Preparing Your Corpus

By Devin Cornell (dcornell@ucsb.edu)

Email me if you have any questions or need help.

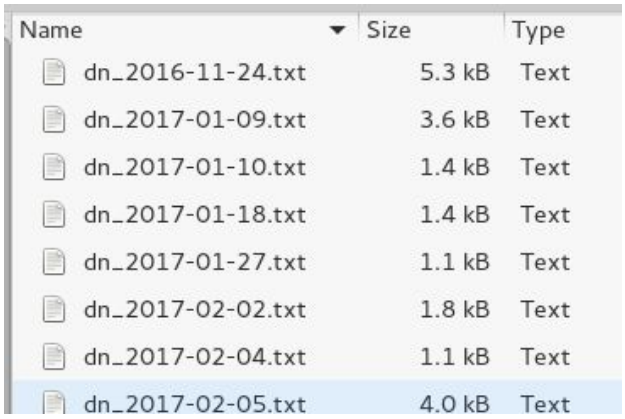
- What does a corpus look like?
- How can I download articles from Lexis Nexus?
- Can I collect my data any other way?

Conclusion

What does a corpus look like?

Text analysis is performed on a *corpus*, or a collection of documents. The selection of your corpus is important - it will determine what kinds of questions you can answer and the arguments you can make. For our purposes, a corpus should be prepared as a series of text files. I'll show you how to construct a corpus related to Betsy DeVos for demonstration. Try to imagine how the same instructions can apply to the questions you'd like to answer for yourself.

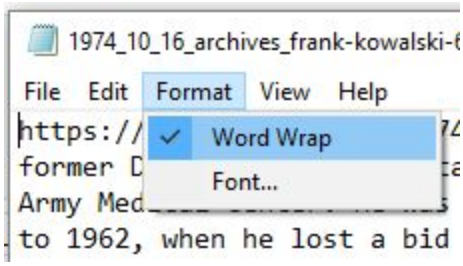
Our corpora will just be a set of text files in a directory. Text files always end with ".txt", (although your computer may not show the file extension to you). It will still have an icon that looks somewhat like a notepad. Note the filenames are organized by source daily news (dn) and new york times (nyt - files not shown here), then by date yr, month, day. This is important because it will allow us to separate the data by sorting by filename. All of the text files should appear in one folder.



The context of the text file will simply be the article then. This is a daily news article appears in one text file by itself.

```
                          Daily News (New York)

                       November 24, 2016 Thursday
                     SPORTS FINAL REPLATE EDITION

TRUMP'S 1ST LADIES * Taps voucher big for Ed. Dept. * S.C. Gov. Haley to
represent U.S. at U.N.

BYLINE: BY ADAM EDELMAN NEW YORK DAILY NEWS With Ben Chapman and News Wire
Services

SECTION: NEWS; Pg. 8

LENGTH: 736 words


DONALD TRUMP moved to fill two key cabinet positions Wednesday - naming GOP
fund-raiser and school-choice advocate Betsy DeVos education secretary and
offering South Carolina Gov. Nikki Haley the position of UN ambassador.

The selections, the first two women to be named to the President-elect's
cabinet, raised eyebrows across the political spectrum - with many questioning
Haley's credentials and DeVos' policy positions.
```

Windows users: If you open the text file in notepad, you may need to turn on ‘word wrap’ for the Betsy Devos articles since the text all appears on one line.

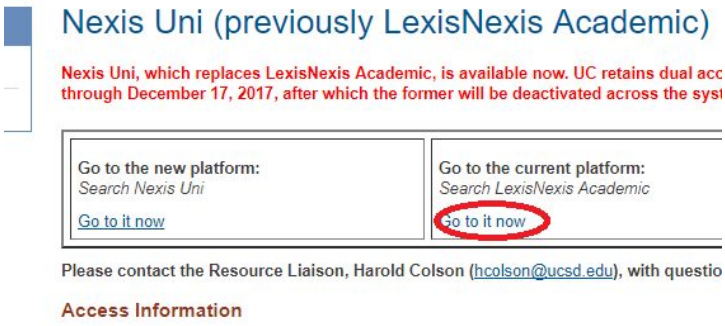


I’ll show you how to build this corpus using Lexis Nexus.

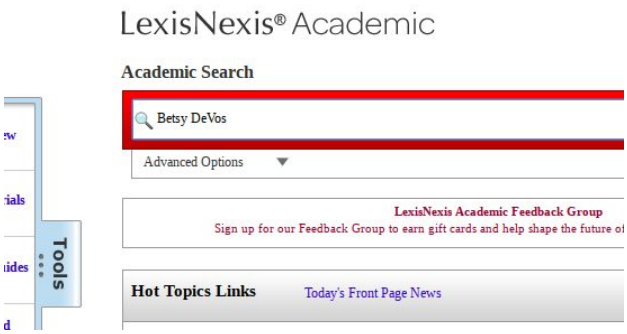
How can I download articles from Lexis Nexus?

I’ll show you how to build a ‘Betsy DeVos’ corpus using the Lexis Nexus database. First, find the link to Lexis Nexus on the library page; this is a shortcut: [NexisLexus]() (this link works if you are on the UCSB campus, not sure about other locations/universities).

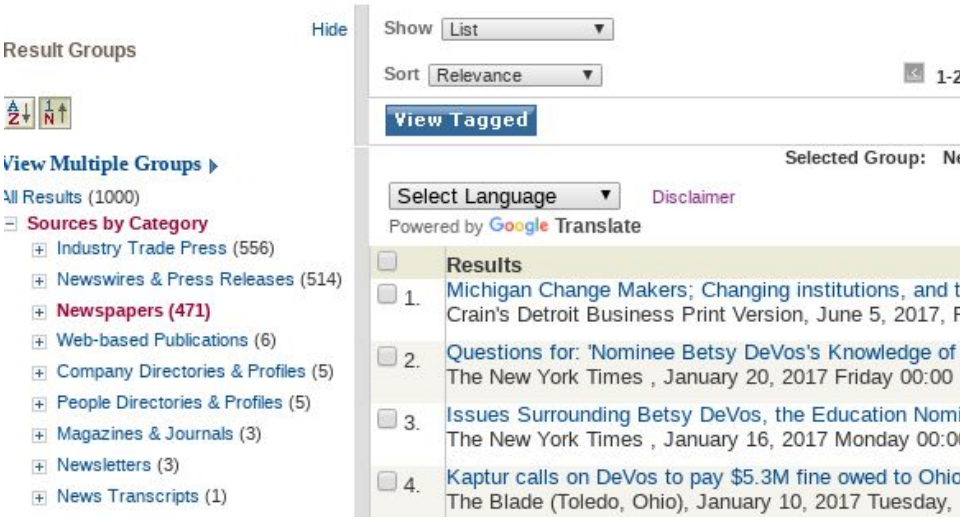
Choose to use the current version (will be out of date in 2018 I guess).



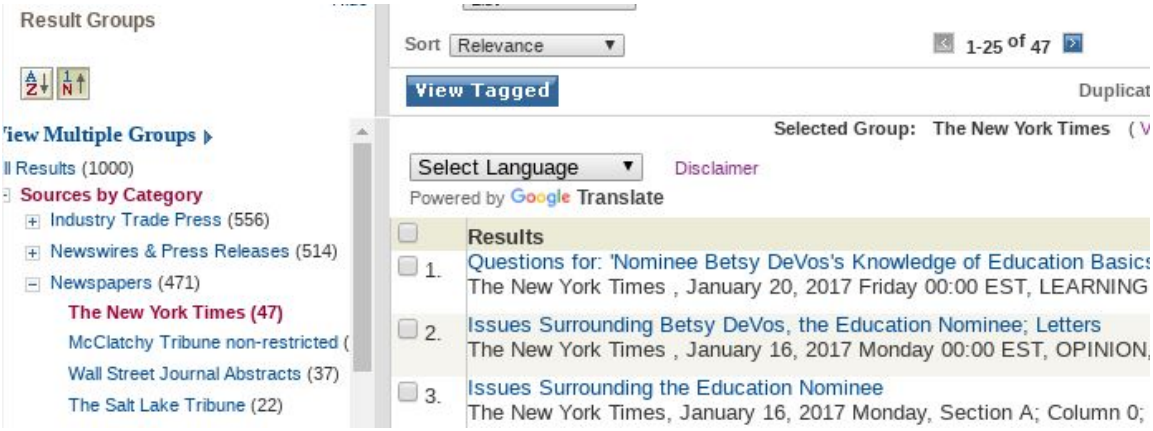
Now start searching. It is up to you to select your query as appropriate for the corpus you want to build. My corpus will be built using “Betsy DeVos” as a search term.



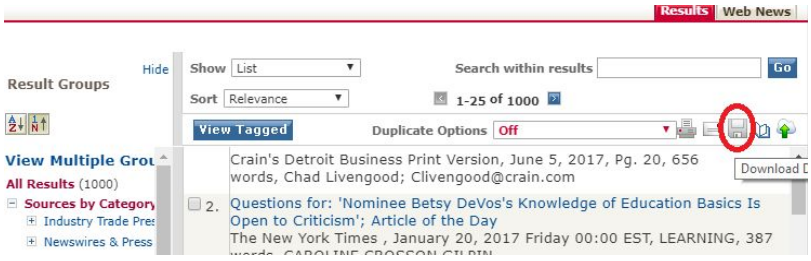
You’ll get the following search results. Since I’m interested in Newspapers, I select that on the left-hand column. It appears that there are 471 documents in that category, so it will narrow down the search results.



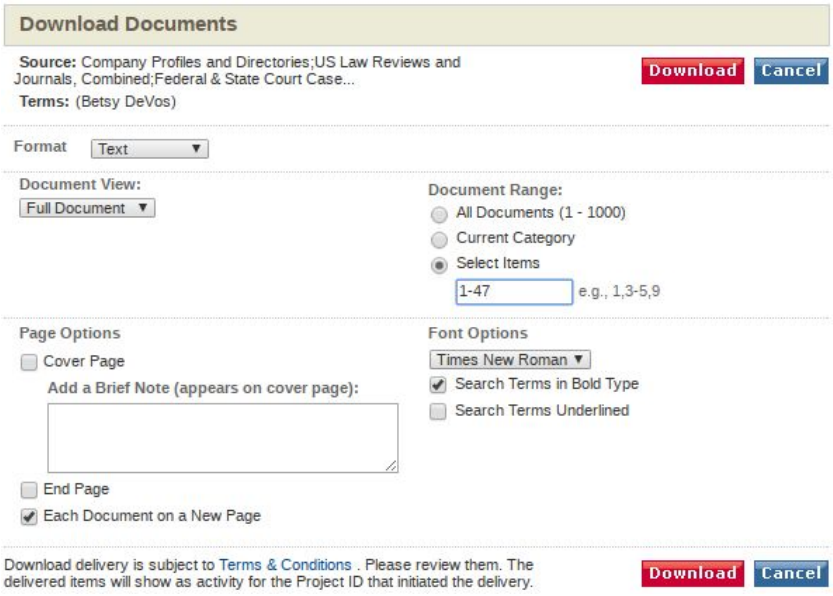
I’ll first build the New York Times part of my corpus. I’ll click on the New York Times option under the Newspapers on the left-hand column to narrow the search term. It appears that there are 47 documents returned in the New York Times category.



Once you have a list of search results you like, download the entire search query using this button.



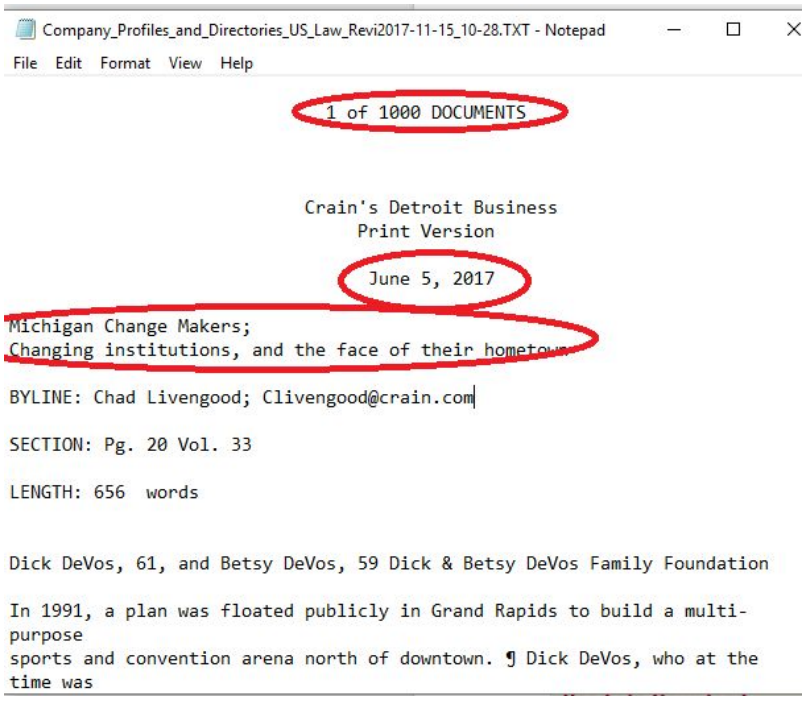
Choose the following download settings for our purposes. Note that it uses text format and I've selected all the documents in this search query (1-47). Once the settings have been configured, hit the download button. It will second to load.



Once the program finishes, you can click the link to download the file. The file will save to somewhere on your computer that you can open - it's just one big text file with all of your articles in it.

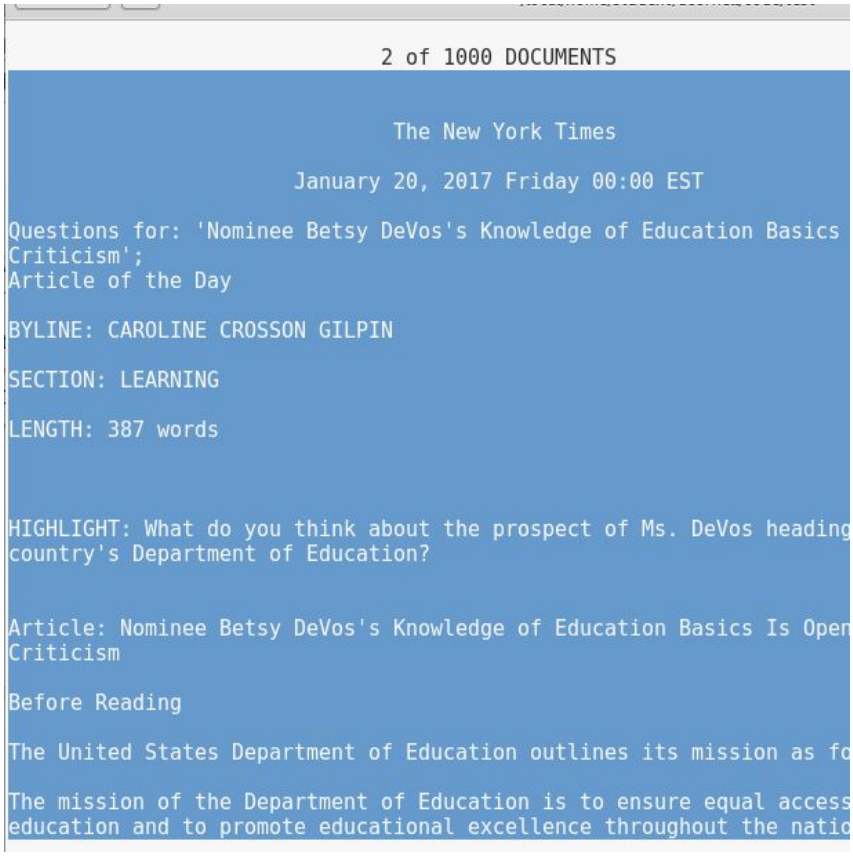
In my case, I wanted two different sub-corpora: one with New York Times, and one with Daily News. In that case, I downloaded the two separate search queries. The next step is to separate the individual articles out of the big file into their own files.

Open the new file and you'll see this.



All the documents in the query have been added into a single text file, so you'll need to separate them to where one article is one text file. You can do that by copy-pasting the articles into new text files. Every document in this text file starts with a "x of 1000 DOCUMENTS", so to get all of

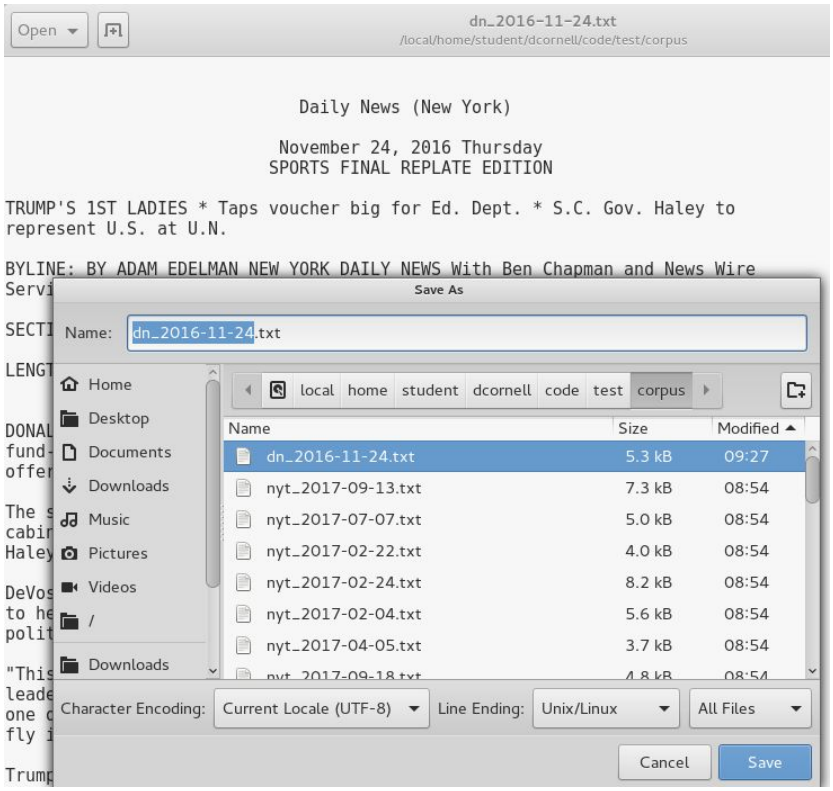
the document content, you should just copy-paste everything in between the headers for consecutive documents.



```
                                2 of 1000 DOCUMENTS

                                The New York Times

                            January 20, 2017 Friday 00:00 EST

Questions for: 'Nominee Betsy DeVos's Knowledge of Education Basics
Criticism';
Article of the Day

BYLINE: CAROLINE CROSSON GILPIN

SECTION: LEARNING

LENGTH: 387 words



HIGHLIGHT: What do you think about the prospect of Ms. DeVos heading
country's Department of Education?


Article: Nominee Betsy DeVos's Knowledge of Education Basics Is Open
Criticism

Before Reading

The United States Department of Education outlines its mission as fo

The mission of the Department of Education is to ensure equal access
education and to promote educational excellence throughout the natio
```

Open up a blank text file in notepad or whatever text editor you use. If you open those programs, it should start with a blank file that you can paste into. Paste in your article, then save the article with an appropriate filename.

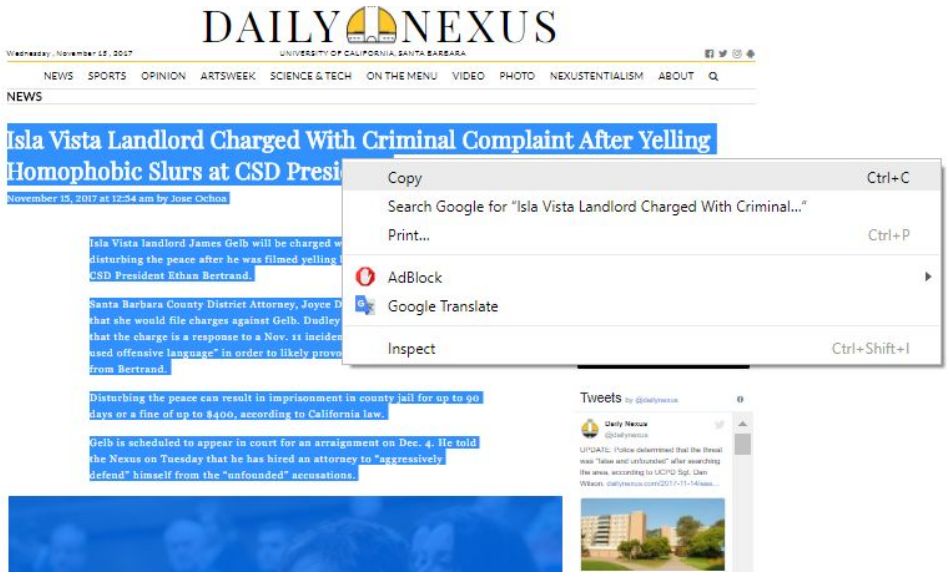
The best filenames include the source name/abbreviation (newspaper), a date (for chronological sorting), and the title of the article (or some other unique identifier). The above text file is used as an example. I didn't use any unique document identifier, but you may want to - especially if you have two documents from the same day.



The naming of the file is important - it is called ‘metadata’. It is data about the data we want to work with. When we go to use the topic modeling software, it will be especially important because it will help us sort through the documents for close reading or quantitative analysis.

Can I collect my data any other way?

Yes, but it all needs to end up in the same format: a series of text files with appropriate names. For instance, instead of copying from the Lexis Nexus download page, you can copy-paste an article from the web in the same way. If you copy into a program like ‘notepad’, it will just ignore the image (this is what we want).



Conclusion

In the end, the most important thing is to have a list of text files you can analyze, with filenames that are useful for your analysis (i.e. convenient names for sorting). It is recommended that you have at least around 100 text articles to perform topic modeling or sentiment analysis (although, feel free to use many more - the software can use as many as you can make).

After you have finished your corpus, upload it to Gauchospace and we'll run the topic modeling sofware on it. You can read the 'instructions overview' to understand how to analyze the data. Get the corpus collected as soon as possible so you can get started with the analysis.

If you have any technical questions, feel free to email dcornell@ucsb.edu.