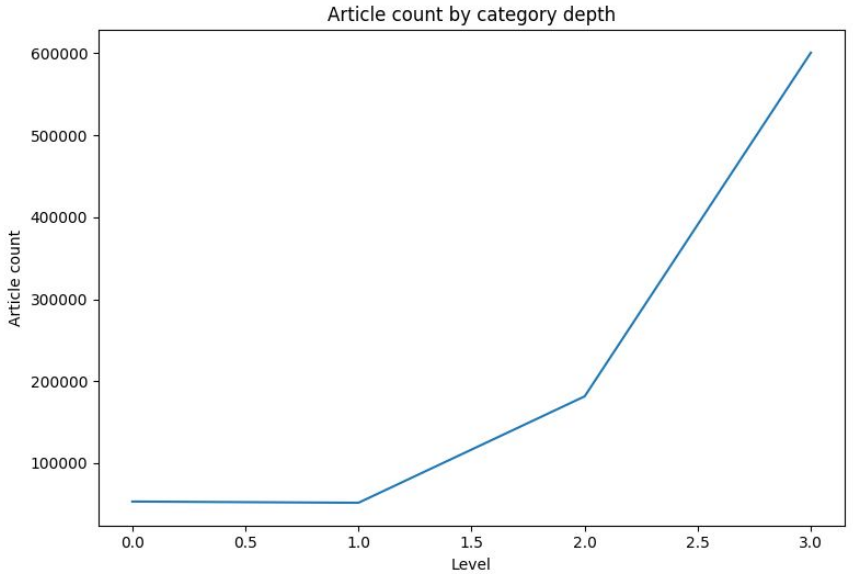
General Updates - 2/19

Category tracking

→ Issues gathering depth of > 3 for a large number of categories (timeout errors)

→ At a depth of 3 (with these categories), there are already a total of 500,000+ articles



Category selection reminder

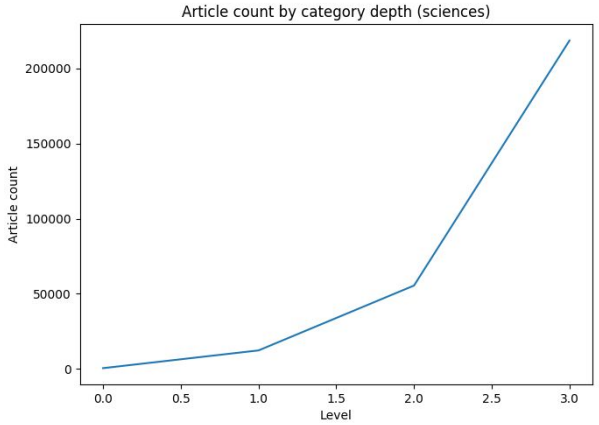
sciences : ['Branches_of_biology', 'Fields_of_mathematics', 'Concepts_in_physics', 'Chemistry']

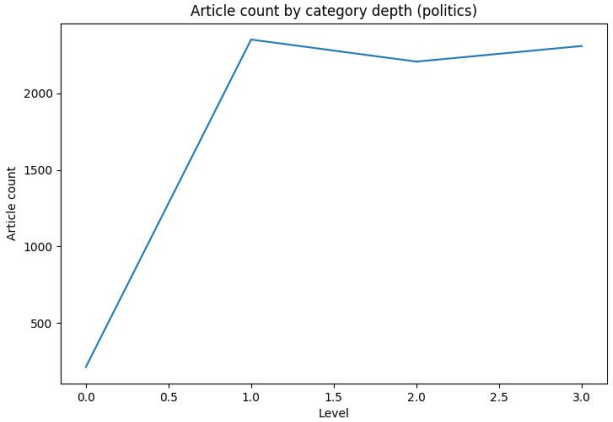
sports : ['Ice_hockey_in_the_United_States', 'American_football_in_the_United_States', 'Basketball_in_the_United_States', 'Baseball_in_the_United_States']

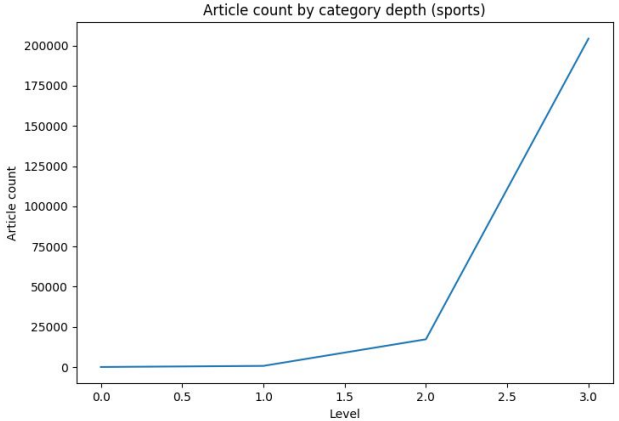
culture : ['Television_in_the_United_States', ‘American_films', 'American_novels']

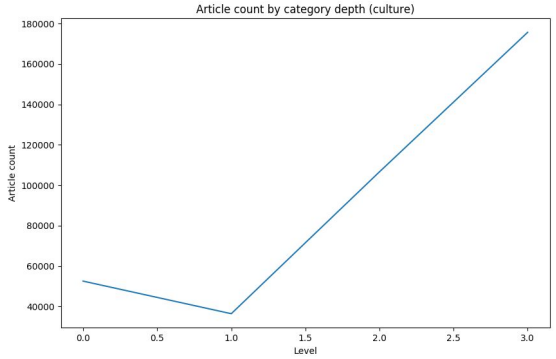
politics = ['Conservatism_in_the_United_States', 'Liberalism_in_the_United_States']

Sizes by category selection









Are articles still relevant after 3 levels?

Short answer: Yes, generally. 225 of the 240 sampled articles were “relevant”.

Methodology: Sample 20 articles from each level from each domain. Confirm whether each article is relevant to the initial domain & category or not.

Long answer: Things get a tiny bit wacky at level 3. There are some stubs (1-2 sentence articles) being collected. There are also some redirect pages. We can toss these easily. Political pages also included some conservative publications (do these represent ideology?). Liberal pages tended to be more historical (e.g. Ben Franklin, Boston Tea Party, Revolutionary War). A solution to this might be to just sample from a broader category. (Remember that relatively few political pages were collected).

Next steps

Use different, larger category for politics

Finish scraping culture using the same style

Sample articles wayyyy down. We were getting good results with a dataset of ~450 articles using the previous politics dataset; it would be wise to sample ~1,000 articles from each category to keep everything managable (and not put too much strain on scraping)