Applied Natural Language Processing

Info 256

Lecture 7: Annotation (Sept. 18, 2023)

David Bamman, UC Berkeley

Modern NLP is driven by annotated data

- Penn Treebank (1993; 1995;1999); morphosyntactic annotations of WSJ
- OntoNotes (2007–2013); syntax, predicate-argument structure, word sense, coreference
- FrameNet (1998–): frame-semantic lexica/annotations
- MPQA (2005): opinion/sentiment
- SQuAD (2016): annotated questions + spans of answers in Wikipedia

Modern NLP is driven by annotated data

- In most cases, the data we have is the product of human judgments.
    - What's the correct part of speech tag?
    - Syntactic structure?
    - Sentiment?

Respect

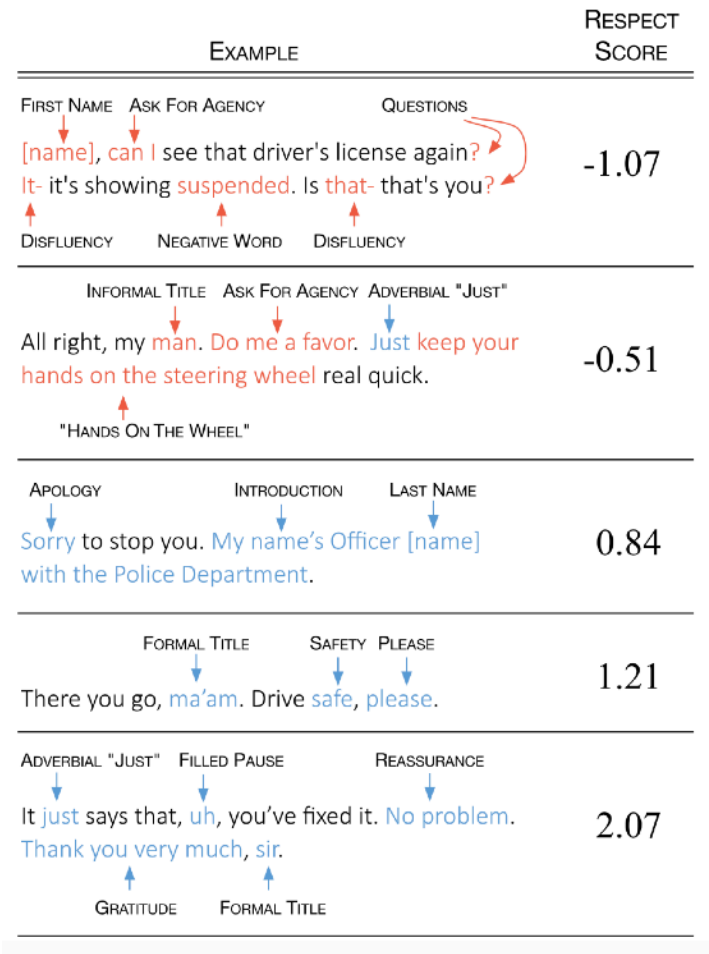
Input: transcripts of 981 OPD traffic stops (everyday interactions)

Output: measure of “respect” directed from officer to driver

Voigt et al. 2017, “Language from police body camera footage shows racial disparities in officer respect”

Respect

- Present one dialogue turn (police/driver) to be rated by people for respect (4-point Likert scale). High IAA.



Dogmatism

Fast and Horvitz (2016), "Identifying Dogmatism in Social Media: Signals and Models"

Given a comment, imagine you hold a well-informed, different opinion from the commenter in question. We'd like you to tell us how likely that commenter would be to engage you in a constructive conversation about your disagreement, where you each are able to explore the other's beliefs. The options are:

(5): It's unlikely you'll be able to engage in any substantive conversation. When you respectfully express your disagreement, they are likely to ignore you or insult you or otherwise lower the level of discourse.

(4): They are deeply rooted in their opinion, but you are able to exchange your views without the conversation degenerating too much.

(3): It's not likely you'll be able to change their mind, but you're easily able to talk and understand each other's point of view.

(2): They may have a clear opinion about the subject, but would likely be open to discussing alternative viewpoints.

(1): They are not set in their opinion, and it's possible you might change their mind. If the comment does not convey an opinion of any kind, you may also select this option.

Dogmatism

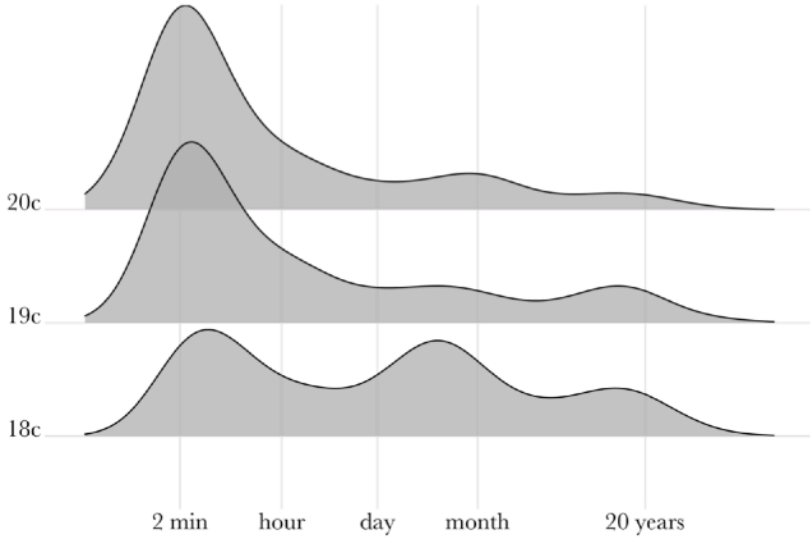
Fast and Horvitz (2016), "Identifying Dogmatism in Social Media: Signals and Models"

Highest	Score	Lowest	Score
cringepics	0.553	photography	0.399
DebateAChristian	0.551	DIY	0.399
DebateReligion	0.540	homebrewing	0.401
politics	0.536	cigars	0.402
ukpolitics	0.533	wicked_edge	0.404
atheism	0.529	guitar	0.406
lgbt	0.527	gamedeals	0.406
TumblrInAction	0.524	buildapc	0.407
islam	0.523	techsupport	0.410
SubredditDrama	0.520	travel	0.410

Table 3: Subreddits with the highest and lowest dogmatism scores. Politics and religion are common themes among the most dogmatic subreddits, while hobbies (e.g., photography, homebrewing, buildapc) show the least dogmatism.

Literary Time

- How many minutes pass in a 250-word passage of fiction?



Underwood 2018, “Why Literary Time is Measured in Minutes”

“Tom!” No answer. “TOM!” No answer. “What’s gone with that boy, I wonder? You TOM!” No answer.

The old lady pulled her spectacles down and looked over them about the room; then she put them up and looked out under them. She seldom or never looked _through_ them for so small a thing as a boy; they were her state pair, the pride of her heart, and were built for “style,” not service—she could have seen through a pair of stove-lids just as well. She looked perplexed for a moment, and then said, not fiercely, but still loud enough for the furniture to hear:

“Well, I lay if I get hold of you I’ll—”

She did not finish, for by this time she was bending down and punching under the bed with the broom, and so she needed breath to punctuate the punches with. She resurrected nothing but the cat.

“I never did see the beat of that boy!”

She went to the open door and stood in it and looked out among the tomato vines and “jimpson” weeds that constituted the garden. No Tom. So she lifted up her voice at an angle calculated for distance and shouted:

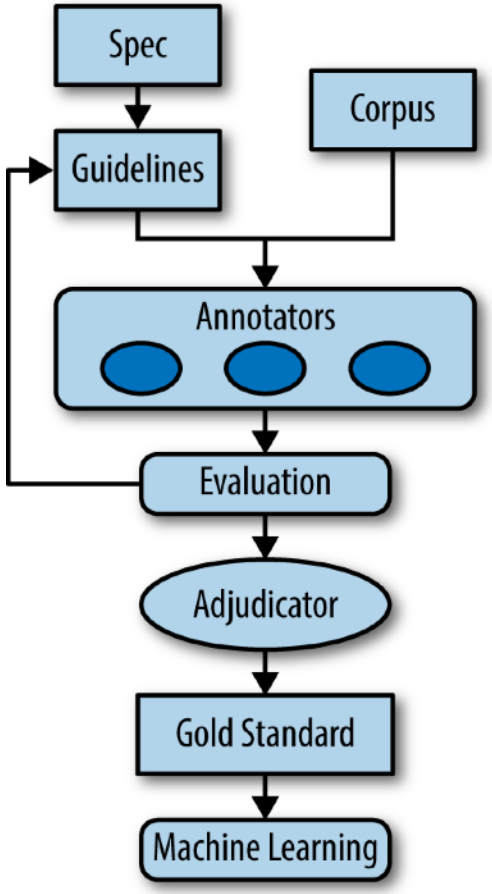
“Y-o-u-u TOM!”

There was a slight noise behind her and she turned just in time to seize a small boy by the slack of his roundabout and arrest his flight.

“There! I might ’a’ thought of that closet. What you been doing in there?”

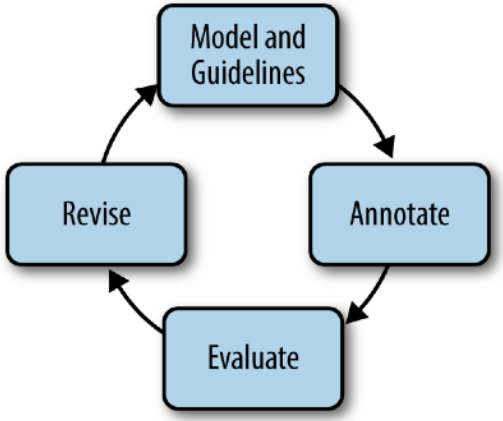

a month
a week
a day
6 hours
an hour
15 min
2 min
1719
1800
1900
2000
Robinson Crusoe
The Monk
East Lynne
A Portrait of the Artist
Mrs Dalloway

Annotation pipeline



Pustejovsky and Stubbs (2012),
Natural Language Annotation for Machine Learning

Annotation pipeline



Pustejovsky and Stubbs (2012),
Natural Language Annotation for Machine Learning

Annotation guidelines

- Our goal: given the constraints of our problem, how can we formalize our description of the annotation process to encourage multiple annotators to provide the same judgment?

Annotation guidelines

- What is the goal of the project?
- What is each tag called and how is it used? (Be specific: provide examples, and discuss gray areas.)
- What parts of the text do you want annotated, and what should be left alone?
- How will the annotation be created? (For example, explain which tags or documents to annotate first, how to use the annotation tools, etc.)

Pustejovsky and Stubbs (2012), Natural Language Annotation for Machine Learning

brat rapid annotation tool

online environment for collaborative text annotation

Learn more:

- What is it?
- What can you do with it?
- What does it do?
- What do I need to run it?

Create your own local brat installation:

Download v1.3

Manage your own annotation effort

Easy to set up: installation instructions

Instructions for upgrading to v1.3 (Crunchy Frog)

Open source (MIT License)

Current version: v1.3 Crunchy Frog (2012-11-08).

https://brat.nlplab.org/

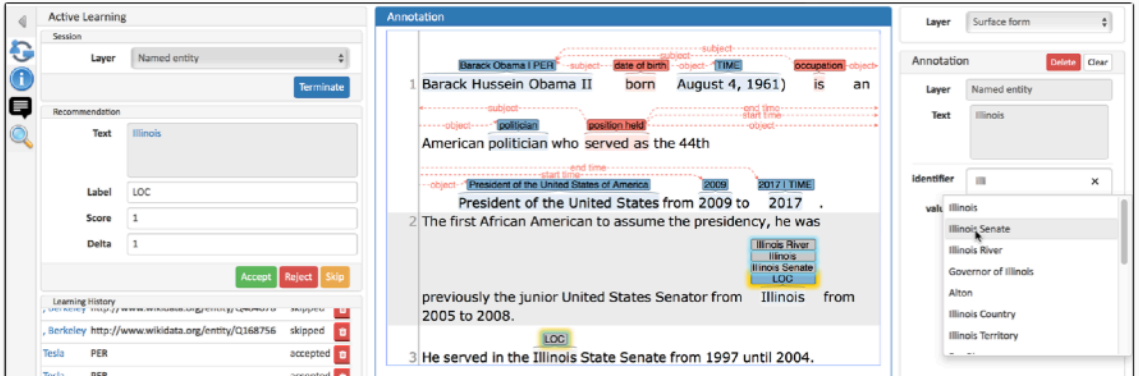
INCEpTION - Welcome

https://inception-project.github.io/

A semantic annotation platform offering intelligent assistance and knowledge management

The annotation of specific semantic phenomena often require compiling task-specific corpora and creating or extending task-specific knowledge bases. Presently, researchers require a broad range of skills and tools to address such semantic annotation tasks.

In the recently funded INCEpTION project, UKP Lab at TU Darmstadt aims towards building an annotation platform that incorporates all the related tasks into a joint web-based platform.



Download
INCEpTION 29.1
(Released on 2023-09-12)

Try
INCEpTION
online

Subjectivity Annotations

File Edit View Insert Format Data Tools Extensions Help

Share

Z53

	A	B
1	Label	Text
2		My favorite food is ice cream.
3		John said "My favorite food is ice cream."
4		Nancy Pelosi is the speaker of the house.
5		The sky is blue.
6		The earth is flat.
7		The earth is round.
8		Joe Biden won the 2020 US presidential election.
9		Donald Trump won the 2020 US presidential election.
10		Arrival is objectively the best movie.
11		Gods exist.
12		Gods do not exist.
13		I saw this movie last month at a free sneak preview and I walked out.
14		This documentary provides the viewers with unique footage about the 2003 coup in Venezuela.
15		I saw this film on its release, and have watched it 3 or 4 more times, including last week.
16		I was really disappointed in this car.
17		If you need cheering up on a cold weekday evening, this is the film for you.
18		The film only consists of a collection of scenes depicting the daily life of the family, such as swimming, taking piano lessons or cooking eggs.
19		For those of you unfamiliar with Jimmy Stewart , this is one of his "lesser" films from later in his career.
20		I read somewhere that when Kay Francis refused to take a cut in pay, Warner Bros. retaliated by casting her in inferior projects for the remainder of her contract.
21		Condensation trails left by airplanes purposely contain chemical agents put there to harm the general public.
22		I believe condensation trails left by airplanes purposely contain chemical agents put there to harm the general public.
23		I'm pretty sure condensation trails left by airplanes purposely contain chemical agents put there to harm the general public.

Why not do it yourself?

- Expensive/time-consuming
- Multiple people provide a measure of consistency: is the task well enough defined?
- Low agreement = not enough training, guidelines not well enough defined, task is bad

Adjudication

- Adjudication is the process of deciding on a single annotation for a piece of text, using information about the independent annotations.
- Can be as time-consuming (or more so) as a primary annotation.
- Does not need to be identical with a primary annotation (both annotators can be wrong by chance)

Interannotator agreement

annotator A

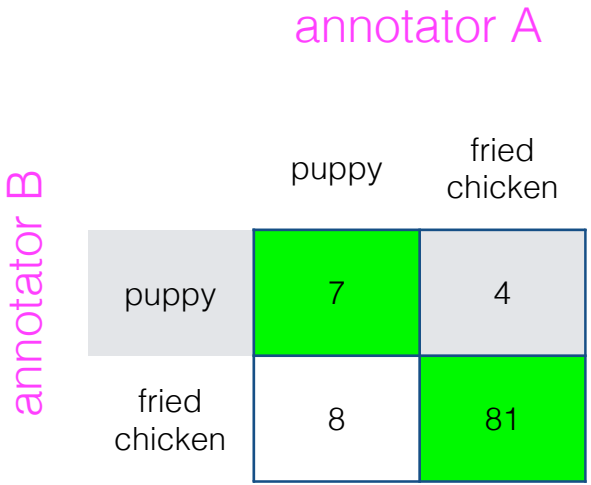
annotator B

	puppy	fried chicken
puppy	6	3
fried chicken	2	5

observed agreement = 11/16 = 68.75%

Cohen's kappa

- If classes are imbalanced, we can get high inter annotator agreement simply by chance

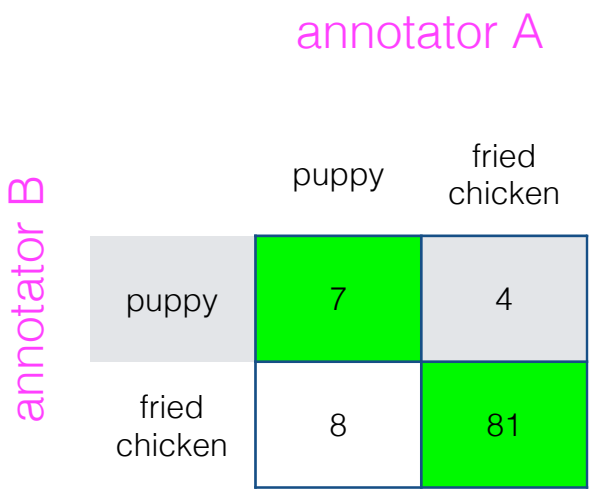


Cohen's kappa

- If classes are imbalanced, we can get high inter annotator agreement simply by chance

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

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



Cohen's kappa

- Expected probability of agreement is how often we would expect two annotators to agree assuming independent annotations

$$p_e = P(A = \text{puppy}, B = \text{puppy}) + P(A = \text{chicken}, B = \text{chicken})$$

$$= P(A = \text{puppy})P(B = \text{puppy}) + P(A = \text{chicken})P(B = \text{chicken})$$

Cohen's kappa

$$= P(A = \text{puppy})P(B = \text{puppy}) + P(A = \text{chicken})P(B = \text{chicken})$$

P(A=puppy)	15/100 = 0.15
P(B=puppy)	11/100 = 0.11
P(A=chicken)	85/100 = 0.85
P(B=chicken)	89/100 = 0.89

$$= 0.15 \times 0.11 + 0.85 \times 0.89$$
$$= 0.773$$

annotator A

annotator B

	puppy	fried chicken
puppy	7	4
fried chicken	8	81

Cohen's kappa

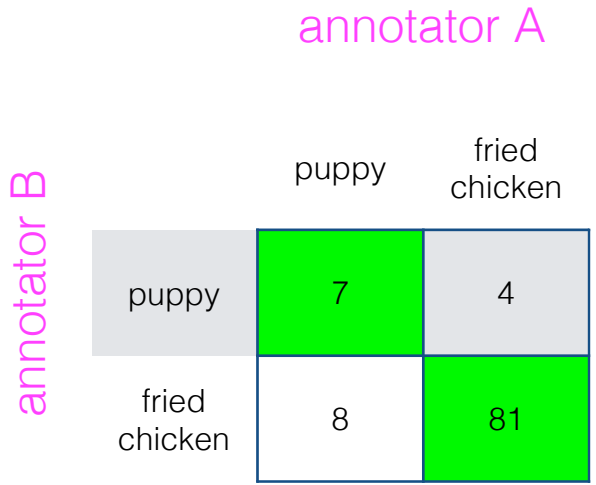
- If classes are imbalanced, we can get high inter annotator agreement simply by chance

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

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

$$\kappa = \frac{0.88 - 0.773}{1 - 0.773}$$

$$= 0.471$$



Cohen’s kappa

- “Good” values are subject to interpretation, but rule of thumb:

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

Interannotator agreement

- Cohen's kappa can be used for any number of classes.
- Still requires two annotators who evaluate the same items.
- Fleiss' kappa generalizes to multiple annotators, each of whom may evaluate different items (e.g., crowdsourcing)

Fleiss' kappa

- Same fundamental idea of measuring the observed agreement compared to the agreement we would expect by chance.

- With N > 2, we calculate agreement among pairs of annotators

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

Fleiss' kappa

Number of annotators who assign category *j* to item *i*

$$n_{ij}$$

For item *i* with *n* annotations, how many annotators agree, among all n(n-1) possible pairs

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^{K} n_{ij}(n_{ij} - 1)$$

Fleiss' kappa

For item *i* with *n* annotations, how many annotators agree, among all n(n-1) possible pairs

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^{K} n_{ij}(n_{ij} - 1)$$

Annotator

A	B	C	D
+	+	+	-

agreeing pairs of annotators →

A-B
B-A
A-C
C-A
B-C
C-B

Label	n_{ij}
+	3
-	1

$$P_i = \frac{1}{4(3)}(3(2) + 1(0))$$

Fleiss' kappa

Average agreement among all items

$$P_o = \frac{1}{N}\sum_{i=1}^{N} P_i$$

Probability of category *j*

$$p_j = \frac{1}{Nn}\sum_{i=1}^{N} n_{ij}$$

Expected agreement by chance — joint probability two raters pick the same label is the product of their independent probabilities of picking that label

$$P_e = \sum_{j=1}^{K} p_j^2$$

Krippendorf's alpha

- Kappa values still require categorical labels
- What about real-valued labels (e.g., Likert ratings, ordinal values)?

Krippendorf's alpha

	doc 1	doc 2	doc 3	doc 4
annotator A	5	5	5	1
annotator B	4	5	4	3

- We'll use the same principle that we used before: how much do our observed labels for a document differ from what we'd expect given the ratings we see?

- For real-valued ratings, we will also use distance metric to quantify how different two ratings are.

Observed

	doc 1	doc 2	doc 3	doc 4
annotator A	5	5	5	1
annotator B	4	5	4	3

… how often do we see another with this label for that same item?

when one annotator gives this label…

	1	3	4	5
1		1		
3	1			
4				2
5			2	2

Expected

	doc 1	doc 2	doc 3	doc 4
annotator A	5	5	5	1
annotator B	4	5	4	3

rating	count
1	1
3	1
4	2
5	4

Given this distribution of ratings overall, how often would we expect to see a pair of ratings together?

$$P(r_1 = 5) = 4/8$$

$$P(r_2 = 1) = 1/7$$

normalize over 7 now instead of 8 because we already selected one

$$P(r_1 = 5, r_2 = 1) = 4/8 \times 1/7 = 1/14$$

Expected

	doc 1	doc 2	doc 3	doc 4
annotator A	5	5	5	1
annotator B	4	5	4	3

rating	count
1	1
3	1
4	2
5	4

$$P(r_1 = 1, r_2 = 1) = ?$$

Expected

	doc 1	doc 2	doc 3	doc 4
annotator A	5	5	5	1
annotator B	4	5	4	3

when one annotator gives this label…

… what's the probability of seeing another with this label for that same item?

	1	3	4	5
1	0	1/56	3/56	1/14
3	1/56	0	1/28	1/14
4	1/28	1/28	1/28	1/7
5	1/14	1/14	1/7	3/14

Expected

	doc 1	doc 2	doc 3	doc 4
annotator A	5	5	5	1
annotator B	4	5	4	3

Transform these into expected counts by multiplying by the total number of annotations (8)

when one annotator gives this label…

… how often do we expect to see another with this label for that same item?

	1	3	4	5
1	0	1/7	3/7	4/7
3	1/7	0	2/7	4/7
4	2/7	2/7	2/7	8/7
5	4/7	4/7	8/7	12/7

Distance

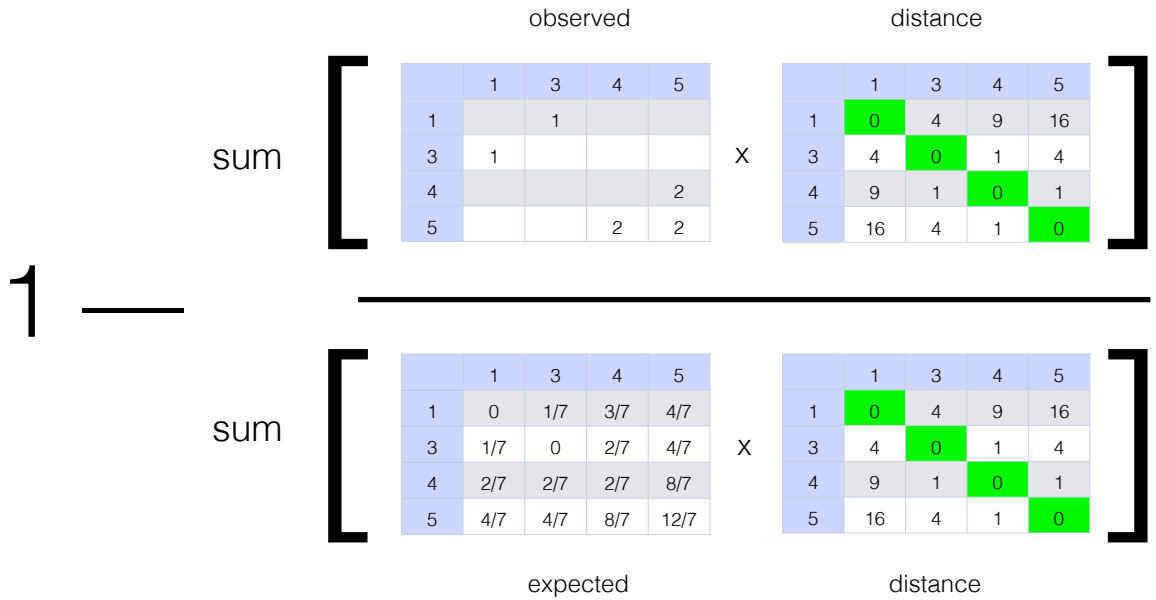
	doc 1	doc 2	doc 3	doc 4
annotator A	5	5	5	1
annotator B	4	5	4	3

For real labels, we can use the squared distance as a measure of cost.

$$(r_1 - r_2)^2$$

	1	3	4	5
1	0	4	9	16
3	4	0	1	4
4	9	1	0	1
5	16	4	1	0

Krippendorf's alpha



Implementation

- https://www.nltk.org/api/nltk.metrics.html

Activity

- Form groups of 2 or 3.

- Decide whether a given piece of text is subjective or objective using either a binary judgment (subjective/objective) or an ordinal one (e.g., 1-5). A subjective statement reflects an opinion held by a belief holder (e.g., "mint chocolate chip ice cream is terrible") while an objective statement relates factual information ("water boils at 212 degrees Fahrenheit").

- In groups in class, independently annotate the data we provide on bCourses (in `Files/Activities/subjective_inclass.tsv`) and discuss any differences you have to come to a consensus about how you will operationalize subjectivity vs. objectivity. Write up 2-3 sentences about your concept to submit at the end of class, along with your group's names + emails.