TODOs:

* Set up environment
* Create git repo

Notebook: list US news World top numbers and random number generator

1 scrape top cs faculty and get random number generator to pick

<https://www.cs.washington.edu/people/faculty/>

collect:

* Get all links
* Get text from page (as unstructured?
* Cv
  + Parse for dates?
* Google scholar
* Linkedin
* ~~Social media~~
* ~~GITHUB?~~
* Run google search?
* Run nyt search?

Things to learn:

* Pdf parsing
* Linkedin api?
* Google scholar scraping [scholarly]
  + Can I just scrapy it give a page?

Final df:

* Timeline of jobs
* Timeline of publications?

Questions:

* Coding in outside academia [hard code as position from CV, linkedin, or other source?

Goal:

* Correlation publication per year [before and after outside academia activity]
* Type of publication [field, topic modelling?]
* Success of publication (google scholar citations [controlling for time])
* Affiliation of co-authors?

8/8/21 update

TODO dc acm/ dlcp scraping

Can run google scholar scraper once, but dc acm gives topics and citation metrics seemingly for free…

Flag: Magdalena Balazinska is id = 4, but currently does not match correctly on linkedin csvs

Sara Mostafavi has error in experience parsing…