**Sherlock**

Should have access for 1 month.

<https://fellowship.sherlockml.net>

Steven and Neil M should have access.

Project name: “Fight for Sight – Fellowship XII”

Online platform for editing and running python scripts in the cloud.

**Python Installation**

Install the “Anaconda” version of python: <https://www.anaconda.com/download/> Anaconda is python with many commonly used libraries pre-installed.

I have tried to make it so you won’t need to do this yourselves, but you may need to install some other libraries, e.g. **wordcloud**, umap-learn, pyldavis, fuzzywuzzy, hdbscan, python-Levenshtein, genism.

To install a library: open “Anaconda Prompt” (search for it in start menu). In the command prompt window type and execute (hit enter) “pip install <library-name>”, where <library-name> is one of the names above.

To start Jupyter notebook: open “Anaconda Navigator”, then click “Launch Jupyter Notebook”

Then use File -> Open to find the .ipynb file you want to run.

If you want to try to learn something about python, start with something about the basics, Jupyter/IPython, and the pandas library.

**Files**

**“data” folder:**

* **CCGrantTracker folder**: Raw data from CCGrantTracker and processed data including a common grant ID (MasterID).
* **EPMC folder**:
  + eye\_in\_new\_keywords: Metadata for all publications returned by eye keyword search.
  + ffs\_papers: All papers acknowledging Fight for Sight or one of its predecessors.
  + ffs\_or\_predecessors\_citations: All papers that cite a paper in the ffs\_papers dataset.
  + .pkl files: Non-human readable format that python can quickly load.
  + .csv files: Human readable “spreadsheet” style data.
  + \_COUNTVEC files: counts of words in abstracts as needed for topic analysis model.
  + \_LDAVEC files: topics for each abstract
  + \_LEMMA files: edited abstracts with uninformative words etc. removed
  + “json” folder: Raw data as returned from queries of the EPMC database.
* **FilingCabinet folder**: Joanne’s spreadsheet + a processed version only including grants present in award master.
* **ResearchFish folder**:
  + Main file – ResearchFish\_master.xlsx (one entry per grant)
  + Other files - Uploaded info on conferences, publications etc. (potentially multiple entries per grant)
  + raw folder – all data files as downloaded from ResearchFish.

**“models” folder:**

* **sklearn:** Topic models.
* **umap:** Model for making topic map visualisations.

**“plots” folder:**

* **epmc**: General publication plots
* **ResearchFish**: from research fish data.
* **Umap**: Topic visualisation maps.
* **Wordcloud**: Important words in each topic.
* **Topics:** General topic related plots.

**Scripts and Notebooks:**

* **UpdatePapers.ipynb (NEW):** Query EPMC to update the dataset of fight for sight publications.
* **PaperStatistics.ipynb (NEW):** Get some basic stats on number of publications and citations per year etc.
* **ResearchVisualisations.ipynb (NEW):** Make visualisation maps, wordclouds and topic variation over time.
* **CCGrantTracker.ipynb:** Process CC Grant Tracker data.
* **Combine.ipynb:** Combine data from award master, filing cabinet, research fish and CC grant tracker.
* **Exploratory.ipynb:** General analysis of award master spreadsheet.
* **PublicationAPIs.txt:** Some info on how to query EPMC.
* **ResearchFish.ipynb:** Process research fish data.
* **ResearchFishAnalysis.ipynb:** General analysis of research fish data.

**General Recommendations**

* Push to get as much out of the relationship with CC Grant Tracker/Digital Science as possible.
* Setup reporting on CC Grant Tracker so that the data from all fields in the applications/reports are present in the exported data.
* Researchers actually uploaded a lot of useful information to ResearchFish, including publications, conferences, further funding etc. etc. Try to setup a similar functionality on CC Grant Tracker/some other system to replace it.
* Settle on a common grant identifier format that is applied in all data sources. Ideally one number/alphanumeric word rather than the 1450/1451 type format.
* Unique identifiers in general are incredibly useful to match up data. E.g. if there was a way to have an ID for things like author, institute etc. that would be great.
* Or if it’s not possible to have a unique ID number, try to limit free text fields as much as possible. E.g. having a dropdown list of institutions to pick from is much less likely to cause issues with typos, different address formats etc. than letting people write it in for themselves.
* Try to settle on a system for grouping disease areas/types.
* If you do another ASI/similar project, try to have one specific, well-defined problem to solve to get as much as possible out of the 6 weeks.
* Possible areas with potential for (partial) automation: application/submission reminders, searching for peer-reviewers, general application checks (are all fields correct, conflicts of interest etc.), support for reviewers (e.g. flagging similar research), transfer of information between systems/sources (e.g. current copy/pasting of information between CC Grant Tracker, Research Fish and Master spreadsheet), monitoring of new research/news.
* Outside the research team there are definitely projects that could be done on the fundraising/marketing side of the charity, depending on the data that’s available.