GALE
A Cengage Company

No Fear Digital Humanities!

Text Mining Historical Documents with the Gale Digital Scholar Lab

10 March 2022

Masaki Morisawa
Senior Product Manager, Gale Asia

About Gale

GALE
A Cengage Company

- HQ in Farmington Hills, Michigan, USA
- One of the foremost library reference publishers
- Has many well-established imprints, including:
  - Gale
  - Charles Scribner's Sons
  - Macmillan Reference USA
- Publishing formats include:
  - **Print** library reference, such as thematic encyclopedias, annual directories, literary biographies/criticisms, etc.
  - **eBook** versions of print publications and an eBook platform
  - Aggregated **Journal databases**
  - **Subject-specific Databases** combining various content together
  - **Microform** collections and serials
  - **Gale Primary Sources**: digital archives of historical material

Today’s Agenda

- What are Gale Primary Sources?
- What is the Gale Digital Scholar Lab?
- Gale Digital Scholar Lab Workflow and Tools
- Live Demonstration
- Which Gale Primary Sources are available to HKBU?
- Wrap up

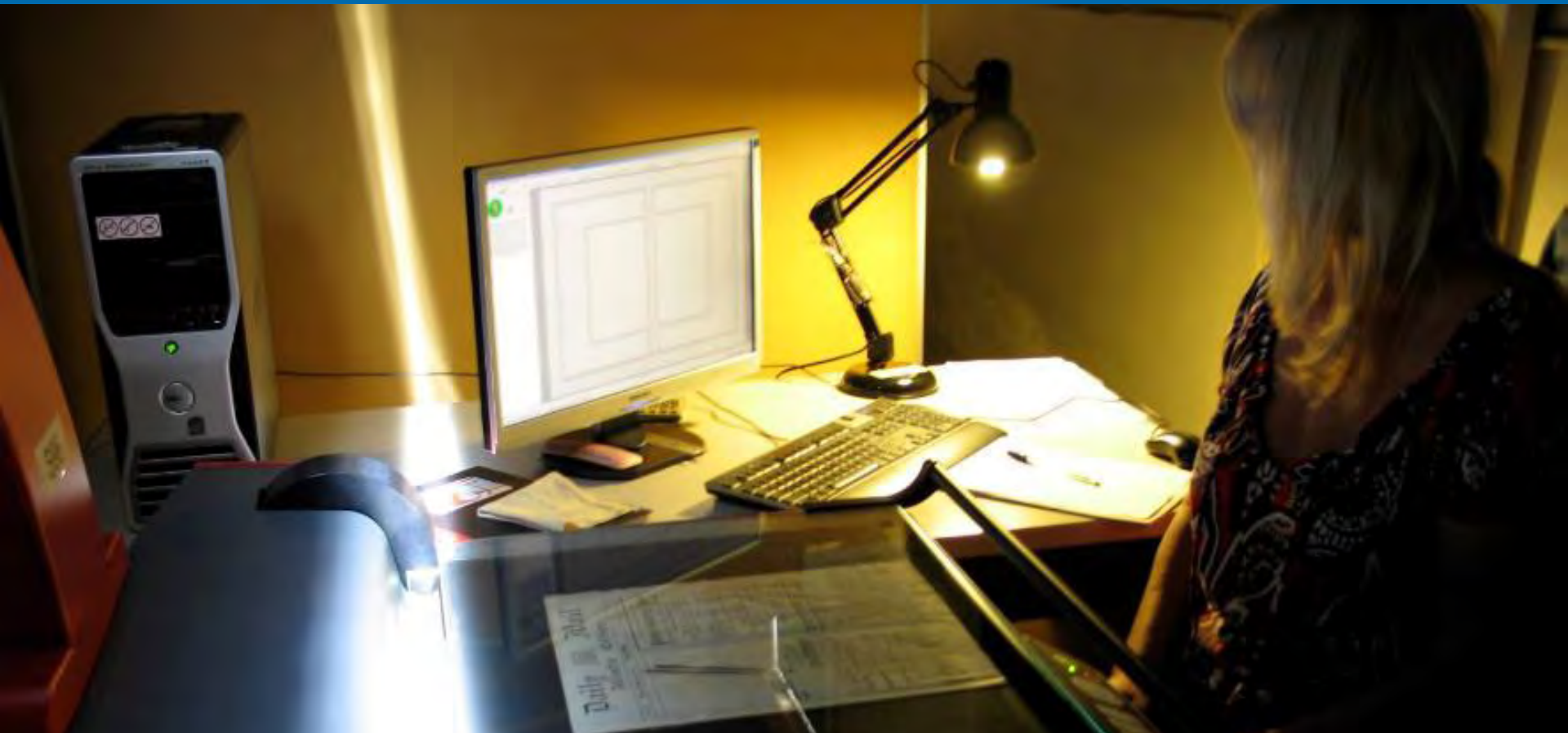
What are Gale Primary Sources?

Gale Primary Sources

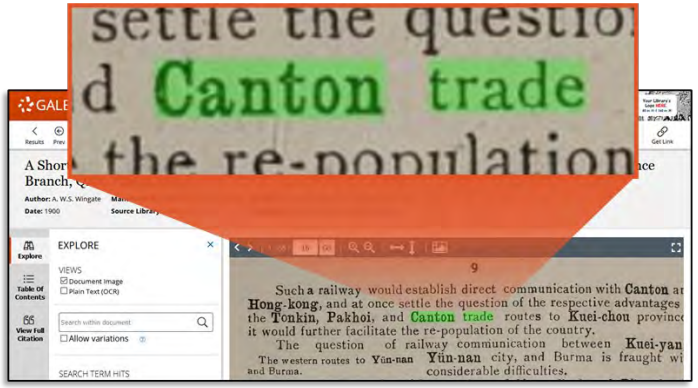
- Gale has spent the past 45 years building one of the world's largest scholarly primary source online libraries, spanning 560 years of global history
- More than 200 digital collections, comprising more than 179 million pages, including:
  - **Nineteenth Century Collections Online**
  - **The Times Digital Archive**
  - **China and the Modern World**
  - **U.S. Declassified Documents Online**
  - **Archives of Sexuality and Gender**
  - ***And many more!***

Examples of Original Holding Institutions

We scan and digitize valuable historical works



Enable the discovery of hidden words

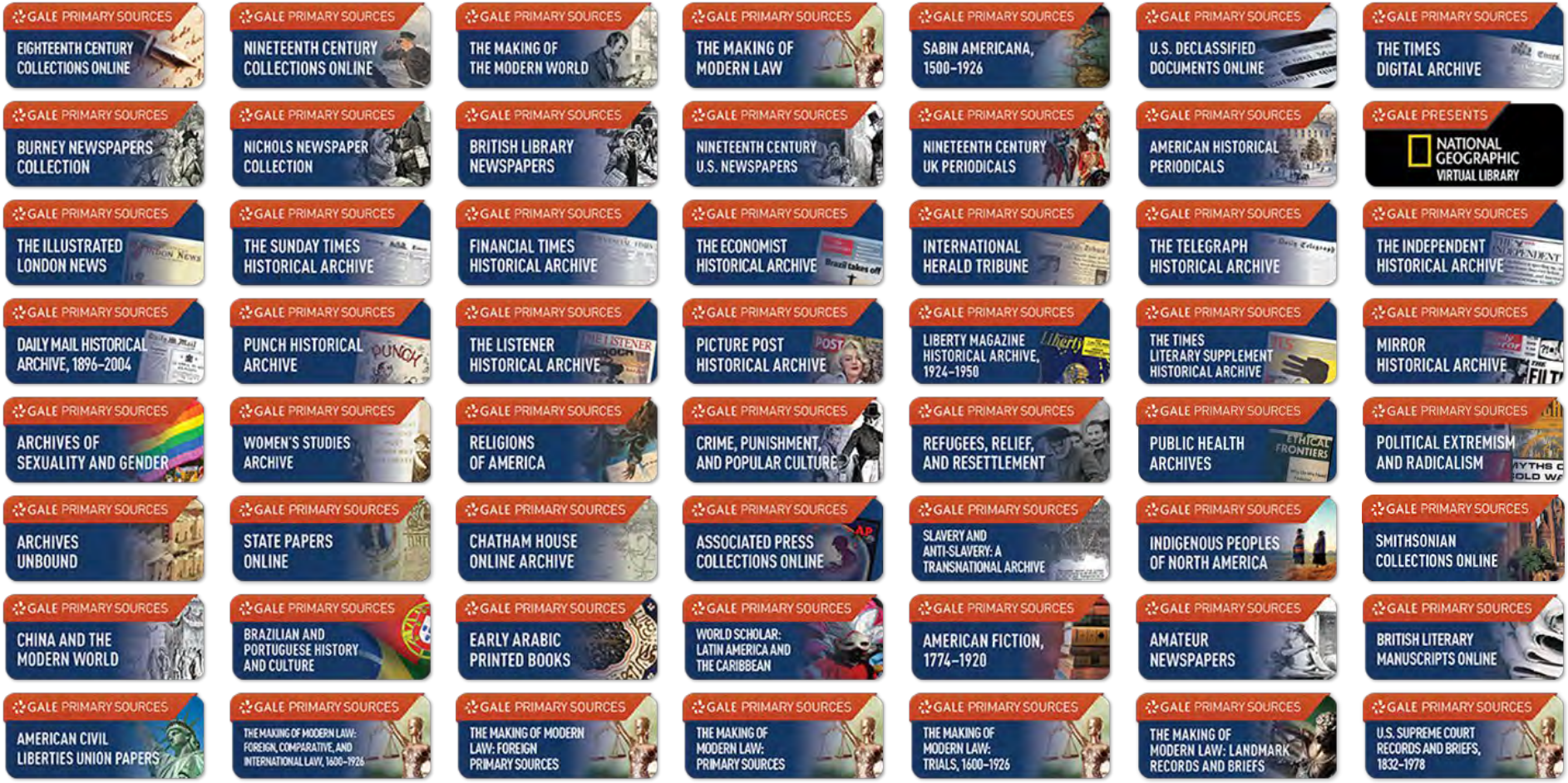


OCR = Optical Character Recognition

HTR = Handwritten Text Recognition*

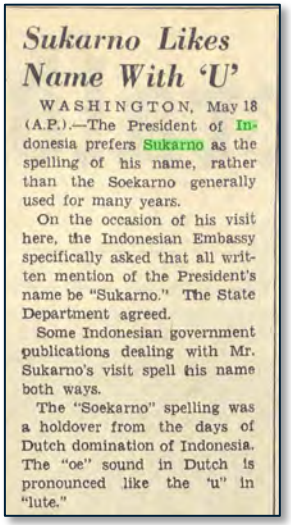
* HTR available in parts of *China and the Modern World* only

We have *many* such products



What is the Gale Digital Scholar Lab?

Some Researchers are Interested in the OCR Text behind the Images



Sukarno Likes Name With 'U'

WASHINGTON, May 18 (A.P.)—The President of Indonesia prefers Sukarno as the spelling of his name, rather than the Soekarno generally used for many years.

On the occasion of his visit here, the Indonesian Embassy specifically asked that all written mention of the President's name be "Sukarno." The State Department agreed.

Some Indonesian government publications dealing with Mr. Sukarno's visit spell his name both ways.

The "Soekarno" spelling was a holdover from the days of Dutch domination of Indonesia. The "oe" sound in Dutch is pronounced like the "u" in "lute."

Page Images (OCR not seen)

Sukarno Likes Name with 'U' Sukarno Likes Name With âUâ,WASHINGTON, May 18 (A.P.).âThe President of Indonesia prefers Sukarno as the spelling of his name, rather than the Soekarno generally used for many years. On the occasion of his visit here, the Indonesian Embassy specifically asked that all written mention of the President's name be âSukarno.â Hie State Department agreed. Some Indonesian government publications dealing with Mr. Sukarnoâs visit spell his name both ways. The âSoekarnoâ spelling was a holdover from the days of Dutch domination of Indonesia. The "oeâ sound in Dutch is pronounced like the âuâ in ālute.â

OCR Text (generated from image)

What If you could combine hundreds of text files like these…

Sukarno Likes Name with 'U' Sukarno Likes Name With âUâ,WASHINGTON, May 18 (A.P.).âThe President of Indonesia prefers Sukarno as the spelling of his name, rather than the Soekarno generally used for many years. On the occasion of his visit here, the Indonesian Embassy specifically asked that all written mention of the President's name be âSukarno.â Hie State Department agreed. Some Indonesian government publications dealing with Mr. Sukarnoâs visit spell his name both ways. The âSoekarnoâ spelling was a holdover from the days of Dutch domination of Indonesia. The "oeâ sound in Dutch is pronounced like the âuâ in âlute.â

+

Eisenhower, Sukarno Agree on 'Freedom' 881 mm President Sukarno of Indonesia, and Mrs. Richard M. Nixon at a Washington reception. Eisenhower, Sukarno ree on 'Freedomâ Ag.By the United Press WASHINGTON, May 20.âPresident Eisenhower said Friday night that the people of the United States must never forget âthat unless we defend the freedom of the ottier fellow we are endangering our own.â Mr. Eisenhower, in a toast to visiting Indonesian President Sukarno, said that he wanted to express his gratitude to Mr. Sukarno for recalling a basic American responsibility. âI want to thank you for what you have reminded us of,â

+

Sukarno Hailed on Broadway Sukarno Hailed On Broadway,NEW YORK, May 23 (A.P.). âNew York today accorded President Sukarno of Indonesia the traditional honor for visiting dignitariesâa ticker tape parade along Lower Broadway through the skyscrapers of the financial district. The President stood in the rear of an open car for the 15-minute drive from Battery Park to City Hall. A light rain began to fall just before the parade started, but even so police estimated the crowd along the line of march and in City Hall Plaza at 80,000. Many persons along the parade route waved the red and white flag of Indonesia. Mr. Sukarno waved repeatedly. Arriving at City Hall, Mr. Sukarno was greeted by Mayor Robert F. Wagner. Both men stood at attention as the national anthems of the United States and Indonesia were played.

+

Sukarno and Pope Hold Talk Sukarno And Pope Hold Talk By Barrett McGurn From the Herald Tribune Bureau g) 1956, New York Herald Tribune, Inc.,ROME, June 13.âPope Pius XII promised President Achmed Sukarno of Indonesia here today that foreign missionaries of the Roman Catholic Church will be withdrawn from his country as quickly as possible, and appealed in effect for âpatienceâ during what he calls the âtransition period.â The Indonesian revolutionary nationalist leader assured the Pope in reply that he considered the 1,000,000 Indonesian Catholics, slightly more than 1 per cent of the population, a loyal part of his country and wished the Catholic religion well in his homeland.

+ …

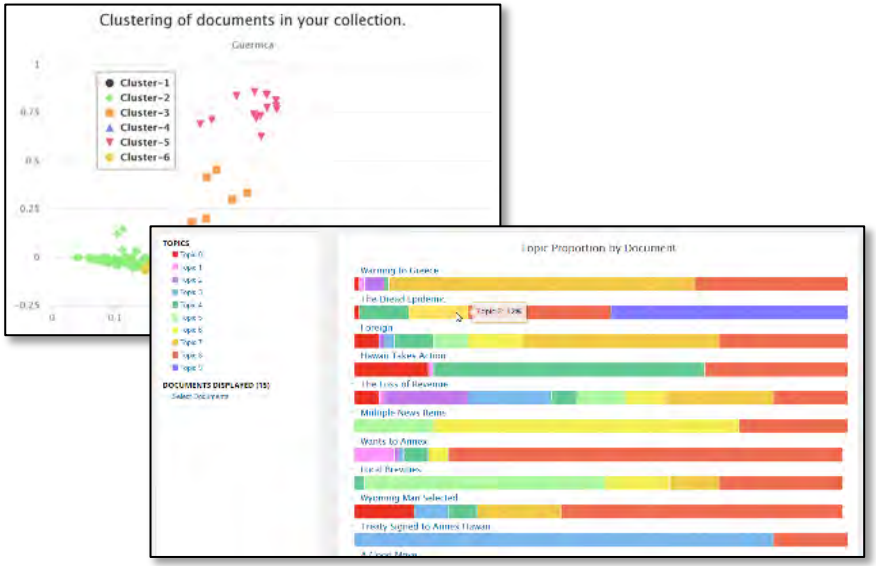
=

?

united states
united nations
new guinea
president sukarno
mr sukarno
foreign minister
indonesian government
communist party
security council
associated press
indonesian republic
west irian
west new
dutch new
air force
news agency
mr subandrio
minister subandrio
president sukarnos
continued page
prime minister
indonesian army
states indonesia
good offices
east indies
east java
herald tribune
united press
central java
dutch troops
defense minister
soviet union
state department
dutch government
republican army
netherlands government
republican government
associated pressbatavia
mr sukarnos
communist china
dutch forces
ceasefire order

Different Workflows

Search text and **Retrieve** images



Gather data and **Analyze** content sets

… Require Different Platforms

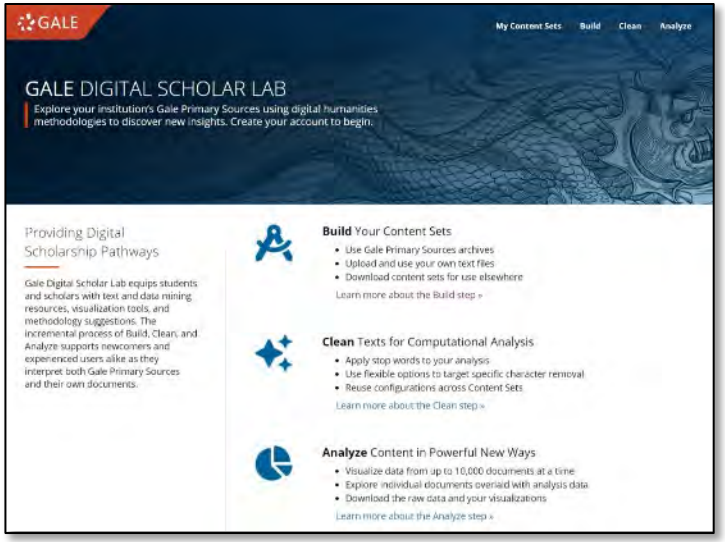
Gale Primary Sources

Gale Digital Scholar Lab

Gale Digital Scholar Lab

Gale's Cloud-based Text and Data Mining Research Environment

- **Access** OCR/HTR text from Gale Primary Sources
- **Build** custom content subsets
- **Clean** OCR text to reduce or control noise
- **Analyse** with simple but powerful tools
- **Export** statistical data, visualizations, OCR text in standard formats
- **Organize** and manage research



Gale Digital Scholar Lab
Workflow and Tools

GALE
A Cengage Company

EMPOWER™ RESEARCH

Gale Digital Scholar Lab: The Workflow

The Six Text Mining Tools Available in the Lab

Name of Tool	Common Uses	Library / Toolkit	Open Source?
Ngram	Extracts frequently occurring words and phrases	Lucene	No
Clustering	Groups documents in “clusters” according to their content	SciKit Learn	Yes
Topic Modelling	Extracts common topics found across many documents	Mallet	Yes
Sentiment Analysis	Classifies documents between positive to negative sentiment	OpenNLP	Yes
Named Entity Recognition	Extracts “Named Entities” (proper names, dates, prices, etc.)	spaCy	Yes
Parts-of-Speech Tagger	Compares the use of grammatical properties (nouns, verbs, etc.)	spaCy	Yes

1. Ngram (Frequent Terms)

The Ngram tool breaks down the text in your content set into sets of n consecutive words (Ngrams) and counts their occurrences. Ngrams are like *phrases* except that the resulting pieces are not always meaningful. They are helpful in finding **frequently occurring words and phrases** in your content set.

For example, if we split the text **“Nobody in Singapore drinks Singapore Slings.”**:

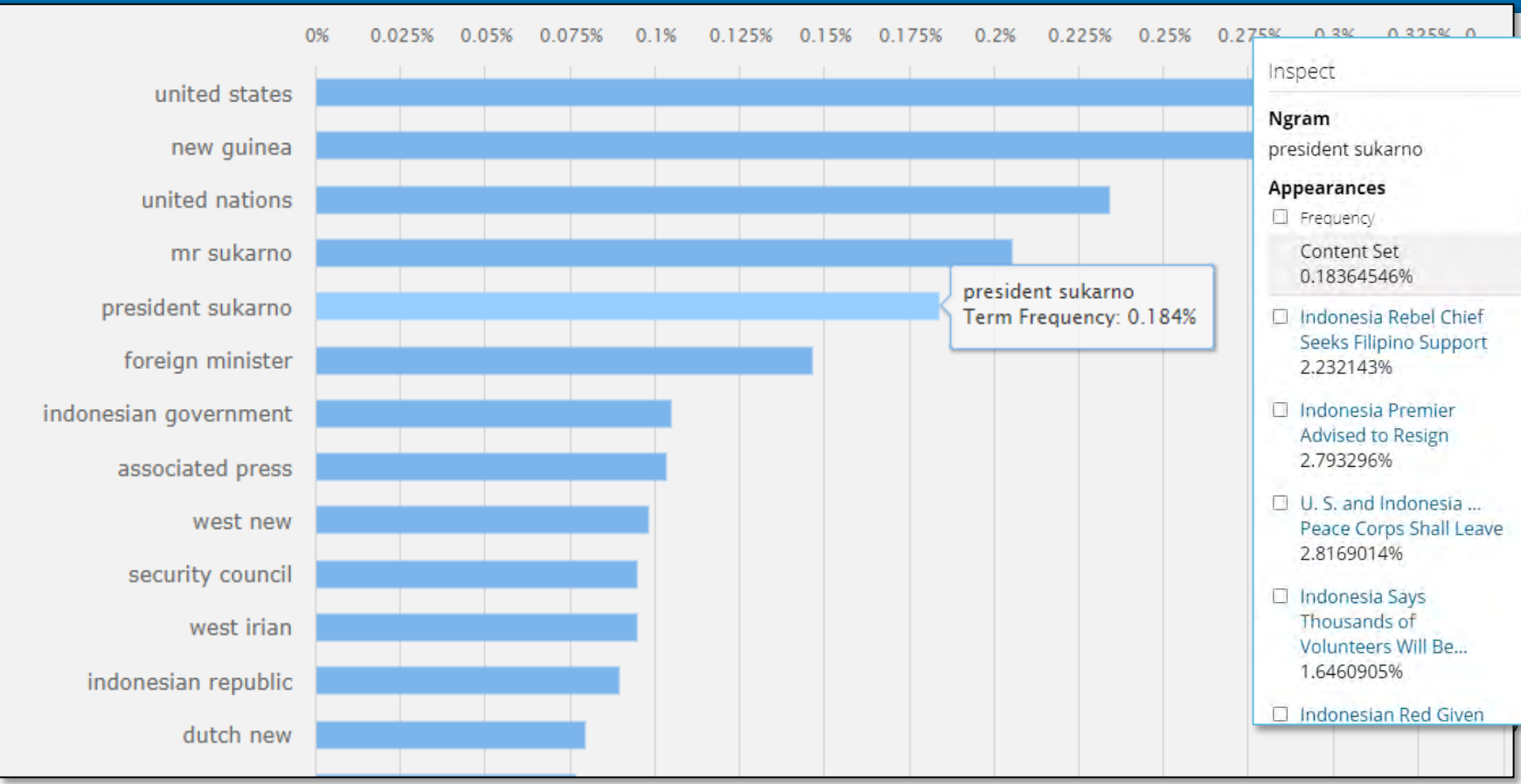
N=1: Unigram
“Nobody", “in", “Singapore", “drinks“, … etc.

N=2: Bigram
“Nobody in", “in Singapore", “Singapore drinks“, “drinks Singapore”, … etc.

N=3: Trigram
“Nobody in Singapore", “in Singapore drinks", “Singapore drinks Singapore“, “drinks Singapore Slings”, … etc.

Note: Ngrams that are composed entirely of stop words will not be considered. Ex. **“To Be or Not to Be”**

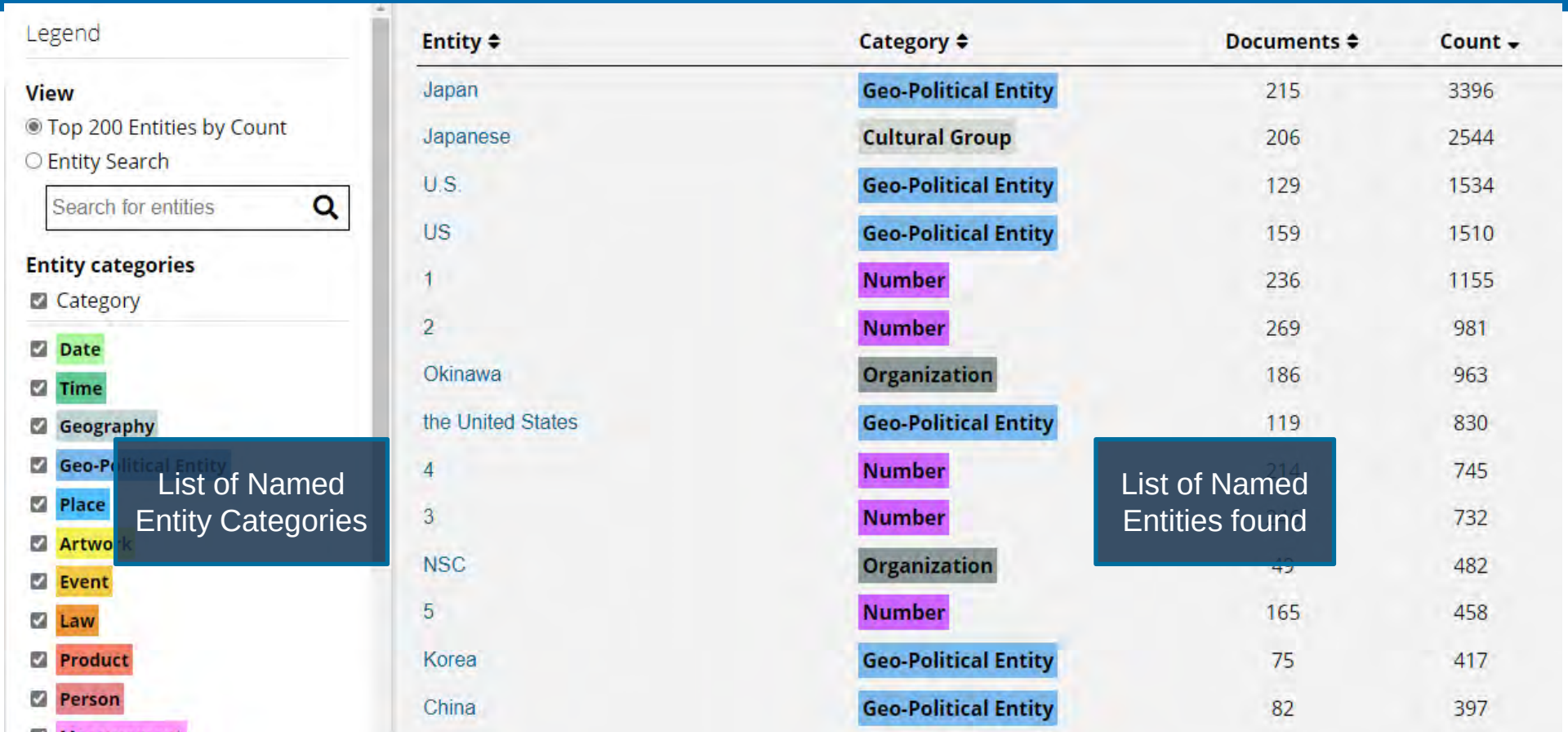
1. Ngram Lab Output (Bar Graph)



2. Named Entity Recognition (NER)

- NER recognises and extracts “Named Entities” (proper names, dates, prices, etc.) from documents within a content set, and output lists of such entities.
- Not designed to analyze textual features, but to extract them, and create new datasets which can then be used to contextualise content sets.
- Example “entity types”:
  - **people** (including fictional),
  - **groups** (nationalities, religious, or political),
  - **organizations** (companies, agencies, institutions),
  - **locations** (countries, states, cities),
  - **products** (objects, vehicles, foods, etc.), works of art (titles of books, songs, etc.),
  - **dates** (absolute or relative dates or periods)

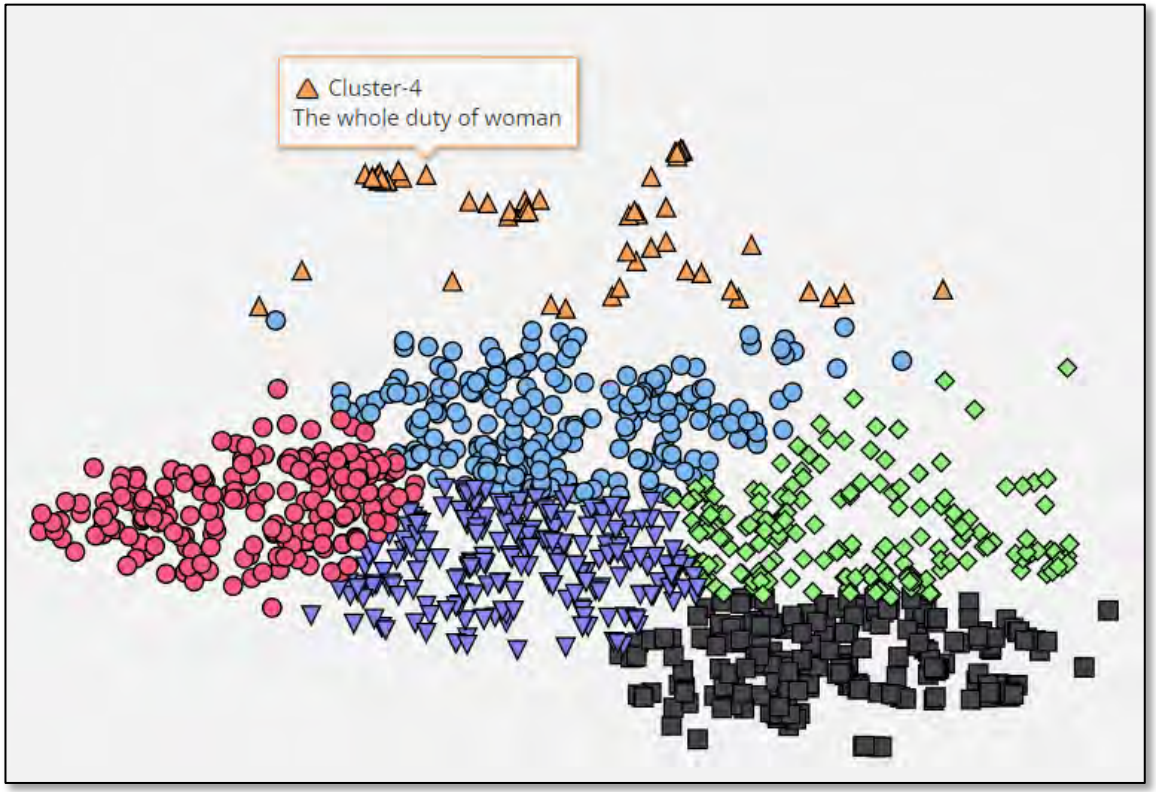
2. Named Entity Recognition Output



3. Clustering

- The Clustering tool looks at a collection of documents, and groups similar documents into clusters
- The user specifies the number of clusters to display
- The tool analyses the documents using statistical methods to group them around particular features

3. Clustering Output



Inspect

Document

The polite instructor; or, youth's museum. Consisting of moral essays, tales, fables, visions, and allegories. Selected from the...

Open Document

Top 20 Most Similar Documents

The smaller the distance, the greater the similarity to the selected document. Download Full List

Distance	Title
0.0001	Instructions for a young lady, in every sphere and period of life. Containing, I. A...
0.0001	The lady's miscellany; or Pleasing essays, poems, stories, and examples, for...
0.0005	The ladies complete letter-writer; teaching the art of inditing letters On every...
0.0013	Instructions for a young lady,

4. Sentiment Analysis

This tool calculates the “sentiment score” of each document, which indicates the use of positive or negative words within them

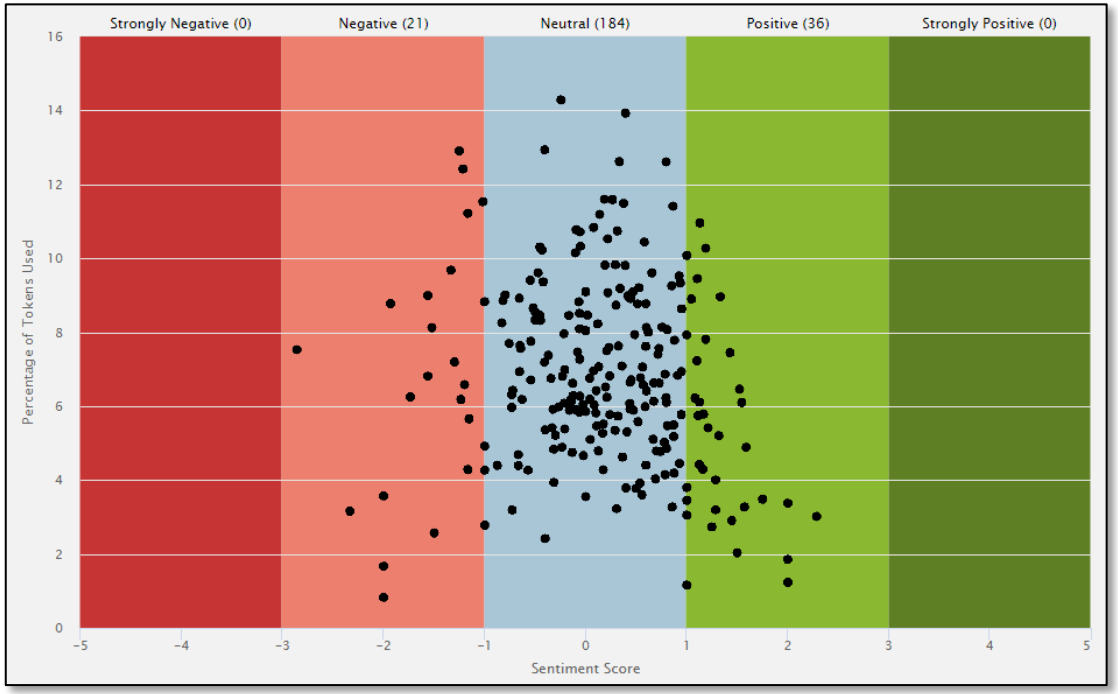
It works by applying a lexicon of terms and their associated sentiment scores to a piece of text

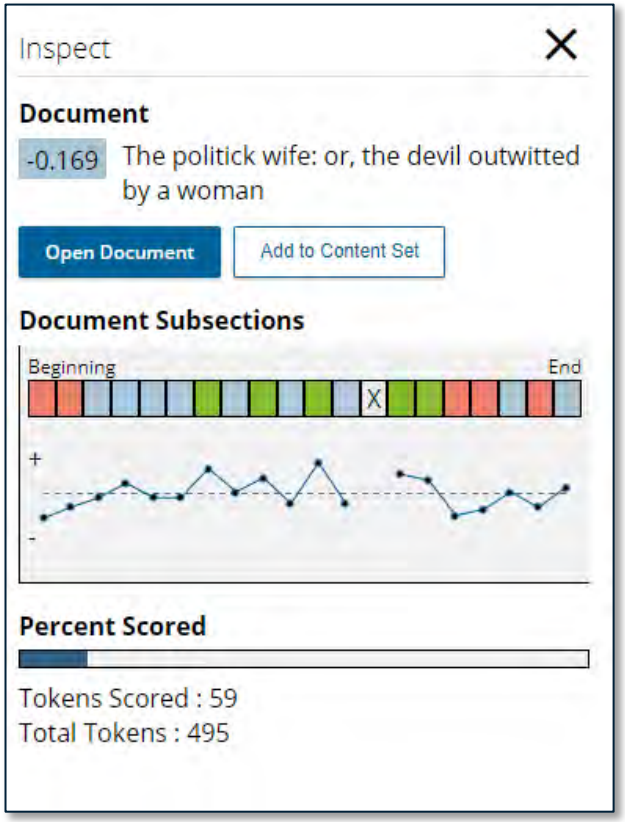
It also divides each document into 20 sections and calculates the sentiment score for each section

4. Sentiment Analysis Output

Negative

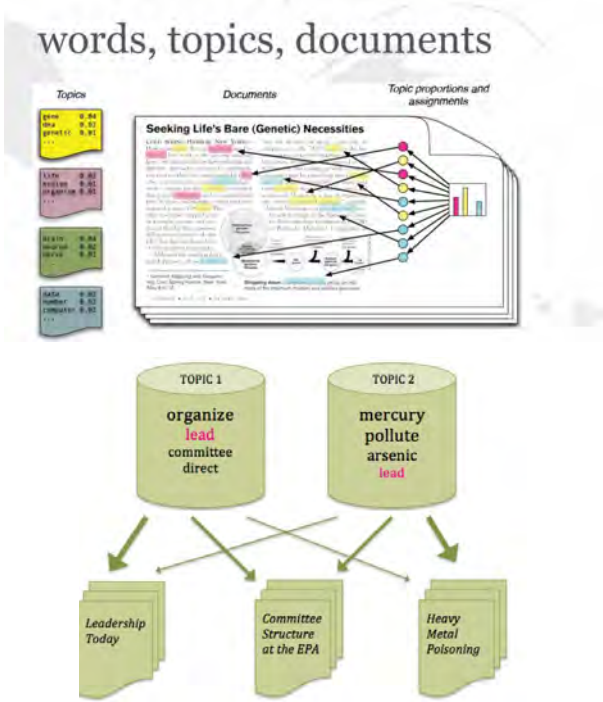
Positive



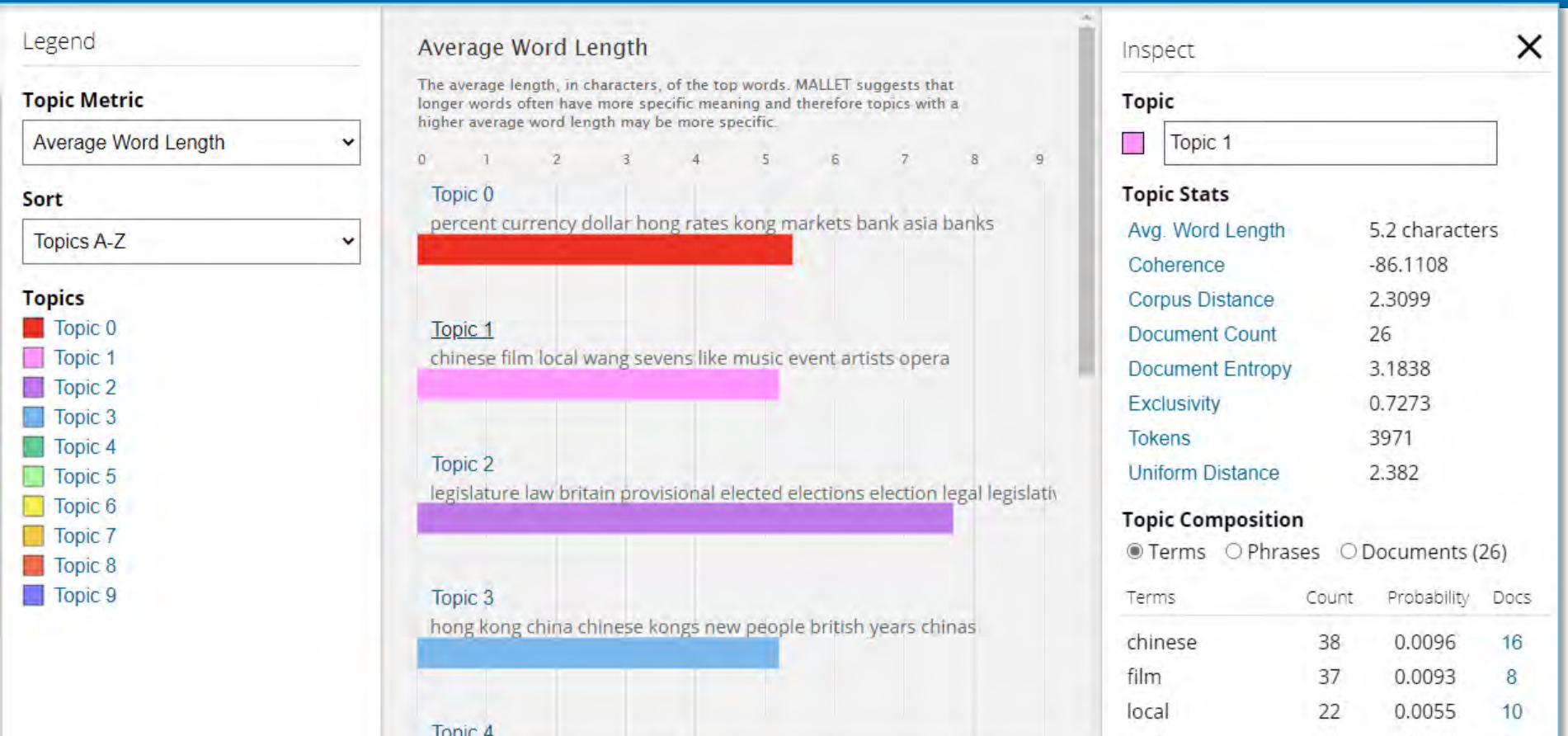


5. Topic Modelling

- Typically used to analyze a large number of documents to find out what "**topics**" occur in common
- "Topic" = a list of keywords that occur in statistically meaningful ways. Topics are not named
- The relationship between documents and topics is a many-to-many relationship
  - Most documents contain multiple topics, and most topics appear in multiple documents, but with different weights assigned

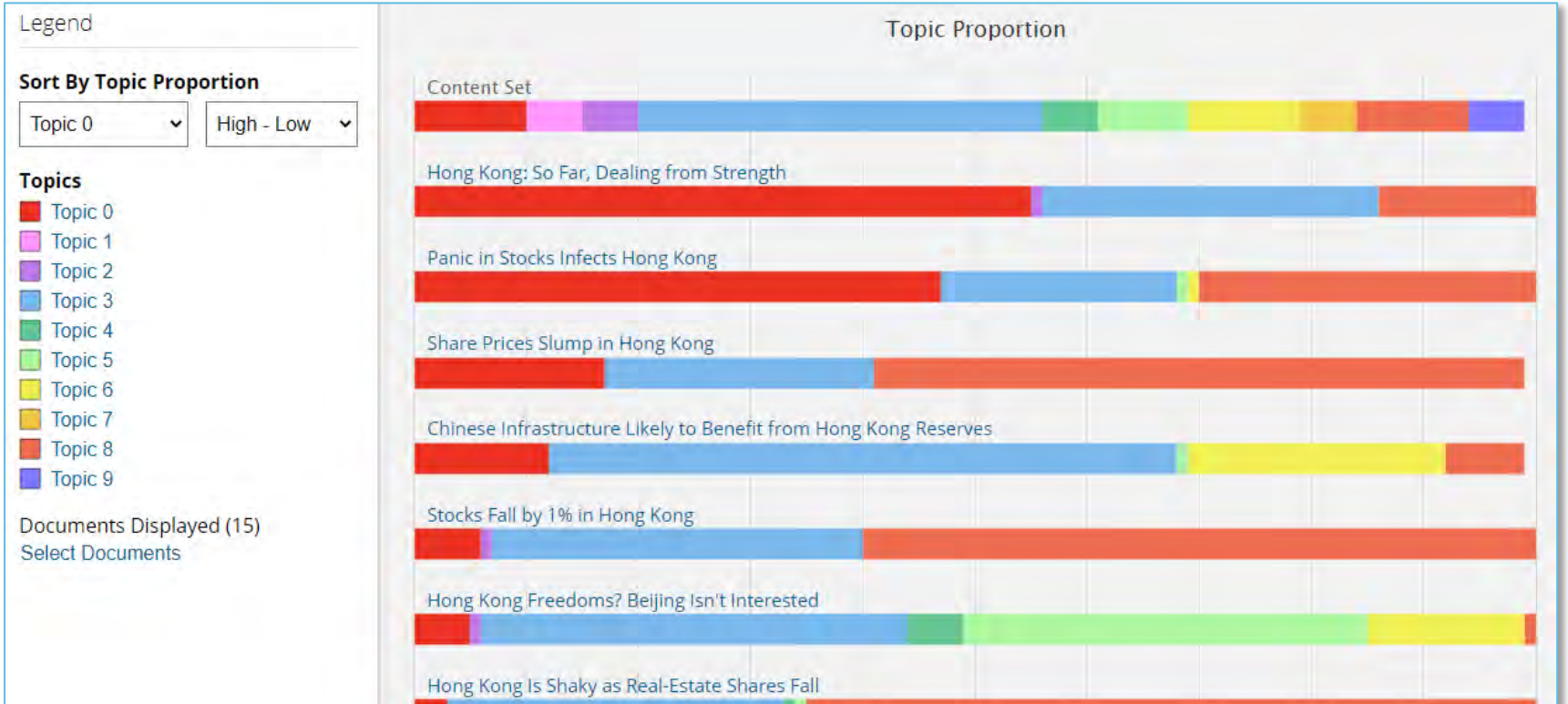


5. Topic Modeling Output (Topic Proportion by Document)



5. Topic Modeling Output (Topic Proportion by Document)

- Show the topic proportion within each of the documents



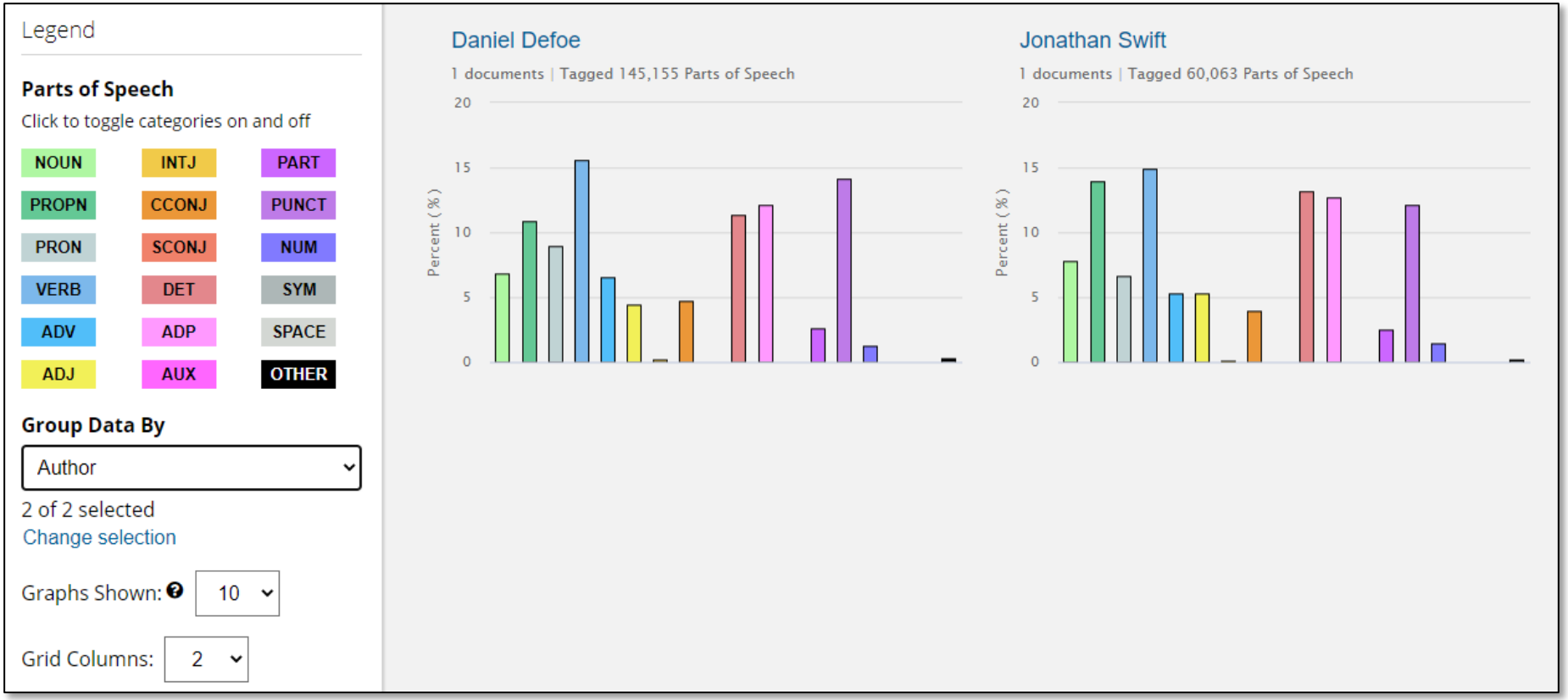
6. Parts-of-Speech Tagger

The Parts-of-Speech tags **grammatical properties** (ex. noun, verb, adjective, adverb, etc.) to the words in documents.

It provides users with the building blocks for looking at how phrases are constructed within each document in a content set.

This tool effectively creates a lexicographical index or dictionary of a content set. In this implementation of Parts of Speech Tagger you may review how authors use of speech varies over time.

6. Parts-of-Speech Tagger Output

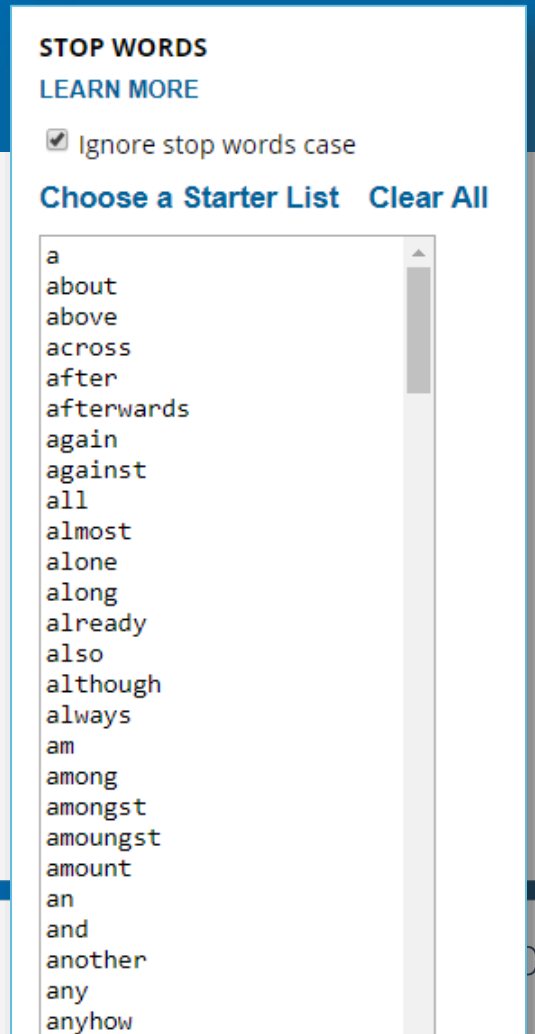


OCR Cleaning Configurations (Optional)

- The Lab has a built-in **OCR cleaning tool**
- Users may **change the default options** by editing stop words, removing special characters, replacing specific words, etc.
- Users can **test-run** their cleaning configurations
- The configurations can be **saved** in the user's account

Stop Words

- Stop words are words that are filtered out because they are too common and do not add much information
- Examples in English are “the” “a” “an” “at” “in” etc., etc.
- **Stop words can be edited by the user** to remove/include more words to meet their purposes
- We currently provide standard Stop Words in 26 different languages including English



GALE
A Cengage Company

Why Clean?

- OCR text can contain considerable “noise”

Document Text

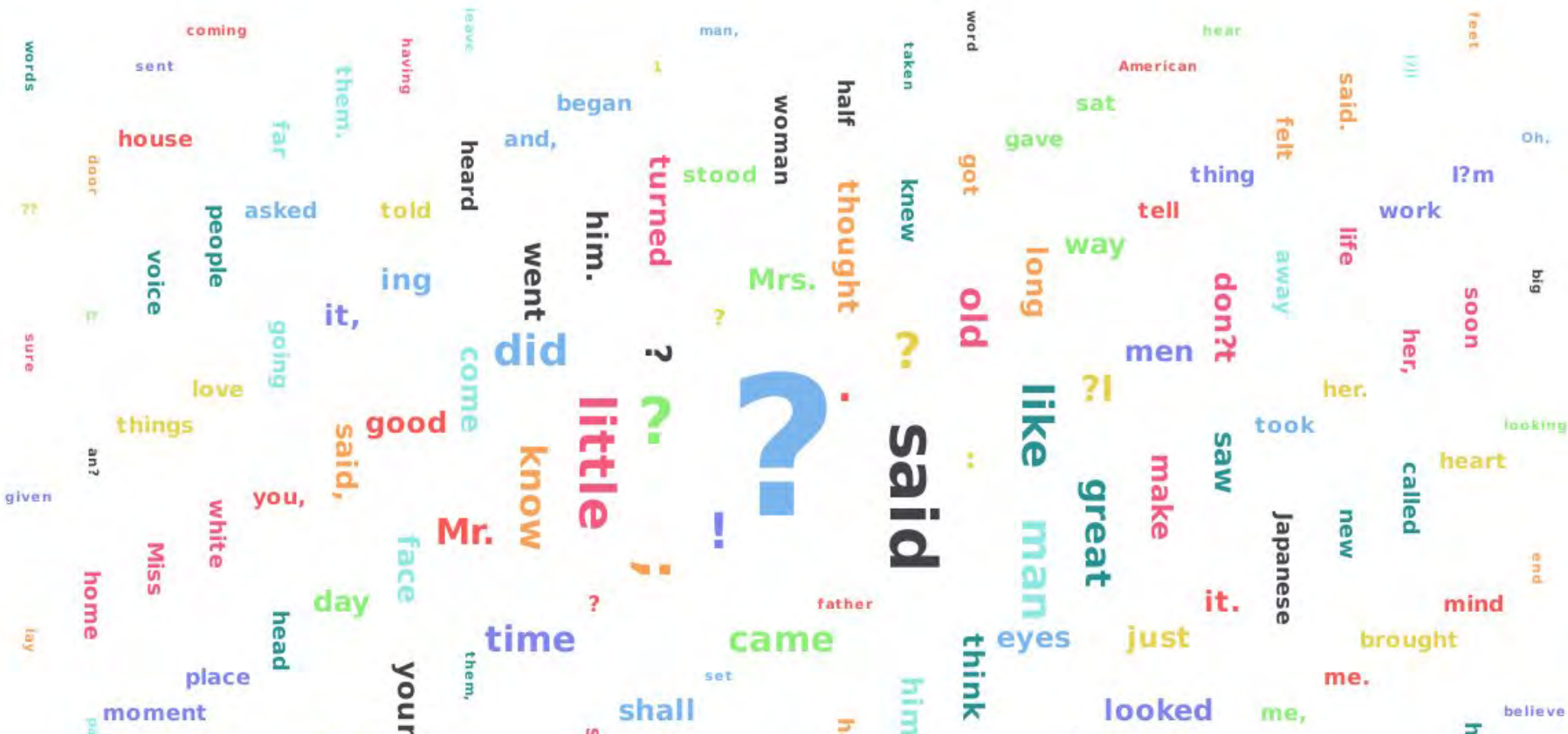
OCR Confidence: **73%** Learn how this text was created

The Language War in Hong Kong Cantonese, Mandarin and English Compete for a Place at School

it m Hi ¥ I m .-A m A g ; * m ¥ I :li# ■.-Ni f itf % ^vf Wv yvX^AxxA ;"¥ f f: x A I I m I si I* ill 3* 7 si WXvSos SSSKS.' ^ * -a? IN i M M i \ ' 1 ■N m M: •41' •Jg m J m >. M |§f§§|§ 'XV lli¥. k'&'S&Xs* sSS\\\|>x: \si X ill ' xxx ' ifisk ¥ ' *1 X ;I KM

A student calling out his answer during a class at the Shatin Tsung Tsin Secondary School in Hong Kong. ' A , ' I Jenni Meili Lau/The Washington Post By Steven Knipp ONG KONG — For years this form¬ er British colony has had the dubious distinction of being a place where it was a serious social stigma not to be able to speak a foreign language fluently. For the city's 6 million Cantonese-speaking majority, that language was English. British rule ended in July, but the people of Hong Kong face a new quandary. Which language should their chil¬ dren learn? Cantonese — the local dialect spoken at home and in the street? Mandarin — H China's official language? Or English — the language of international business? Whatever path the govern¬ ment takes, it will have far- reaching effects on a gener¬ ation of young people. The language issue here is wrapped in conflicting influ¬ ences and emotions, not the least of which is the local Cantonese pride that is res¬ istant to Beijing's hopes of integrating Hong Kong into the “motherland” as quickly as possible, residents, some 60 million people in Guangdong Prov- ince nearby speak Cantonese, But Beijing considers the southern tongue just

Before Cleaning



After Cleaning

Demonstration

Gale Digital Scholar Lab: The Workflow

Build

- Search
- Select and Curate
- Create Content Set (Max 10,000 items per content set)
- View Complete Search History
- Duplicate/Merge Content Sets
- Alternatively, upload your own text

Clean

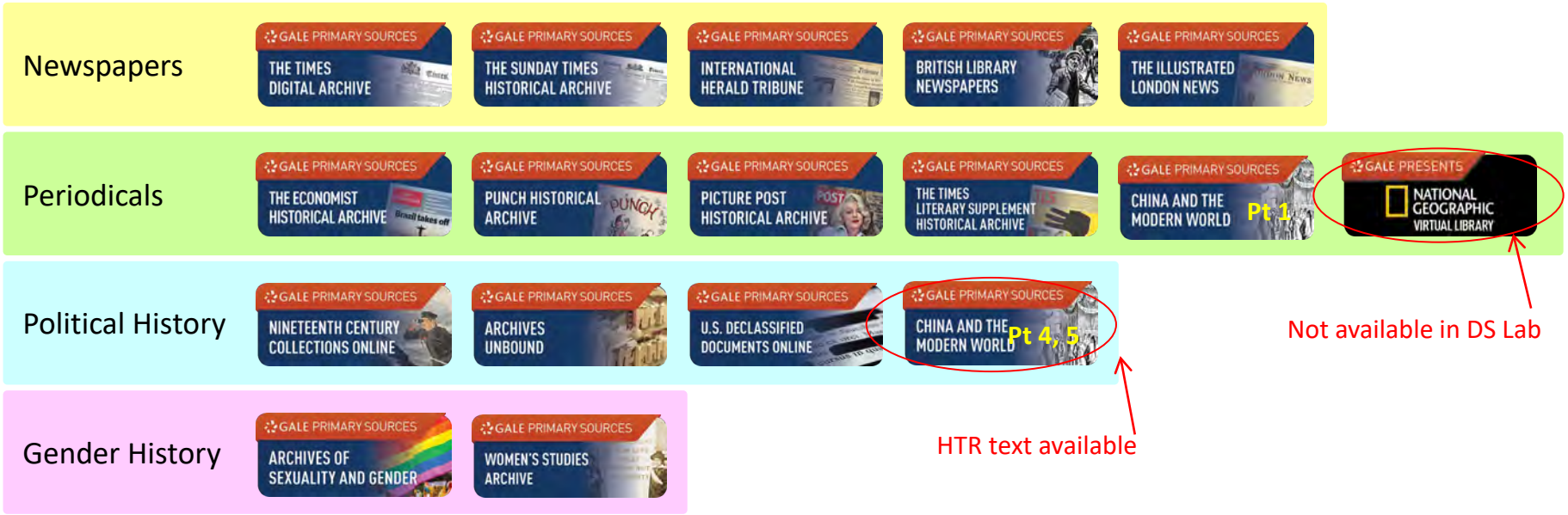
- Apply default cleaning options
- Or, set your own cleaning configuration
- Edit stop words, remove special characters, etc.
- Save your cleaning configurations for later use

Analyze

- Run analyses on the fly
- Six Tools for different purposes
- Download visual and tabular output
- Download OCR for max 5,000 items per content set
- Run history

Which Gale Primary Sources are available to HKBU?

Gale Primary Sources Available to HKBU



Newspapers

Top UK quality paper with longest history and enormous influence, Years 1785-2014

Originally unrelated to *The Times*; analysis, investigative reporting, and culture, 1822-2006

US-owned global paper based in Paris with international readership, 1887-2013

The greatest illustrated weekly newspaper in the UK, 1842-2003

Collections of 70+ English newspapers from the 19th century, 1801-1900

Periodicals

Top UK business weekly with global readership among business elites, Years 1843-2015

The greatest satirical weekly magazine in the UK, 1841-1992

Short-lived but important UK weekly photojournalistic magazine, 1938-1957

Aka “TLS” - famous and influential weekly literary review, 1902-2013

Part 1: 17 English-language missionary, sinology, literary periodicals from China, 1817-1949

Visual US monthly magazine on nature, peoples and science, 1888-1994 *

* National Geographic not available in Digital Scholar Lab

Political & Gender History

Diplomatic reports from Asia to the US/UK, missionary reports, periodicals, and others

Part 4: British Colonial Office papers CO 129 on Hong Kong, 1841-1951 *
Part 5: British Foreign Office papers FO 17 on China, 1815–1881 *

* China and the Modern World, Parts 4-5 contain HTR text

Declassified US government documents, mainly from post-WWII 20th century

14 small sized collections related to China, Japan, Southeast Asia, and missionary activities

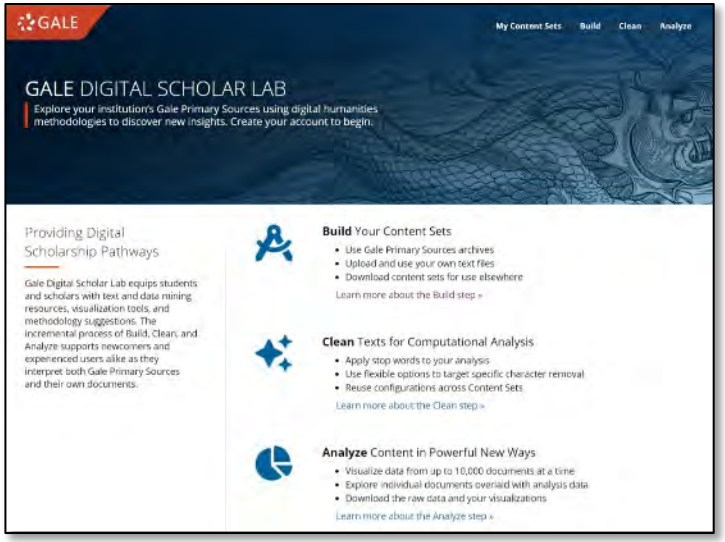
Publications and documents related to LGBTQ movements and their oppression

Publications and documents related to women’s movements, feminism, and other topics

Gale Digital Scholar Lab

Gale's Cloud-based Text and Data Mining Research Environment

- **Access** OCR/HTR text from Gale Primary Sources
- **Build** custom content subsets
- **Clean** OCR text to reduce or control noise
- **Analyse** with simple but powerful tools
- **Export** statistical data, visualizations, OCR text in standard formats
- **Organize** and manage research



Learning Center

LEARN ABOUT THE LAB

What Texts are Available?

See the full list Gale Primary Sources made available to you through your institution and learn more about each.

View the Archives

What Analyses Can I Run?

Get an overview of analysis methods, available tools, text and data mining concepts, and more.

Explore the Analysis Tools

Try Out Sample Projects

Sample Projects are a great way to acquaint yourself with the tasks you can undertake and results you can get with the Digital Scholar Lab.

View Sample Projects

What is the Gale Digital Scholar Lab?
01:16

Transforming Scholarly Research
01:10

Supporting Faculty Research
02:43

A Few Key Points that Might Help

- **Search/Retrieve before Gather/Analyze**

- **Using the Lab is a Cyclical Trial-and-Error Process**

- **Noise in OCR is to be expected**

- **The more contextual knowledge you have the better**

Thank you!