Week of:	Description of Activities:

8/24	• TOTAL 2-3 hours
	• Met with Brandon, Viv, Mike. Introductions. Will meet weekly on Tuesdays 7:30am PT
	  ○ Teams meeting link for Tuesdays
	  ○ Reviewed Student Project FY21 Notes (mission areas, project description, etc.)
	  ○ Don't have access to https://code.usgs.gov/bserna/usgs-dash-project-generator (need login)

8/31	• TOTAL 5-6 hours
	• Tuesday 1 hour
	  ○ Met with Brandon. Will connect on Slack, install CookieCutter, Goals: open dashboard code template and play around with sample data, share a GitHub project with Brandon by the weekend.
	• Sunday 1 hour
	  ○ Read about CookieCutter, tried to unzip and open the usgs-dash-project-generator-master.zip.cpgz file without success.
	  ○ Set up Slack #usgs_science_data_internship
	  ○ Set up GitHub https://github.com/PortfolioSQA/USGS_Catalog_Dash
	• Monday 3-4 hours
	  ○ Set-up environment and became more familiar with dash
	  ○ Created a simple dashboard using sample data using my own code. Will request new cookiecutter template file again Tuesday morning.

9/7	• TOTAL 7 hours
	• Tuesday 1 hour
	  ○ Met with Brandon. Created Slack Channel. Resent CookieCutter file and opened together. Goals for the week: check out the dash template, look at xml from Science Base Catalog (project -> more details) and think about summary graphics/tables.
	  ○ Notes: 8 USGS regions, 6-7 mission areas, many science centers! Connect with elastic search index to fill contents (makes aggregation harder), new elastic search package works with pandas dataframe.
	• Friday/Sunday 2 hours
	  ○ Downloaded the cookiecutter template
	  ○ Played with the dashboard app
	  ○ Looked through Science Data Catalog xml files to determine which data is available
	  ○ Proposed a new idea for dashboard to Brandon- Query/Summarize by map boundaries:
	  ▪ User first filters by Mission Area, Science Topic, or none (dropdown menus)
	  ▪ User then demarcates geographic boundary requirement on interactive US Map (default is entire US)
	  ▪ Return record count and all records using geographic location and other filters
	  ▪ User may filter results further by Science Center, date, or sort by most frequently accessed records
	• Monday 4 hours
	  ○ Read about Dash, Mapbox, & Leafly
	  ○ Installed Shapely, Created Mapbox account

	○ Comments: Polygons are not easy to work with for the user, could choose by selecting geographic boundary by county, state, USGS region, center, or select point and radial distance. ○ What to include in sample data: name, keyword, center, lat/long, dates
9/14	• TOTAL 11 hours • Tuesday 3 hours ○ morning meeting canceled – will meet next Tuesday ○ Installed faker and created sample dataset to use in dashboard ○ Spent time understanding the bounding coordinates in the xml files <bounding> <westbc>-123.60443114739</westbc> <eastbc>-97.764587398423</eastbc> <northbc>45.032773794668</northbc> <southbc>38.897445871348</southbc> </bounding> • Sunday/Monday 8 hours ○ Worked on sample data & dashboard • Questions for Brandon/to figure out: ○ How to use a styled map from MapBox ○ Pricing on MapBox – Feasible for USGS? ○ Other options ○ Discuss best way to map boundaries of data collection (& data collected for entire US) ○ Which summaries are most important for USGS? Why? (record count by mission area/science center/popular/keywords) ○ Layer USGS regions on map? ○ Use rows from data table without displaying all need columns (join?)
9/21	• TOTAL 16 hours • Tuesday 1 hour ○ Met with Brandon ○ To do: Get pricing information for Mapbox, get map working with filtering, fix table, complete aggregate graphs, & place code in template ○ Next week: Get SDC data to play with mapping, use styled maps • Friday 3 hours ○ 'Learning App' Got map filtering working with sample data, fixed table, aggregated by visits ○ Posted mapbox pricing ○ Explored SDC data • Saturday/Sunday 10 hours ○ Graphed SDC data, but need to clean it up to be able to filter by date/keyword etc. ○ ETL SDC data ▪ Dropped NA for missing spatial (8 rows) ▪ Replaced NA for dates, kept year (see questions) ▪ Combined keyword columns ○ 'Sketch' of where to place map/table/graphs in template • Monday 2 hours ○ 'learning_app' committed to Github – has SDC data graphed in map by date/science center, ETL python file in folder ○ 'SDC_Map_Dash' template committed to Github - wireframe: ▪ Will start working on this version once we decide on layout, filtering ▪ Filter by date, map, filter by keyword (not sci center)

	▪ In each tab – Datatable + Summary Stats • mappable US data (not nationwide) • Nationwide datasets – mapped? • world/earth/international datasets – mapped? • Questions: o If nan data is filtered by date, we lose about 10% (600+ rows). Lose the rows? Save them? Include Map/graph? I replaced nan with 1900-present o Best way to determine global, continental, local data (using PlaceKeyword). Agree with three tabs? o What to do with keywords? o Could color markers by … o Moved away from selecting map area to fill table – table fills map (zooming allows user to determine if data is available in a particular area). Do we want map selection ability? Seems messier. o Using integer for year – to_datetime doesn't like 1500s o Next steps
9/28	• TOTAL 7-8 hours • Tuesday 1 hour o Met with Brandon o Notes: ▪ Viv may send invitation to bi-weekly meetings Mondays 9am PT – have not received invite ▪ Filter by Science Center & USGS Thesaurus Keyword (not by date) - done ▪ Figure out a 'pretty' way to select all key words/all science centers – note: dash doesn't seem to have one, we can select all, but shows all science centers… ▪ For now, separate global, US, mappable data by place keyword in three tabs - problematic ▪ Create data table and aggregate statistics for each tab ▪ Color markers by science center o Other: Fix map so when table returns nothing, no markers are mapped, when global or US data tab is selected – map shows what? Need to ignore rows in data table without lat/lon for the map graph data? Nan Science Center? • Tuesday 2 hours o Updated journal, separated keywords, created unique USGS thesaurus keyword and science center list for dropdowns • Monday 4-5 hours o Broke the app trying to add keyword search and spent a couple hours trying to fix it. Started from scratch. o Have not successfully separated international data (see graph) so the tabs don't make sense right now. I could work on separating by lon/lat if we want to do this. o If we still want to separate by area in three tabs, I need to place the map and science center choices inside each tab to be able to change the map/selection o So many Science Centers… can use a select all/deselect etc. need to change font size. How should we display this?
10/5	• TOTAL 10 hours • Tuesday 1 hour o Dropdown with select all – return all data o Remove tabs – or use for summaries o Zoom into populated marker area

	○ Color markers ○ Fill null Science Data Center with 'Undetermined' ○ Counts/Graph ○ Keywords ○ https://data.usgs.gov/modelcatalog/ ○ https://sciencebase.gov/datarelease/summary/ ○ Word cloud for the datatable keywords (% top 50 terms) • Friday 4 hours ○ Combined all data (instead of mappable, US, global) ○ Fixed dropdown & added All Science Centers ○ Added keyword search ○ Removed tabs ○ Fixed null SC -> Undetermined ○ Return Count • Saturday 4 hours ○ Fixed & formatted datatable ○ WordCloud ○ Next Steps ▪ Zoom to populated markers ▪ Color markers by science center ▪ Make sure key words are searched lower case ▪ Speed up wordcloud ▪ Heroku • Sunday 1 hour ○ Researched zoom to markers and marker color
10/12	• TOTAL 9 hours • Tuesday 5 hours ○ 1-hour meeting with Brandon. Notes: ▪ 10/19 check-in biweekly informal 8-8:30 am PT ▪ 10/26 SDM Biweekly 9-10:30 PT ▪ ~~Text changes (see github)~~ ▪ ~~place filters in same search box~~ ▪ ~~update datatable headers~~ ▪ ~~remove browse these data~~ ▪ ~~cache wordcloud for 'all'~~ ▪ ~~keywords – lowercase, remove punctuation~~ ▪ ~~controlled terms – USGS, etc.~~ ▪ ~~bug – CA/iso, count 2, returns 1??~~ ▪ Color by center/zoom into markers ▪ ~~update button to download CSV~~ ▪ ~~Place 'downloading state' into dt and wc~~ ▪ ~~Heroku~~ ○ 4 hours - completing above list • Friday 4 hours ○ 2 hours – completing above list ○ 2 hours - Connect with git SSH, fix files and attempted to deploy with Heroku
10/19	• TOTAL 2.5 hours • Monday 1.5 hours ○ Meeting 0.5 hours ▪ 8am Join Microsoft Teams Meeting ▪ CDI: https://my.usgs.gov/confluence/display/cdi/Home ▪ Model catalog: https://data.usgs.gov/modelcatalog/

	▪ Sciencebase data releases: https://www.sciencebase.gov/catalog/items?q=&filter=systemType%3DData+Release&filter=browseCategory%21%3DData+Release+-+In+Progress ○ 1 hour – tried to get the app to deploy on Heroku ▪ Manage.py file? heroku ps:scale web=1 ??? • Tuesday – 1 hour ○ Met with Brandon, deployed on Heroku
10/26	• TOTAL 4 hours • Monday 3 hours • Join Microsoft Teams Meeting 9am ○ Missed the meeting this morning. I'll be there next time. • 3 hours ○ Fixed loading state for count ○ If I place a loading state on map, it doesn't map correctly!? Can you help me figure out why that would be? ○ sci_center colors – works with static data!!! Talk about how to fix this, if we want it. • Tuesday 1 hour ○ Meet with Brandon and Lisa Zolly ○ Data needed: ▪ Science center ▪ Latest harvest ▪ Status ▪ Doi citation info ▪ DOI ○ Filter by: ▪ Science Center ▪ Dates (Beg, End, Updated, Latest Harvest) ▪ Status – Active/Inactive
11/2	• TOTAL 11 hours • Monday 1.5 hours ○ 8am Join Microsoft Teams Meeting • Tuesday Meeting (0.5 hours) • Looked at data (1 hour) • To Do: ○ Filter by Science Center (datasource) ○ Filter by Status - Active/Inactive (where is this?) ○ Filter by Dates (avail: first_harvest_date, last_harvest_date, last_mdate_check_date, mdate) which? ○ Filter: doi: None Do we use this? ○ Return Table: SciCenter, # citations, doi, status, date? ○ Return Count: dataset count with filters ○ Return # of citations in the last month (don't have by date) ○ Return: Pie Chart of Active/Inactive (if not filtered by this) ○ Flag certain datasets? inactive with citation? ○ Other Questions: ▪ How do I know if data is active/inactive? Is this if the data has been harvested? What would the row look like if it weren't? ▪ For the ORCID availability for retrospective DOI assignment - Do I just check if doi is None?

	▪ Lisa talked about rate of growth of the collection over time (monthly growth? Which dates?) Bar chart ▪ Do we want to show the related primary publications? ▪ Include non-primary pubs for citation count? ▪ Layout similar to dash template? Use tabs? • Friday 3.5 hours o etl o New file from Cookiecutter template using sample data o Filter by Science Center (datasource) o Filter by Status - Active/Inactive o Filter by Dates (beg, end, last updated (mdate), last harvest) o Return Table: SciCenter, # citations, doi, status • Saturday/Sunday 6 hours o Filter by Dates o Return Count: dataset count with filters o Return: Pie Chart of Active/Inactive (if not filtered by this) • Questions/Issues: o When I align the CSV download link to the right, it breaks! Why? o Using mdate instead of last_harvest_date doesn't work… Worked for a while trying to figure out problem with date type, couldn't find it o Fix tab1 so it gives the datatable for selection (doesn't return to all). Why do the other tabs retain filter info, but not datatable?
11/9	• TOTAL 8 hours • Monday 2.5 hours o 9am Join Microsoft Teams Meeting 9am o Fix count and commit to new repository on GitHub: https://github.com/PortfolioSQA/SDC_Manager_Dashboard • Tuesday Meeting 1 hour o TODO: Fix table (so it doesn't change to default when tabs are changed), ~~Fix mdate~~ o Talked about NLP project - creating a custom NER to identify models from epubs warehouse?? for the Model catalog. If time permits start researching this. o Also talked about volunteering for usability testing for Sophie • Saturday 3 hours o Fixed mdate o Worked on datable persistence for a couple hours… I can't figure out what to do to keep the table from reverting to default in tab. o Changed 0/1 => inactive/active • Sunday 1.5 hours o Research on how to create custom NER model o Tried for another 1-2 hours to debug the table persistence. Used dcc.store, persistence = True, and filtering in another callback. Didn't get it to work ☹
11/16	• TOTAL 4.5 hours • Monday 0.5 hours o Monday 8am Join Microsoft Teams Meeting • Tuesday – meeting cancelled due to power outage • Wednesday 3 hours o Deploy to Heroku o Tab update works in Heroku. Weird! o Fixed datatable sort o Need to fix table so we can sort by status and right align download link

	• Sunday 1 hour ○ Created random dates in Excel for sample data
11/23	Thanksgiving Week – WEEK OFF
11/30	Final Exam – WEEK OFF

12/7	• TOTAL 12.5 hours
	• 1.5 hours Monday Join Microsoft Teams Meeting 9am
	• 1 hour Tuesday meeting
	  ○ Methods: CNN, SVM, etc...
	  Tasks: Image classification, image segmentation, etc...
	  Applications (use cases): precipitation-induced landslide warning, tracking rainfall thresholds,
	• 3 hours Tuesday
	  ○ Create sample data (dates for beg, end, update, harvest), change code to incorporate new 'dates'
	• 3 hours Wednesday
	  ○ Debug code changes for new sample dates, deploy on Heroku, update journal, e-mail Sophie with updates
	• 2 hours Thursday
	  ○ Create notes document for model scraping
	  ○ Review text
	• 2 hours Sunday
	  ○ extract abstract, identify method, task and application
	  ○ Create txt files with abstracts

12/14	• TOTAL 10 hours
	• 0.5 hours
	  ○ Monday 8am Join Microsoft Teams Meeting
	• 1 hour
	  ○ 2pm - Meet with Lisa, Brandon, and Sophie
	• TO DO LIST:
	  ○ ~~Rename dashboard~~
	  ○ ~~Link – learn how metrics are calculated~~
	  ○ Table – font choice?
	  ○ ~~Org of interest -> Choose your science center/program~~
	  ○ ~~Select dataset status~~
	  ○ Explanation of active/inactive
	  ○ ~~Put date selection in same box~~
	  ○ Justify radio buttons
	  ○ ~~Colon after headers~~
	  ○ ~~Remove beg/end dates~~
	  ○ ~~Last harvested, last updated -> start date?~~
	  ○ Flexibility in search by date? Calendar year?
	  ○ ~~Dataset Count - > Active, Inactive, Total, pie chart~~
	  ○ ~~Make stacked bar graph for active inactive by date~~
	  ○ ~~Remove pie chart tab~~
	  ○ ~~doi – links?~~
	  ○ ~~Alternate IDs for datasets?~~
	  ○ ~~Delete widgets in graph area~~
	  ○ ~~Format date label~~
	  ○ ~~Download -> more usable file name~~
	  ○ Help & documentation (Tool tips)
	  ○ Errors: If it's not working what happens?
	• 8.5 hours

	○ Complete TODO list, redesign site, insert links for DOI, alternate identifiers, stacked bar chart for counts by date, delete widgets in graphs
12/21	WEEK OFF - HOLIDAY
12/28	• TOTAL 7-8 hours • 30 min -Monday 8am Join Microsoft Teams Meeting ○ https://data.usgs.gov/datacatalog/ • 1 hour – Tuesday Meeting with Brandon ○ Questions: title, ~~markdown – open link in new tab/window for DOI,~~ filter by active/inactive?, explanation for active inactive, Errors ○ **TODO for dashboard**: ~~table -> dbc, tab (dash -> SDC Dashboard), Sentence – bold or large numbers,~~ ○ ~~https://dash-bootstrap-components.opensource.faculty.ai/docs/components/table/ - dbc table~~ ○ https://data.usgs.gov/datacatalog/api/docs/v1 , swagger API – look at documentation to get data except citation (leave as null for now) ○ **TODO for NER**: ~~exploratory functional prototype: look at https://paperswithcode.com/ , format text in jsonl, get prodigy (costs $390),~~ {"text": "This is a text"} {"text": "This is another text"} https://data.usgs.gov/modelcatalog/search Training Data 1. https://pubs.usgs.gov/tm/14/a2/tm14a2.pdf 2. https://pubs.usgs.gov/of/2008/1159/downloads/pdf/OF08-1159.pdf 3. https://pubs.usgs.gov/of/2016/1136/ofr20161136.pdf 4. https://pubs.usgs.gov/of/2007/1088/pdf/of07-1088_508.pdf 5. https://www.mdpi.com/2073-4441/8/1/17 6. https://pubs.usgs.gov/wri/1990/4130/report.pdf 7. https://pubs.usgs.gov/tm/12b1/ 8. https://pubs.usgs.gov/tm/2006/tm6b3/ 9. https://www.mdpi.com/1999-4893/1/2/52 10. https://data.usgs.gov/modelcatalog/data/5eb4485782ce25b5135abf28 11. https://data.usgs.gov/modelcatalog/data/5eb4485082ce25b5135abee3 12. https://data.usgs.gov/modelcatalog/data/5f6240eb82ce38aaa2361498 13. https://data.usgs.gov/modelcatalog/data/5eb4485e82ce25b5135abf70 14. https://data.usgs.gov/modelcatalog/data/5eb4485f82ce25b5135abf86 15. https://data.usgs.gov/modelcatalog/data/5eb4485f82ce25b5135abf7c 16. https://data.usgs.gov/modelcatalog/data/5eb4485582ce25b5135abf16 17. https://data.usgs.gov/modelcatalog/data/5f036b3f82ce0afb2446e04a 18. https://data.usgs.gov/modelcatalog/data/5eb4486182ce25b5135abfb0 19. https://data.usgs.gov/modelcatalog/data/5eb4485282ce25b5135abef4 20. https://data.usgs.gov/modelcatalog/data/5eb4486182ce25b5135abfaa • 3 hours – Wednesday ○ Update journal, look at prodigy ($390 license), get 10 more text files, create jsonl file (see above links) • 3 hours --Thursday ○ Dashboard updates, dbc tables are SUPER SLOW - decided to format dash data table (I have another file with the dbc table I can show if you'd like)

1/4	• TOTAL 9 hours
	• 2 hours - Monday:
	  ○ Doctor's Appt- Couldn't join meeting Join Microsoft Teams Meeting 9am
	  ○ 2 hours Install prodigy, en_core_web_sm, etc. Label text, git commit, etc.
	• 1 hour - Tuesday
	  ○ Meet with Brandon 7:30 – will meet with Mike & Viv soon, Ruby Gem linguist shows which languages in code,
	  ○ TODO: ~~10-20 unlabeled texts for evaluation data (good and bad examples),~~ run in prodigy (see slack), try again using shorter NEs, use en_core_web_lg, future: establish list of methods, concept of datasets on prodigy?, deploy dbc table on Heroku and see if its faster
	• 3 hours – Thursday
	  ○ 1+ hours Meet with Sophie to talk about Usability Testing
	  ○ 2 hours Evaluation Data (in excel & jsonl)
	• 3 hours – Friday
	  ○ Train/evaluate ner – 0% accuracy, relabeled training set 0% accuracy, read more documentation, tried with models only, frustrating - next combined to 40 for training set and re-labeled. Didn't help
	1. https://doi.org/10.1126/science.aat4723
	2. https://pubs.usgs.gov/of/2001/ofr-01-0002/
	3. https://data.usgs.gov/modelcatalog/data/5ff62dc1d34ea5387df035fa
	4. https://data.usgs.gov/modelcatalog/data/5eb4485982ce25b5135abf3c
	5. https://doi.org/10.1111/gwat.12397
	6. https://data.usgs.gov/modelcatalog/data/5eb4485482ce25b5135abf0e
	7. https://data.usgs.gov/modelcatalog/data/5eb4485682ce25b5135abf1c
	8. https://data.usgs.gov/modelcatalog/data/5eb4485382ce25b5135abefc
	9. https://doi.org/10.3133/wri874163
	10. https://doi.org/10.1002/2017JC013204
	11. https://doi.org/10.1029/2011JB008968
	12. https://doi.org/10.3133/wri974022
	13. https://pubs.usgs.gov/tm/tm4f2/
	14. https://doi.org/10.1016/j.ocemod.2010.07.010
	15. https://doi.org/10.3133/ofr20151009
	16. https://data.usgs.gov/modelcatalog/data/5ef3952782ced62aaae3ef55
	17. https://pubs.usgs.gov/tm/tm6a37/
	18. https://doi.org/10.1002/joc.3625
	19. https://doi.org/10.3133/tm14A1
	20. https://doi.org/10.3133/tm6A43

1/11	• TOTAL 9.5 hours
	• 2 hours - Monday:
	  ○ 30 min. Monday 8am Join Microsoft Teams Meeting
	  ○ 1.5 hours – tried again with prodigy, read a bit on topic analysis, dbc table. No pagination, can't sort, slower
	• 2 hours - Tuesday
	  ○ Meet with Brandon 7:30
	  ○ TODO: Tutorials, analyze n-grams, constrain to better examples
	  ○ Peer review for a "resource review" that Sophie preparing for the CDI Usability Collaboration Area
	• 2 hours - Wednesday
	  ○ Prodigy tutorials https://prodi.gy/docs/named-entity-recognition (Food Ingredient entities)

	• 1.5 hour - Thursday
	  ○ Continued Prodigy tutorial
	• 1 hour – Friday
	  ○ Continued Prodigy tutorial
	• 1 hour – Saturday
	  ○ N-gram analysis of text
	  ○ https://github.com/PortfolioSQA/USGS_Catalog_Dash/blob/master/ngram_text_analysis.ipynb

1/18	• TOTAL 10 hours
	Martin Luther King Jr Holiday Monday
	• 0.5 hour - Tuesday
	  ○ Meet with Brandon 7:30 – 8:00
	  ○ TODO: methods NER: try water balance, transport model, etc. separate and then try DE, Lin reg, random forests, etc. (may have to use regex for that), Graphical interface - Keep track of articles that may not be models, New Github directory with files and commands
	• 2.5 hours – Friday
	  ○ New model for methods (1) Used all methods
	• 4 hours - Saturday
	  ○ New model for methods (2) Used geologic models (not stats and ML methods)
	• 3 hours – Sunday
	  ○ Model 3 + notes

1/25	• TOTAL 10.5 HOURS
	• 1.5 hours Monday
	  ○ 30 min Monday 8am Join Microsoft Teams Meeting
	  ○ Notes, github commit, ml model
	• 0.5 hours Tuesday
	  ○ Meet with Brandon 7:30-8:00
	  ○ TODO:
	  ▪ ~~Get another 20 texts for testing (abstracts)~~
	  ▪ ~~Train again with 'bad examples' see if make-gold improves~~
	  ▪ ~~Text (jsonl) files for entity seeds (separate models)~~
	  ▪ ~~Train for words before 'model' using verb 'model' as bad examples~~
	  ▪ ~~Research existing work for ML NER, articles and/or list for seed terms~~
	  ▪ Goal: blog post (search for other blogs pertaining to the topic)
	• 1.5 hours Friday
	  ○ Gather 20 new texts from Model Catalog
	• 4 hours Saturday
	  ○ Train new model – geological terms with 'bad examples'
	  ○ Train new model – words that are prior to 'model'
	• 2 hours Sunday
	  ○ Anaconda environment broke after update. Spent an hour trying to fix it. (Didn't count this as hours). Took a break from Prodigy to research.
	  ○ Research available ML model detection algorithms, seed lists

2/1	TOTAL 11 HOURS
	• 1.5 hours Monday Join Microsoft Teams Meeting 9am
	• 1 hour Tuesday
	  ○ 1 hour Meet with Brandon
	  ○ New model – train ALGO model with arxiv examples (don't use acronyms)
	• 3.5 hours Wednesday
	  ○ Downloaded and put ML arxiv data in jsonl format for prodigy (train 3k+, test 1k+)

	○ Seeded, labeled about 140 texts, pre-trained (started at 4:45 pm, ended at 12:45 next day = 20 hours) ○ Using the ignore button incorrectly! Used if you don't know values, not if they are all wrong. Deselect any text that is correct and click reject. • 2 hours Friday ○ Use seeds from last model, redo manual train using buttons correctly (209 texts), 1025 make-gold texts and resulted in 44% accuracy. Tried to use the same pre-training since it takes so long, but neat to start again clean. First reduce the size of the train set so it doesn't take so long to pretrain, then re-seed and make new model ▪ Tips for training (consistency is key, take pictures to remind yourself of how you trained the model or go the jsonl file and search for the text, the model fails if you are inconsistent in labeling the entities) • 3 hours Sunday ○ Tried training a new model and ran into error (in commands)
2/8	TOTAL 11.5 HOURS • 3.5 hours Monday ○ 8am Join Microsoft Teams Meeting ○ Attempt 3 with the arxiv texts/algo model, lots of labeling ☺ ○ Watched NER videos: https://www.youtube.com/watch?v=sqDHBH9IjRU https://www.youtube.com/watch?v=UxzyD6gVlC8 ○ Link to documentation: file:///Users/sashaqanderson/Downloads/PRODIGY_README.html • 1 hour Tuesday ○ 7:30 am Meet with Brandon • 1.5 hours Wednesday ○ Virtual Student Federal Service (VSFS) career & fellowship programs - I guess this was just recruiting, didn't know what kind of meeting it was exactly so I attended ☺ • 6 hours Friday ○ Train ALGO model4 – see highlighted models below. ▪ First work on arxiv – algo model, then GEO model again. ▪ MODEL 1: (GEO + ML Methods): 64% (seemed to label everything) ▪ MODEL 2 (GEO + ML Methods): 68% (ran into labeling problem) ▪ MODEL 3 (GEO): 69% - better results than I thought – Could work on this model a bit more… *Prodigy ner.print-stream geo3_model /Users/sashaqanderson/Dropbox/USGS/NER_Work/ner_text_test20.jsonl --label GEO* ▪ Model 4: ML: 0% didn't make gold (skipped model 5 oops :-) ▪ MODEL 6 (GEO): 46% accuracy (ran into labeling problem) ▪ Model 7: (MDL) 25% accuracy – more of GEO model than ML model - maybe poor performance because of seeds ▪ ALGO MODEL COMMANDS – 44% (used different pretrain model – bust) ▪ ALGO MODEL2 – 25% (training error – better model than 3 but ran into the error. Use these rules) *Prodigy ner.print-stream algo_model2 /Users/sashaqanderson/Dropbox/USGS/NER_Work/ner_text_test20.jsonl --label ALGO* *Prodigy ner.print-stream algo_model2 /Users/sashaqanderson/Dropbox/USGS/NER_Work/algo_model3/arxiv_test.jsonl --label ALGO*

	▪ ALGO MODEL3 – 37% (does a good job with some texts, terrible with others) ○ Got 65% accuracy after about 5 hours of labeling/training, then went to 33% after more make-gold. ○ New saved model: *Prodigy ner.print-stream algo_model4 /Users/sashaqanderson/Dropbox/USGS/NER_Work/algo_model3/arxiv_test.jsonl --label ALGO*
2/15	TOTAL • 2 hours Monday ○ Read about Spacy & Prodigy and worked on NER Summary (thinking about why my model went from 65 to 33% after a bit more make gold for the ALGO model4 ○ 10-15 weeks left of VSFS internship?? Can I set some intermediate goals for myself?
2/22	Monday 8am Join Microsoft Teams Meeting
3/1	Monday Join Microsoft Teams Meeting 9am
3/8	Monday 8am Join Microsoft Teams Meeting
3/15	Monday Join Microsoft Teams Meeting 9am
3/22	Monday 8am Join Microsoft Teams Meeting
3/29	Monday Join Microsoft Teams Meeting 9am
4/5	Monday 8am Join Microsoft Teams Meeting
4/12	Monday Join Microsoft Teams Meeting 9am
4/19	Monday 8am Join Microsoft Teams Meeting
4/26	Monday Join Microsoft Teams Meeting 9am
5/3	