* Necessary MIMIC data transferred to /projects/opennotes/fli49/files.
* Cleaning copy-forward: Find out patients with at least 2 clinical notes. Sort them by ID and store time. Then, for each patient, apply the following pipeline.

1. Use spacy’s English tokenizer, tagger, parser and NER.

nlp = spacy.load("en\_core\_web\_sm")

2. For each note, compare it with the previous note and store the cleaned version. Use the uncleaned the version for comparison with next note.

def remove\_copied\_per\_patient(notes\_series, threshold=0.85):

cleaned\_notes = []

prev\_note = ""

for curr\_note in notes\_series:

cleaned = remove\_copied\_chunks(curr\_note, prev\_note, threshold)

cleaned\_notes.append(cleaned)

prev\_note = curr\_note

return cleaned\_notes

3. For each note, parse current note into sentences. Find the most similar sentence in the previous note. If max similarity > 0.8, the remove this sentence. Otherwise, keep it. (\* Time consuming step --- takes ~30s for each patient.) Combine all kept sentence together into cleaned note.

max\_similarity = max(

[difflib.SequenceMatcher(None, curr\_sent, prev\_sent).ratio()

for prev\_sent in prev\_sentences],

default=0

)

if max\_similarity < threshold:

filtered\_sentences.append(curr\_sent)

Question: What to do about the first note? Compare it with the second note and remove the shared content? Or keep it as a baseline and only compare the change in other notes.

* Patients with increasing note length: 39. Patients with decreasing note length: 27. If we have exact date information, we can just compare average note length of the same sample before and after open note policy.
* Compared the Shannon entropy, excluding the first note for all patients. Again, with exact date, we can compare the entropy before and after policy.

Average entropy: 8.909598168636803

Average change from previous note: -0.017437615855062188

* Count the number of repetitive words (deleted in the copy-forward cleaning) before and after policy.

Average number of words removed per note: 219.55

Average % of words removed per note: 8.40%

Length difference (min, max): 0 3355

To do:

1. Search for how copy-forward happen in clinical notes. Find some literature about it. How to evaluate the effect of cleaning? Manual vs synthetic data? Discuss in the meeting with the larger group.
2. LCS for sequence matching for now.
3. Is there race information?
4. Run on discovery.