BY KATHARINE MILLER

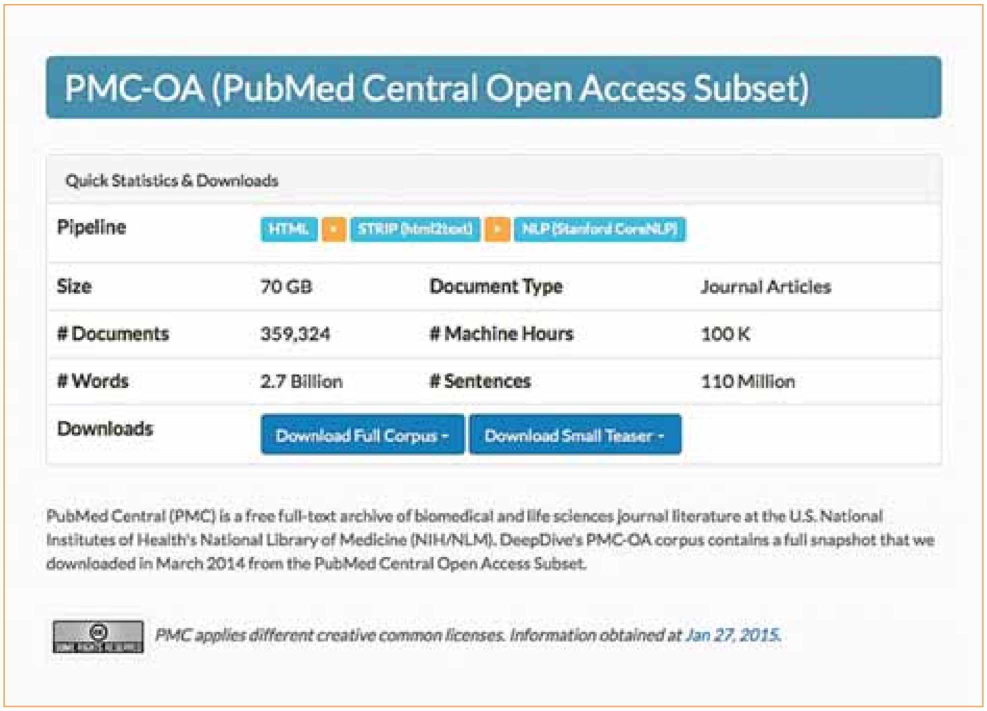
Giving Away the NLP Store

In order for computers to extract knowledge from the vastness of all biomedical literature, the machines must first determine the structure of the natural language text—what are the nouns and verbs and what is their role in each sentence. But transforming such a large body of literature into a series of dependency trees not unlike the sentence diagrams of old-time grammar classes requires millions of CPU hours. Fortunately, researchers associated with the Mobilize Center have completed this preprocessing task for several large biomedical literature datasets (including PubMed Central's Open Access Subset, PLoS and BioMed Central)—and are offering the marked-up resources for use by others.

"We're really excited that we've been able to preprocess these datasets and make them available," says **Chris Ré, PhD**, assistant professor in the computer science department at Stanford University. "Now other researchers who want to prospect the scientific literature and perform natural language processing can get all that detailed markup for free. We hope they find something interesting with it."

Most labs can't afford to preprocess such a huge amount of literature, Ré says. But in collaboration with the Center for High Throughput Computing at the University of Wisconsin, Ré's group gained access to many hours of computational time using the Open Science Grid. This allowed them to mark up large volumes of creative commons literature for their own work—and now offer it to others.

Before the end of 2015, Ré plans to use DeepDive, a free, open-source, probabilistic inference engine developed by his lab, to go a step further with his natural language processing (NLP) analysis of the biomedical literature and release those results for free as well. If one thinks of NLP preprocessing as identifying the nouns, verbs, and objects (x inhibits y, for example) then DeepDive might be determining the nature of the entities—whether x and y are genes or proteins, for example—as well as the genes' or proteins' relationship to some specific disease term described in the same paper. "These inference or entity resolution problems are very challenging computationally as well as challenging to get high quality on," Ré says. But DeepDive has proven adept at the task, sometimes outperforming expert annotators.



PMC-OA (PubMed Central Open Access Subset)

Quick Statistics & Downloads

Pipeline	HTML › STRIP (html2text) › NLP (Stanford CoreNLP)		
Size	70 GB	Document Type	Journal Articles
# Documents	359,324	# Machine Hours	100 K
# Words	2.7 Billion	# Sentences	110 Million
Downloads	Download Full Corpus ▾	Download Small Teaser ▾	

PubMed Central (PMC) is a free full-text archive of biomedical and life sciences journal literature at the U.S. National Institutes of Health's National Library of Medicine (NIH/NLM). DeepDive's PMC-OA corpus contains a full snapshot that we downloaded in March 2014 from the PubMed Central Open Access Subset.

PMC applies different creative common licenses. Information obtained at Jan 27, 2015.

Screenshot from the DeepDive Open Datasets web page offering preprocessed NLP analysis of the open access portion of PubMed Central.

How it works: Domain experts specify the kinds of relationships or features they are interested in. They might provide examples from ontologies or a sample of manually curated data, or they might just explain to DeepDive researchers how to reason with the data. "We take that specification and translate it into a large probabilistic inference problem," Ré says. "We solve that and produce data for the researchers." It's an iterative process—the researchers look at what comes out and give feedback that is used to train the system over time, so it can extract the entities or relationships more robustly.

But the release of the NLP datasets should be useful to people even if they don't use DeepDive, Re says. Right now, people are still just downloading the datasets and poking around. "The excitement is: we're giving away data," Re says. "If it's useful to anyone, send us a note." □

DETAILS

The DeepDive Open Datasets are available at http://deepdive.stanford.edu/doc/opendata/