

Application Programming Interfaces (API) for Web-Scraping & Text Analysis

Jeff Sternberg
Intro to Sociology
Laura Senier
Spring 2020



Northeastern University
NULab for Texts, Maps, and Networks

Workshop Agenda

- Learn about Web-Scraping & APIs for Data Collection
- Introduce and Explore the New York Times API
- Collect “Opioid Epidemic” News Coverage
- Conduct a Text Analysis of Opioid Epidemic News Coverage in Lexos

Slides, handouts, and data available at

<http://bit.ly/diti-spring2020-senier>

Learning Objectives

- Understand the definition and purpose of an API and web-scraping
- Understand the importance of API documentation
- Understand the affordances and limitations of using APIs to build a corpus
- Start to understand how to use digital tools to pull out novel insights and findings from text data

Discussion

Webster et al.'s “A critical content analysis of media reporting on opioids: The social construction of an epidemic”

- What did we learn from this article about the discourses and attitudes represented in news coverage of the opioid epidemic?
- Who are the actors involved in the opioid epidemic?
- What explanations for and solutions to the epidemic are given?
- How would an inductive approach differ from this study's deductive approach?

What is an API and what is web-scraping?

An API, or application programming interface, is a set of subroutine definitions, communication protocols, and tools for building software that ultimately allows applications to communicate with one another. An API may be for a web-based system, operating system, database system, computer hardware, or software library.

Web-scraping is the process of extracting large amounts of data from an internet source and downloading the data to a local repository. The scraping process can be done manually, but is usually automated by using software because of the large amount of data typically involved.

API Documentation

- When using APIs for web-scraping, it is necessary to refer to the API documentation and a link is usually found on the API homepage.
- Why?
 - While the concepts remain roughly the same, APIs differ and the syntax for accessing data can be very different.
 - You will likely need an API key, and the links for registering for the key will be found in the documentation.
 - There may be other unaccounted for differences and API specifics that require a close understanding of the API's structure.

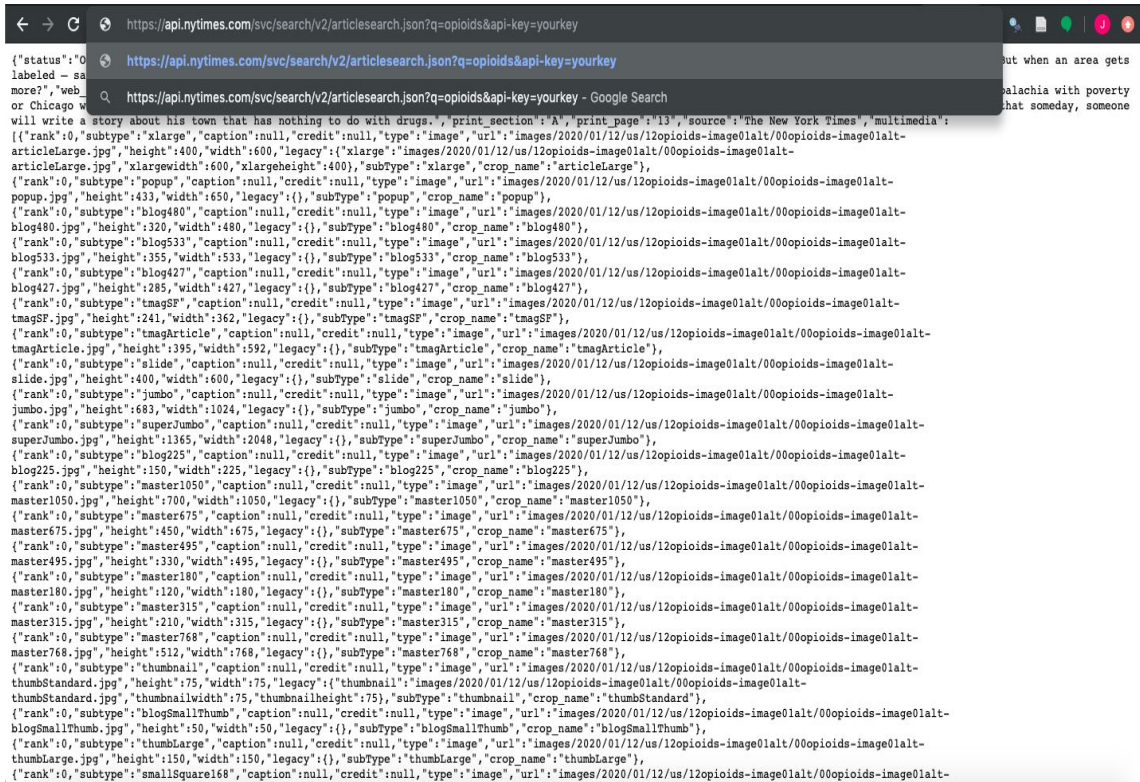
Popular APIs

- New York Times: <https://developer.nytimes.com/>
- Reddit: <https://www.reddit.com/dev/api/>
- IMDB: <http://www.omdbapi.com/>
- FBI: <https://crime-data-explorer.fr.cloud.gov/api>
 - Other Federal government APIs:
<https://api.data.gov/docs/>
- Twitter: <https://developer.twitter.com/en/docs.html>

New York Times API

- The New York Times has many different active [APIs](#) providing access to a variety of different text data sources, holding Articles, Movie Reviews, Book Reviews, User Comments, etc
- For our purposes, investigating news article coverage of the opioid epidemic, we will be utilizing the [New York Times Article Search API](#)

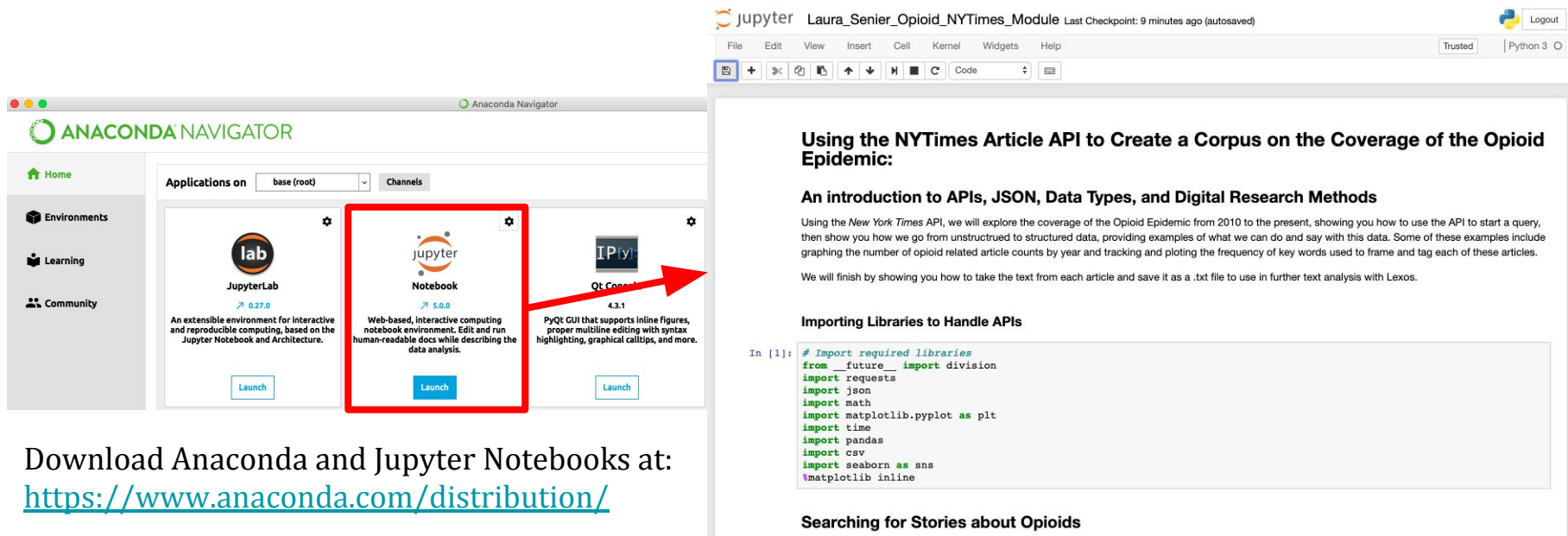
Querying Using URL and Browser



```
{
  "status": "OK",
  "labeled": "sa",
  "more": "web",
  "or": "Chicago",
  "will write a story about his town that has nothing to do with drugs.",
  "print_section": "A",
  "print_page": "13",
  "source": "The New York Times",
  "multimedia": [
    {
      "rank": 0,
      "subtype": "xlarge",
      "caption": null,
      "credit": null,
      "type": "image",
      "url": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-articleLarge.jpg",
      "height": 400,
      "width": 600,
      "legacy": {
        "xlarge": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-articleLarge.jpg",
        "xlargewidth": 600,
        "xlargeheight": 400
      },
      "subtype": "xlarge",
      "crop_name": "articleLarge"
    },
    {
      "rank": 0,
      "subtype": "popup",
      "caption": null,
      "credit": null,
      "type": "image",
      "url": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-popup.jpg",
      "height": 433,
      "width": 650,
      "legacy": {
        "popup": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-popup.jpg",
        "popupwidth": 650,
        "popupheight": 433
      },
      "subtype": "popup",
      "crop_name": "popup"
    },
    {
      "rank": 0,
      "subtype": "blog480",
      "caption": null,
      "credit": null,
      "type": "image",
      "url": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-blog480.jpg",
      "height": 320,
      "width": 480,
      "legacy": {
        "blog480": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-blog480.jpg",
        "blog480width": 480,
        "blog480height": 320
      },
      "subtype": "blog480",
      "crop_name": "blog480"
    },
    {
      "rank": 0,
      "subtype": "blog533",
      "caption": null,
      "credit": null,
      "type": "image",
      "url": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-blog533.jpg",
      "height": 355,
      "width": 533,
      "legacy": {
        "blog533": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-blog533.jpg",
        "blog533width": 533,
        "blog533height": 355
      },
      "subtype": "blog533",
      "crop_name": "blog533"
    },
    {
      "rank": 0,
      "subtype": "blog427",
      "caption": null,
      "credit": null,
      "type": "image",
      "url": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-blog427.jpg",
      "height": 285,
      "width": 427,
      "legacy": {
        "blog427": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-blog427.jpg",
        "blog427width": 427,
        "blog427height": 285
      },
      "subtype": "blog427",
      "crop_name": "blog427"
    },
    {
      "rank": 0,
      "subtype": "tmsqSF",
      "caption": null,
      "credit": null,
      "type": "image",
      "url": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-tmsqSF.jpg",
      "height": 241,
      "width": 362,
      "legacy": {
        "tmsqSF": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-tmsqSF.jpg",
        "tmsqSFwidth": 362,
        "tmsqSFheight": 241
      },
      "subtype": "tmsqSF",
      "crop_name": "tmsqSF"
    },
    {
      "rank": 0,
      "subtype": "tmsqArticle",
      "caption": null,
      "credit": null,
      "type": "image",
      "url": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-tmsqArticle.jpg",
      "height": 395,
      "width": 592,
      "legacy": {
        "tmsqArticle": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-tmsqArticle.jpg",
        "tmsqArticlewidth": 592,
        "tmsqArticleheight": 395
      },
      "subtype": "tmsqArticle",
      "crop_name": "tmsqArticle"
    },
    {
      "rank": 0,
      "subtype": "slide",
      "caption": null,
      "credit": null,
      "type": "image",
      "url": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-slide.jpg",
      "height": 400,
      "width": 600,
      "legacy": {
        "slide": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-slide.jpg",
        "slidewidth": 600,
        "slideheight": 400
      },
      "subtype": "slide",
      "crop_name": "slide"
    },
    {
      "rank": 0,
      "subtype": "jumbo",
      "caption": null,
      "credit": null,
      "type": "image",
      "url": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-jumbo.jpg",
      "height": 683,
      "width": 1024,
      "legacy": {
        "jumbo": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-jumbo.jpg",
        "jumbowidth": 1024,
        "jumboheight": 683
      },
      "subtype": "jumbo",
      "crop_name": "jumbo"
    },
    {
      "rank": 0,
      "subtype": "superJumbo",
      "caption": null,
      "credit": null,
      "type": "image",
      "url": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-superJumbo.jpg",
      "height": 1365,
      "width": 2048,
      "legacy": {
        "superJumbo": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-superJumbo.jpg",
        "superJumbowidth": 2048,
        "superJumboheight": 1365
      },
      "subtype": "superJumbo",
      "crop_name": "superJumbo"
    },
    {
      "rank": 0,
      "subtype": "blog225",
      "caption": null,
      "credit": null,
      "type": "image",
      "url": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-blog225.jpg",
      "height": 150,
      "width": 225,
      "legacy": {
        "blog225": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-blog225.jpg",
        "blog225width": 225,
        "blog225height": 150
      },
      "subtype": "blog225",
      "crop_name": "blog225"
    },
    {
      "rank": 0,
      "subtype": "master1050",
      "caption": null,
      "credit": null,
      "type": "image",
      "url": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-master1050.jpg",
      "height": 700,
      "width": 1050,
      "legacy": {
        "master1050": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-master1050.jpg",
        "master1050width": 1050,
        "master1050height": 700
      },
      "subtype": "master1050",
      "crop_name": "master1050"
    },
    {
      "rank": 0,
      "subtype": "master675",
      "caption": null,
      "credit": null,
      "type": "image",
      "url": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-master675.jpg",
      "height": 450,
      "width": 675,
      "legacy": {
        "master675": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-master675.jpg",
        "master675width": 675,
        "master675height": 450
      },
      "subtype": "master675",
      "crop_name": "master675"
    },
    {
      "rank": 0,
      "subtype": "master495",
      "caption": null,
      "credit": null,
      "type": "image",
      "url": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-master495.jpg",
      "height": 330,
      "width": 495,
      "legacy": {
        "master495": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-master495.jpg",
        "master495width": 495,
        "master495height": 330
      },
      "subtype": "master495",
      "crop_name": "master495"
    },
    {
      "rank": 0,
      "subtype": "master180",
      "caption": null,
      "credit": null,
      "type": "image",
      "url": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-master180.jpg",
      "height": 120,
      "width": 180,
      "legacy": {
        "master180": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-master180.jpg",
        "master180width": 180,
        "master180height": 120
      },
      "subtype": "master180",
      "crop_name": "master180"
    },
    {
      "rank": 0,
      "subtype": "master315",
      "caption": null,
      "credit": null,
      "type": "image",
      "url": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-master315.jpg",
      "height": 210,
      "width": 315,
      "legacy": {
        "master315": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-master315.jpg",
        "master315width": 315,
        "master315height": 210
      },
      "subtype": "master315",
      "crop_name": "master315"
    },
    {
      "rank": 0,
      "subtype": "master768",
      "caption": null,
      "credit": null,
      "type": "image",
      "url": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-master768.jpg",
      "height": 512,
      "width": 768,
      "legacy": {
        "master768": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-master768.jpg",
        "master768width": 768,
        "master768height": 512
      },
      "subtype": "master768",
      "crop_name": "master768"
    },
    {
      "rank": 0,
      "subtype": "thumbStandard",
      "caption": null,
      "credit": null,
      "type": "image",
      "url": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-thumbStandard.jpg",
      "height": 75,
      "width": 75,
      "legacy": {
        "thumbStandard": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-thumbStandard.jpg",
        "thumbStandardwidth": 75,
        "thumbStandardheight": 75
      },
      "subtype": "thumbStandard",
      "crop_name": "thumbStandard"
    },
    {
      "rank": 0,
      "subtype": "blogSmallThumb",
      "caption": null,
      "credit": null,
      "type": "image",
      "url": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-blogSmallThumb.jpg",
      "height": 50,
      "width": 50,
      "legacy": {
        "blogSmallThumb": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-blogSmallThumb.jpg",
        "blogSmallThumbwidth": 50,
        "blogSmallThumbheight": 50
      },
      "subtype": "blogSmallThumb",
      "crop_name": "blogSmallThumb"
    },
    {
      "rank": 0,
      "subtype": "thumbLarge",
      "caption": null,
      "credit": null,
      "type": "image",
      "url": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-thumbLarge.jpg",
      "height": 150,
      "width": 150,
      "legacy": {
        "thumbLarge": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-thumbLarge.jpg",
        "thumbLargewidth": 150,
        "thumbLargeheight": 150
      },
      "subtype": "thumbLarge",
      "crop_name": "thumbLarge"
    },
    {
      "rank": 0,
      "subtype": "smallSquare168",
      "caption": null,
      "credit": null,
      "type": "image",
      "url": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-smallSquare168.jpg",
      "height": 168,
      "width": 168,
      "legacy": {
        "smallSquare168": "images/2020/01/12/us/12opioids-image01alt/00opioids-image01alt-smallSquare168.jpg",
        "smallSquare168width": 168,
        "smallSquare168height": 168
      },
      "subtype": "smallSquare168",
      "crop_name": "smallSquare168"
    }
  ]
}
```

- We access the API through our web-browser, giving it the API URL with our query as q=opioids and our API key after that.
- This returns us a mess of a json file presented in html in our browser.
- What do we do with this? How do we make it useable?

The Answer? Parsing using Python and Jupyter Notebooks!



The image shows two overlapping screenshots. The background screenshot is the Anaconda Navigator application window. On the left is a sidebar with 'Home', 'Environments', 'Learning', and 'Community'. The main area is titled 'Applications on base (root)' and shows three options: 'JupyterLab' (0.27.0), 'Jupyter Notebook' (5.0.0), and 'Qt Console' (4.3.1). The 'Jupyter Notebook' option is highlighted with a red rectangle. A red arrow points from this rectangle to the foreground screenshot. The foreground screenshot is a Jupyter Notebook interface. The top bar shows the notebook name 'Laura_Senier_Opioid_NYTimes_Module' and a 'Last Checkpoint: 9 minutes ago (autosaved)' status. The menu bar includes 'File', 'Edit', 'View', 'Insert', 'Cell', 'Kernel', 'Widgets', and 'Help'. The toolbar contains icons for file operations and code execution. The main content area has the title 'Using the NYTimes Article API to Create a Corpus on the Coverage of the Opioid Epidemic:' followed by a subtitle 'An introduction to APIs, JSON, Data Types, and Digital Research Methods'. The text describes using the New York Times API to explore opioid epidemic coverage from 2010 to the present. Below this is a section titled 'Importing Libraries to Handle APIs' which contains a code cell with the following Python code:

```
In [1]: # Import required libraries
from __future__ import division
import requests
import json
import math
import matplotlib.pyplot as plt
import time
import pandas
import csv
import seaborn as sns
%matplotlib inline
```

Below the code cell is the section title 'Searching for Stories about Opioids'.

Download Anaconda and Jupyter Notebooks at:

<https://www.anaconda.com/distribution/>

Link to the [Jupyter Notebook](#), follow along!

Computational Text Analysis

Computational text analysis refers to an array of methods that can be used to “read” texts with a computer. This form of analysis can range from basic word frequency counts to more advanced techniques like machine learning.

Text analysis is often used on a **corpus**, or a collection of multiple texts, and provides a glimpse into patterns across the texts. Some people also perform text analysis on larger individual documents, like novels.

Why Computational Text Analysis?

Computational text analysis can help us analyze a **ton** of data and discover **patterns** in texts.

Particular disciplines care **deeply** about the language used and how this language may reach intended audiences. Text analysis provides another method for approaching these discourses.

Notes on Creating a Corpus (in General)

1. Choose the texts you want to include in your corpus
2. Create a folder on your computer titled “corpus” or something even more specific
3. Copy and paste your texts into a **plain text editor** (on Macs: Text Edit; on Windows: Notepad)
 - a. Mac users, you will need to make your Text Edit into a plain text editor. Open Text Edit, go to Preferences, and make sure “plain text” is selected
4. Save each text as a different plain text file (with a .txt extension). Name your files so you know what is in them!

Lexos

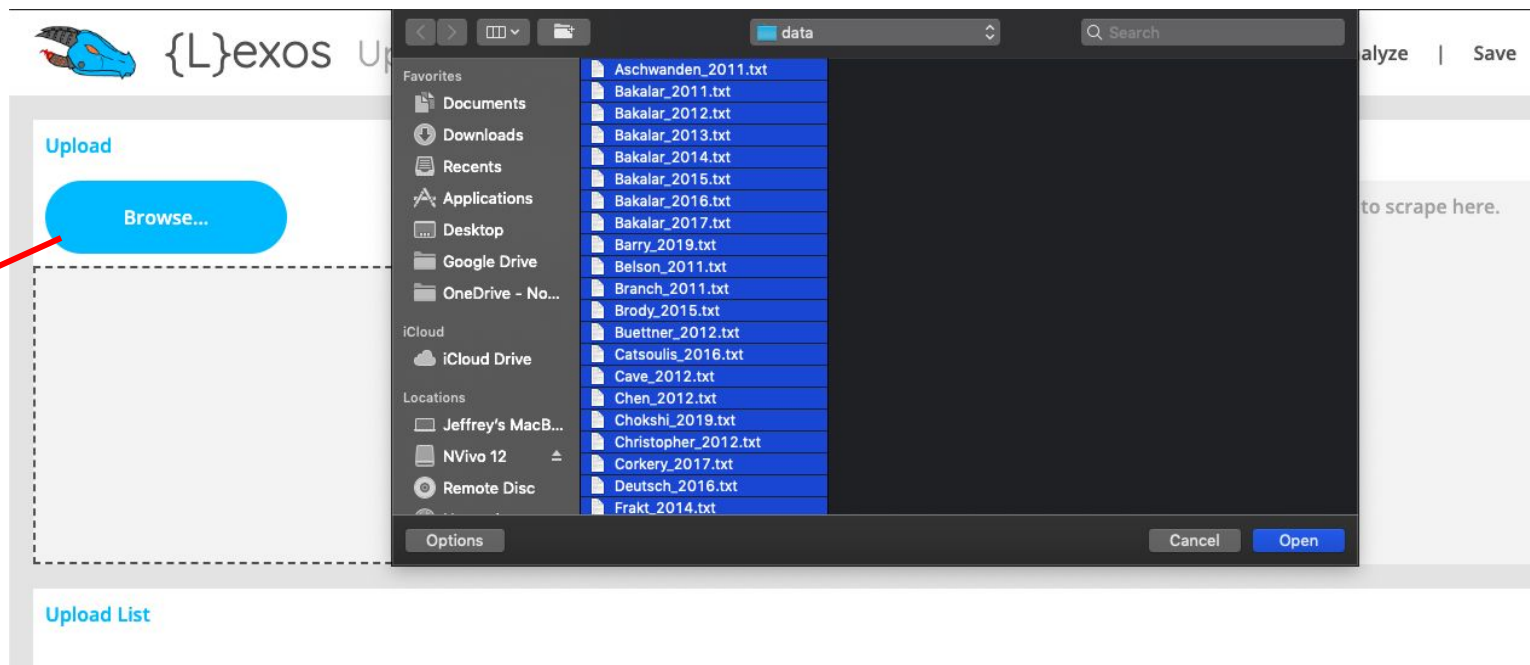
Lexos: <http://lexos.wheatoncollege.edu/upload>

Lexos provides a step-by-step guide for corpus uploading, preparation, and analysis.

- **Upload:** upload your corpus (your separate .txt files)
- **Manage:** select the files you want to prepare and analyze
- **Prepare:** prepare your corpus for analysis
- **Visualize:** create visualizations of patterns across your corpus or in single texts
- **Analyze:** analyze your corpus, including comparing texts

Lexos: Upload

Click Browse
and select your
entire corpus
(or drag and
drop)



Lexos: Manage



{L}exos Manageⁱ

Upload **Manage** Prepare Visualize Analyze | Save Reset | Help

Make sure all the documents in the corpus you want to use are selected (blue = selected, gray = not selected)

Active	#	Document	Class	Source	Excerpt	Download	?
<input checked="" type="radio"/>	1	Aschwanden_2011		Aschwanden_2011.txt	More than 20 studies, including a large analysis of data on more than 200,000 children, have produced results that link acetami... ...alysis of data on more than 200,000 children, have produced results that link acetaminophen use to an increased risk of asthma.		
<input checked="" type="radio"/>	2	Bakalar_2011		Bakalar_2011.txt	A study found that 49 percent of patients over age 75 were given pain medication, compared with about 65 percent of those under... ...found that 49 percent of patients over age 75 were given pain medication, compared with about 65 percent of those under age 75.		
<input checked="" type="radio"/>	3	Bakalar_2012		Bakalar_2012.txt	The drug accounted for 1.7 percent of the 257 million prescriptions written in 2009 for opioid pain relievers, but it was invol... ...he 257 million prescriptions written in 2009 for opioid pain relievers, but it was involved in 31.4 percent of overdose deaths.		

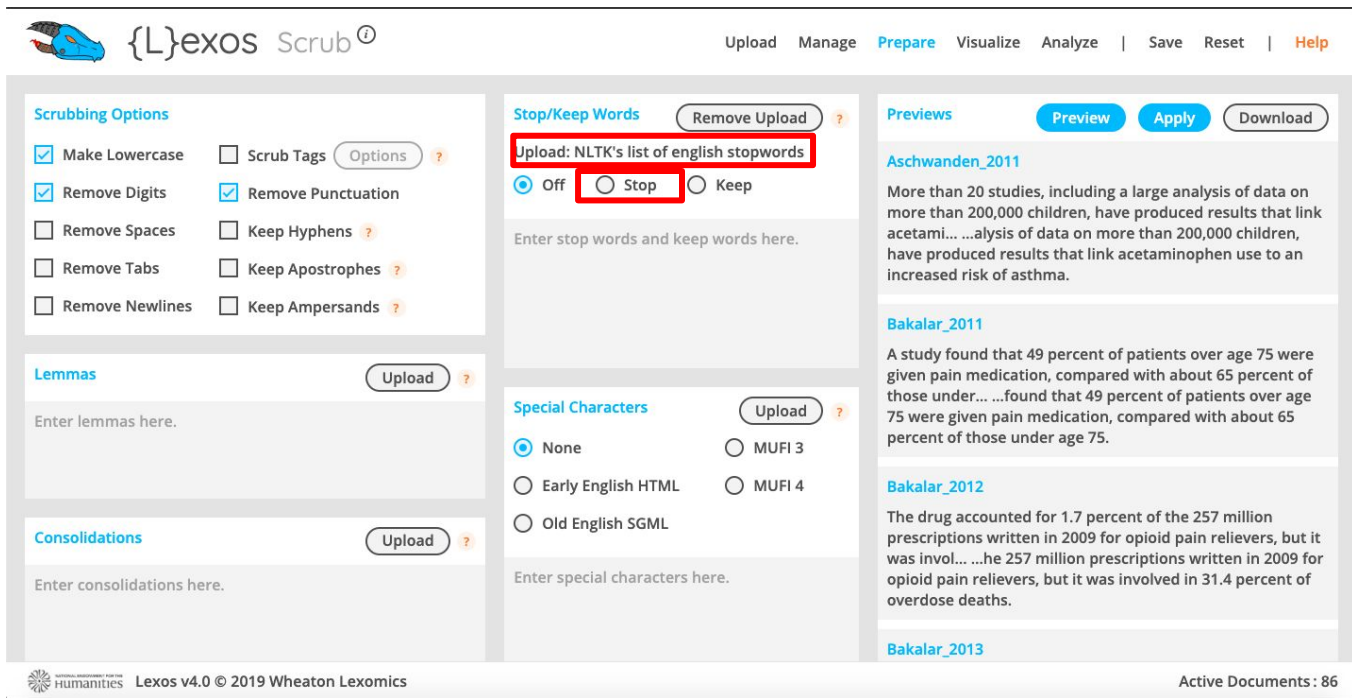
Lexos: Prepare (scrub)

Lexos demonstrates the different options you have for preparing your corpus. By “scrubbing,” you are transforming the texts in your corpus and making choices that will impact your results. Here are some possibilities:

- **Make Lowercase:** make all your letters lowercase. Even though you know “A” and “a” are the same letter, the computer treats these as two separate characters. Lowercasing removes this distinction.
- **Remove Punctuation:** remove punctuation, which may influence your results.
- **Stop/Keep Words:** remove a list of words. Usually these would be **stopwords**, or the most common words in a language (English: the, a she, her, it, him, they, etc).
- **Lemmas:** standardize to the *stem* of word. For example, you can stem all forms of talk: talking, talked, talks, etc. to “talk”

Lexos: Removing Stopwords

Get a list of English stopwords here: <https://gist.github.com/sebleier/554280> (we also sent you a .txt file). Copy and paste the stopwords (or upload the .txt file) into the “Stop/Keep Words” box then select “Stop”



The screenshot shows the Lexos Scrub interface. The top navigation bar includes 'Upload', 'Manage', 'Prepare', 'Visualize', 'Analyze', 'Save', 'Reset', and 'Help'. The main interface is divided into several sections:

- Scrubbing Options:** Includes checkboxes for 'Make Lowercase', 'Remove Digits', 'Remove Spaces', 'Remove Tabs', 'Remove Newlines', 'Scrub Tags', 'Remove Punctuation', 'Keep Hyphens', 'Keep Apostrophes', and 'Keep Ampersands'. There is an 'Options' button with a question mark.
- Lemmas:** A text input field with an 'Upload' button and a question mark.
- Consolidations:** A text input field with an 'Upload' button and a question mark.
- Stop/Keep Words:** This section is highlighted with a red box. It contains a 'Remove Upload' button, a text input field with the text 'Upload: NLTK's list of english stopwords', and three radio buttons: 'Off' (selected), 'Stop' (highlighted with a red box), and 'Keep'. Below the radio buttons is a text input field labeled 'Enter stop words and keep words here.'
- Special Characters:** Includes radio buttons for 'None' (selected), 'Early English HTML', 'Old English SGML', 'MUFI 3', and 'MUFI 4'. There is an 'Upload' button and a question mark. Below the radio buttons is a text input field labeled 'Enter special characters here.'
- Previews:** Includes 'Preview', 'Apply', and 'Download' buttons. It shows three preview cards for documents: 'Aschwanden_2011', 'Bakalar_2011', and 'Bakalar_2012'.

The footer of the interface includes the 'Lexos v4.0 © 2019 Wheaton Lexomics' logo and the text 'Active Documents : 86'.

Lexos: Applying your Preparations

Once you have made decisions about your preparations, click “**Apply**” and wait a few minutes. Because the program is going through each document and completing all the processes you selected, it needs some time. Then, you will see the final results of your preparation! You can also **download** your new corpus.

BEFORE PREP

Previews

Preview

Apply

Download

Aschwanden_2011

More than 20 studies, including a large analysis of data on more than 200,000 children, have produced results that link acetami... ..analysis of data on more than 200,000 children, have produced results that link acetaminophen use to an increased risk of asthma.

Bakalar_2011

A study found that 49 percent of patients over age 75 were given pain medication, compared with about 65 percent of those under... ..found that 49 percent of patients over age 75 were given pain medication, compared with about 65 percent of those under age 75.

Bakalar_2012

The drug accounted for 1.7 percent of the 257 million prescriptions written in 2009 for opioid pain relievers, but it was invol... ..he 257 million prescriptions written in 2009 for opioid pain relievers, but it was involved in 31.4 percent of overdose deaths.

AFTER PREP

Previews

Preview

Apply

Download

Aschwanden_2011

studies including large analysis data children produced results link acetaminophen use increased risk asthmathe hypothesismore studies including large analysis data children produced results link acetaminophen use increased risk asthma

Bakalar_2011

study found percent patients age given pain medication compared percent age older people go emergency room pain less likely ge... ..people similar levels distress new analysis founda study found percent patients age given pain medication compared percent age

Bakalar_2012

drug accounted percent million prescriptions written opioid pain relievers involved percent overdose deathsmethadone accounted... ..esearchers foundthe drug accounted percent million prescriptions written opioid pain relievers involved percent overdose deaths

Lexos: Statistics



{L}exos Statistics ⓘ

Upload Manage Prepare Visualize **Analyze** | Save Reset | Help

Generate All

Tokenize ⓘ

- ☒ By Tokens
- ☐ By Characters

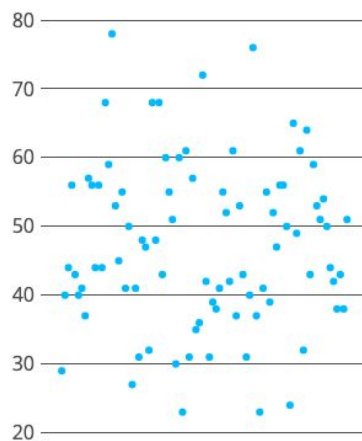
Grams

1

Cull ⓘ

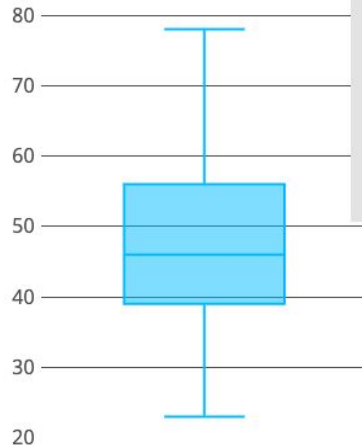
- ☐ Use the top 100 terms ⓘ
- ☐ Must be in 1 documents ⓘ

Document Sizes



PNG

SVG



Statistics ⓘ

- Dendrogram
- K-Means
- Consensus Tree
- Similarity Query
- Top Words
- Content Analysis

Statistics

on: 12.05
ge: 17

st

on_2011 (small),
no_last_name_8_6_2018 (small),
Chen_2012 (large), Louis_2015
(large), no_last_name_7_4_2017

Interquartile Range Test

No Anomalies

Document Statistics

Order: ☒ Ascending ☐ Descending

Generate

Download

Name

Single-Occurrence Terms

Total Terms

Vocabulary Density

Distinct Terms

Aschwanden_2011

3

29

0.552

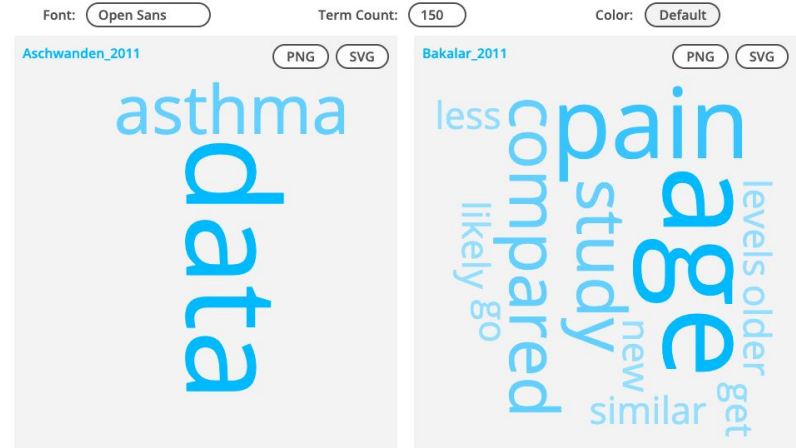
16

Lexos: Visualize



Multi Cloud: visualize wordclouds for each individual document/text

Word Cloud:
visualize a
wordcloud across
the entire corpus.



Lexos: Top Words



{L}exos Top Words ⓘ

Upload Manage Prepare Visualize **Analyze** | Save Reset | Help

Comparison Method ⓘ

- ☒ Each Document to the Corpus ⓘ
- ☐ Each Document to Other Classes ⓘ
- ☐ Each Class to Other Classes ⓘ

Tokenize ⓘ

- ☒ By Tokens
- ☐ By Characters

Grams:

Cull ⓘ

- ☐ Use the top
- ☐ Must be in

Statistics
Dendrogram
K-Means
Consensus Tree
Similarity Query
Top Words
Content Analysis

Class Divisions ⓘ

Top Words

Generate

Download ⓘ

Document "Aschwanden_2011" Compared To The Corpus

produced	11.7505
results	11.7505
analysis	10.4734
data	10.4734
link	9.5275
asthma	8.3068

Document "Bakalar_2011" Compared To The Corpus

age	11.9541
percent	9.1888
medication	8.1181
compared	8.0677
given	8.0677
study	7.4326

Document "Bakalar_2012" Compared To The Corpus

accounted	11.6428
relievers	10.7411
written	9.504
percent	8.7298
involved	8.4551
million	7.0688

Lexos: K-Means Clustering



{L}exos K-Meansⁱ

[Upload](#)[Manage](#)[Prepare](#)[Visualize](#)[Analyze](#)[Save](#)[Reset](#)[Help](#)

Options

Clusters

☒ Voronoi

☐ 2D Scatter

☐ 3D Scatter

Advanced

☒ K-Means++ ☐ Random

Maximum Iterations

Different Centroids

Relative Tolerance

Tokenize

☒ By Tokens

☐ By Characters

Grams:

Normalize

☒ Proportional

☐ Raw

☐ TF-IDF

Statistics

☐ Dendrogram

☒ K-Means

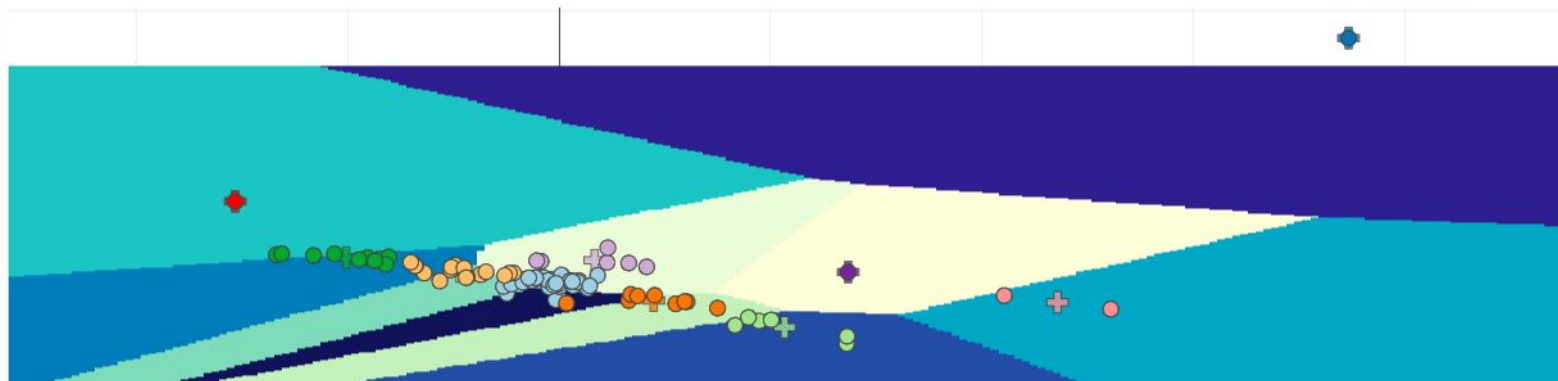
☐ Consensus Tree

☐ Similarity Query

☐ Top Words

☐ Content Analysis

K-Means

[Generate](#)[PNG](#)[SVG](#)[CSV](#)

- + Centroid 1
- + Centroid 2
- + Centroid 3
- + Centroid 4
- + Centroid 5
- + Centroid 6
- + Centroid 7
- + Centroid 8
- + Centroid 9
- + Centroid 10
- Cluster 1
- Cluster 2

Lexos: Save or Reset Your Results

Lexos allows you to **save** your results as a Lexos file. If you do this, you can re-upload the Lexos file any time to access your cleaned-up corpus as well as the different analyses you've done.

You can also save individual visualizations as images (PNGs).

Finally, if you want to start over, you can “Reset” your Lexos dashboard.

Contact and Resources

If you have any questions, contact me at:

Jeff Sternberg

DITI Research Fellow

sternberg.je@husky.neu.edu

Garrett Morrow

DITI Research Fellow

morrow.g@husky.neu.edu

Slides and data available at <http://bit.ly/diti-spring2020-senier>

Sign up for office hours at <https://calendly.com/sternberg-je/15min>