwork

Paid, small tasks

Choose the best category for this image [View Instructions ↓](#)

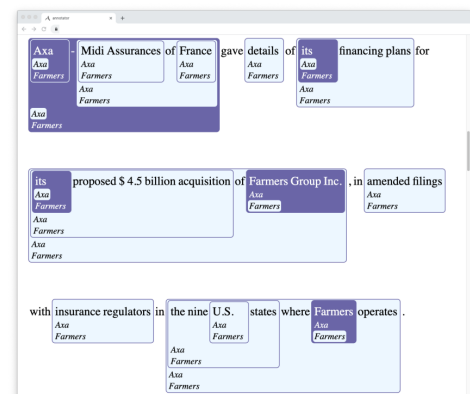


- ☐ kitchen
- ☐ living
- ☐ bath
- ☐ bed
- ☐ outside

You must ACCEPT the HIT before you can submit the results.

Select the room location in home for this picture. Seating areas outside are outside not living. Offices or dens are living not bedrooms. Bedrooms should contain a bed in the picture.

amazon
mechanical turk



60



COMP 4446 / 5046
Lecture 8

Data Sources
Annotation
Crowdsourcing
Workshop Preview



menti.com 7824 5802

Games with a Purpose

Unpaid, make task fun



61



COMP 4446 / 5046
Lecture 8

Data Sources
Annotation
Crowdsourcing
Workshop Preview



menti.com 7824 5802

Games with a Purpose

Competition with prize money



Sherlink Holmes went to the shop.
He couldn't believe it was closed.

If you think that the word in orange, **It**, refers to the same object as the phrase in blue, **the shop**, you should say that you agree with this decision.

[Further instructions](#)



62



COMP 4446 / 5046
Lecture 8

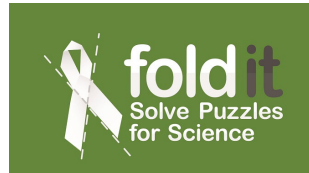
Data Sources
Annotation
Crowdsourcing
Workshop Preview



menti.com 7824 5802

Citizen Science / Collaboration

Unpaid, maybe fun? Satisfying



63



COMP 4446 / 5046
Lecture 8

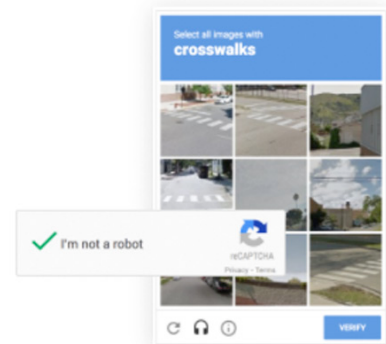
Data Sources
Annotation
Crowdsourcing
Workshop Preview



menti.com 7824 5802

Implicit Work

Unpaid, not fun,
provides a benefit



64



COMP 4446 / 5046
Lecture 8

Data Sources
Annotation
Crowdsourcing
Workshop Preview



menti.com 7824 5802

Microwork

Paid, small tasks

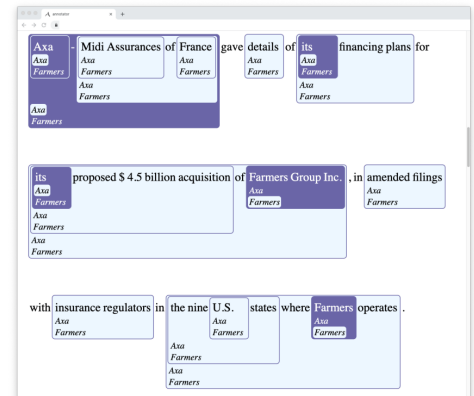
Choose the best category for this image [View Instructions](#)



Select the room location in home for this picture. Seating areas outside are outside not living. Offices or dens are living not bedrooms. Bedrooms should contain a bed in the picture.

- ☐ kitchen
- ☐ living
- ☐ bath
- ☐ bed
- ☐ outside

You must ACCEPT the HIT before you can submit the results.



amazon
mechanical turk



65



COMP 4446 / 5046
Lecture 8

Data Sources
Annotation
Crowdsourcing
Workshop Preview



menti.com 7824 5802

Microwork can have ethical issues



66

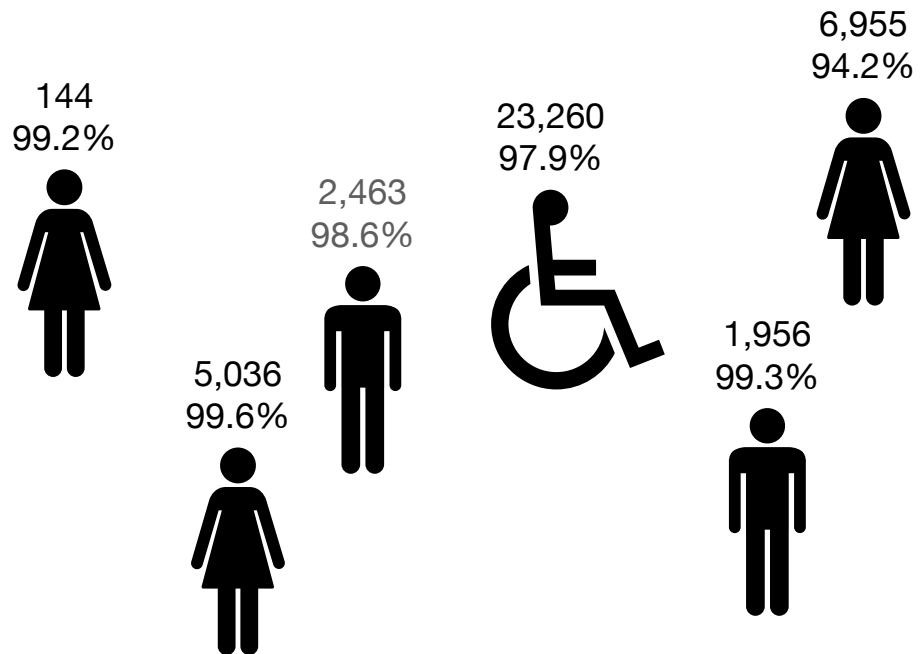


Data Sources
Annotation
Crowdsourcing
Workshop Preview



menti.com 7824 5802

What are qualifications?

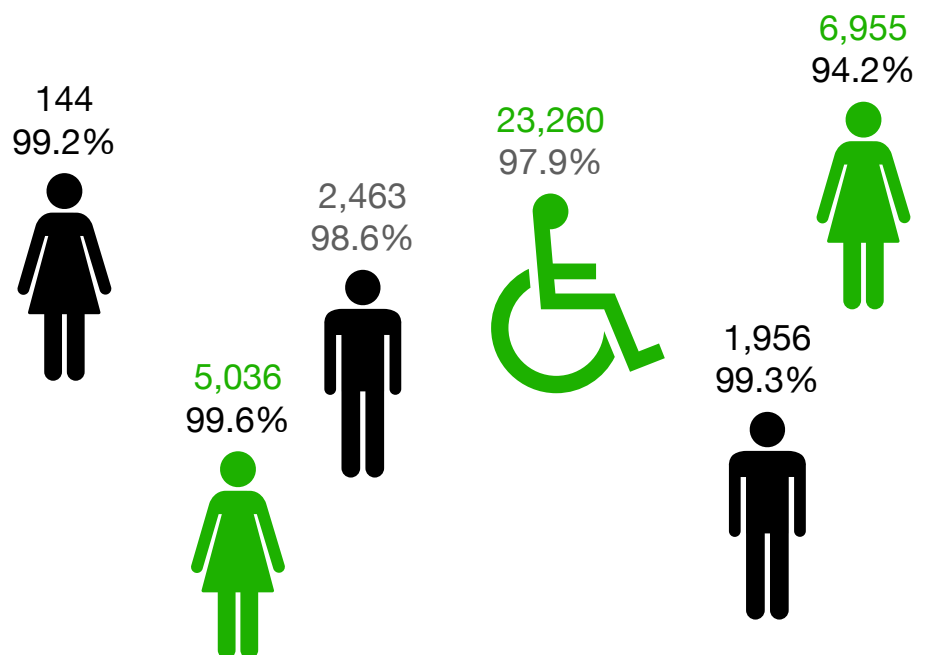


Data Sources
Annotation
Crowdsourcing
Workshop Preview



menti.com 7824 5802

What are qualifications?





Data Sources
Annotation
Crowdsourcing
Workshop Preview



menti.com 7824 5802

What are qualifications?

144
99.2%



5,036
99.6%



2,463
98.6%



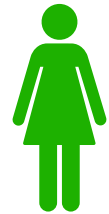
23,260
97.9%



1,956
99.3%



6,955
94.2%



Data Sources
Annotation
Crowdsourcing
Workshop Preview



menti.com 7824 5802

Conventions for qualifications



Conventions for qualifications

	100	1,000	5,000	10,000
amazon				
Blogs				
Papers and Courses				



Conventions for qualifications

	100	1,000	5,000	10,000
amazon				
Blogs				
Papers and Courses	*			



Conventions for qualifications

	100	1 000	5 000	10,000
amazon				
Blind				
Papers and Courses	*			

For your first 1000 HITs you may want to concentrate on approval milestones rather than \$\$\$... most of the better-paying requesters require 1000/5000/10000+ approved HITs



Recap: Crowdsourcing

What is crowdsourcing? A range of methods fit under this heading. In all cases, they involve a large group of people contributing to the creation of a resource.

The most well known and widely used is 'microwork', where people are paid to do small tasks online. Other variants have had incredible success in specific areas (e.g., creating Wikipedia) but are harder to create.



Data Sources
Annotation
Crowdsourcing
Workshop Preview



menti.com 7824 5802

Workshop Preview

75



Data Sources
Annotation
Crowdsourcing
Workshop Preview



menti.com 7824 5802

Pre-work: Set up two annotation tools on your computer

In-class: Annotate a small amount of data with each one

Very valuable for
assignment 4!

76



COMP 4446 / 5046
Lecture 8

Data Sources
Annotation
Crowdsourcing
Workshop Preview



menti.com 7824 5802

Muddy Card

Open now, closes Thursday 11:59pm

<https://canvas.sydney.edu.au/courses/72300/quizzes/332405>



If you do not wish to participate in the study,
use the Microsoft form instead
Go to Ed → Lessons → Muddy Cards → Form