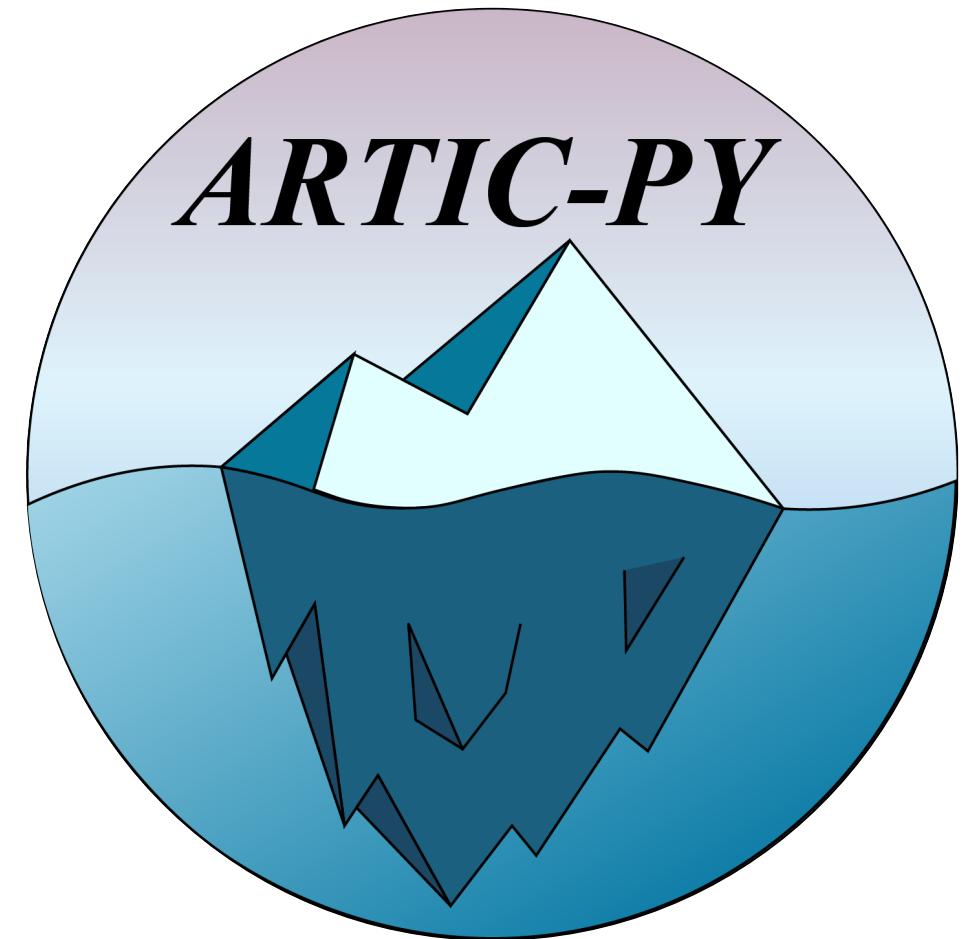


Coding Challenge 2021

Topic 2: News Analysis

Police National Service Department

- Ong Jung Yi
- George Tay



Project Objective

• Problem Statement & Context

Problem Statement 1

To analyse public sentiment from news articles to investigate how sentiment correlates to the crime index

Reference Literature

Studies show that negatively polar sentiments can lead to a higher risk of social unrest

Based on: A Tool to Predict the Possibility of Social Unrest Using Sentiments Analysis of Zimbabwe Politics 2017-2018 (Elliot Mbunge - Chinhoyi University of Technology)

Reasons for Low Sentiments

- Poor economy affected the local's standard of living
- Increase in crime – E.g. corruption, misuse of funds
- Lower sentiments in the community due to high inflation and unemployment rates

Problem Statement 2

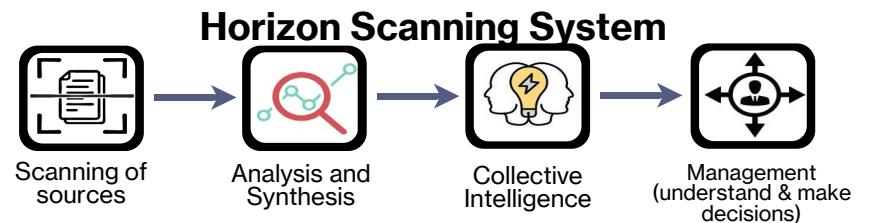
To analyse public sentiments from news articles to identify topics or issues within society.

Importance of understanding Public Sentiment

- Ground Sensing and surveys are used by officers most to understand sentiment
 - Communities that distrust police/ government often will not participate in such interactions. Concerns harder to detect

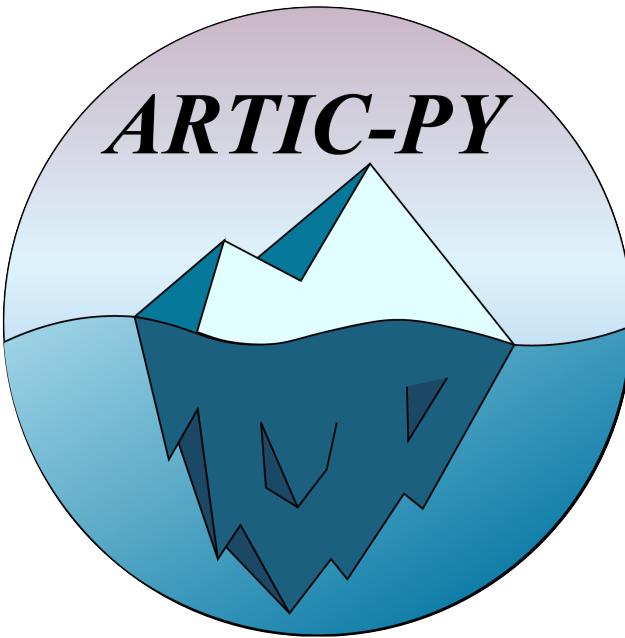
Based on: Sentiment Analysis: The Missing Link in Police Performance Management Systems - Police Chief Magazine

- Sentiment Analysis for horizon scanning can be used to complement ground sensing
- News analysis can identify issues that are harder to detect by ground sensing and understand emerging issues of concern



ARTIC-PY targets **Analysis and Synthesis**, and facilitates **Collective Intelligence**





Application Overview



App Overview

- Initiating ARTIC-PY – 3 Methods available

Application Source Code

(Recommended)

Set-up Difficulty:	Medium
Set-up Control:	High
Method Security:	Medium

Benefits

1. **Most Balanced**– Method gives users the most control over the initiation of the app.
2. Access to our source code:
<https://github.com/asdfghjkxd/ArticPy>

Docker platform

Set-up Difficulty:	High
Set-up Control:	Medium
Method Security:	High

Benefits

1. **Secure** as it is containerized
2. Easy to orchestrate the running of apps on multiple devices (*Docker Compose, Swarm, Kubernetes*)

Streamlit Web Application



Set-up Difficulty:	Low
Set-up Control:	Low
Method Security:	Low

<https://share.streamlit.io/asdfghjkxd/articpy/main/app.py>.

Benefits

1. Very convenient to initiate the app

Limitations

1. Limited volume of data due to hardware limitation of running the app online
2. Run through Source code and Docker to make use of system CPU/ GPU/ RAM



App Overview

- Load, Clean and Visualise Module

ArticPy *

An app built to simplify and condense NLP tasks into one simple yet powerful Interface.

App Modules

Select the following available modules:

NLP Functions

Load, Clean and Visualise

Load, Clean and Visualise

NLP Toolkit

Load, Clean and Visualise Data

Module Description

Origin of Data File

Local

Define the Data Input Format

CSV

Load CSV File

Drag and drop file here
Limit 1000GB per file • CSV

Browse files

Warning: Your Dataset file is not loaded.

Processing Mode

Choose the type of processing you want to apply to your dataset. You may choose between the three processes: Cleaning, Modification (Country Extraction) and Query.

Choose Data Processing Mode

Data Cleaning

Data Cleaning Mode Selected

Data Cleaning

Data Modification

Data Query

ARTIC-PY

Program Language: Python3

Software Platform: Docker

App UI: Streamlit

Module Functions

- Cleaning and visualising raw data from the news articles
- Data querying and Data tagging functions





App Overview

- Natural Language Processing Toolkit Module

ArticPy *

An app built to simplify and condense NLP tasks into one simple yet powerful Interface.

App Modules

Select the following available modules:

NLP Functions

NLP Toolkit

NLP Toolkit

Module Description

+

Upload Data

Origin of Data File

Define the Data Input Format

Local

CSV

Load CSV File

Drag and drop file here
Limit 1000GB per file • CSV

Browse files

Warning: Your Dataset file is not loaded.

NLP Operations

Select the NLP functions which you wish to execute.

Select the NLP Operation to execute

Topic Modelling

Topic Modelling Selected

Topic Modelling

Ensure that your data is lemmatized and properly cleaned; data should not be tokenized for this step. Use the Load, Clean and Visualise module to clean and lemmatize your data if you have not done so already.

Module Functions

- Tools to conduct In-depth analysis and visualisation of the data

Topic Modelling

Topic Classification

Analyse Sentiment

Word Cloud

Named Entity Recognition

POS Tagging

Summarise



Project Objective



Application Overview



Analysis of PS1



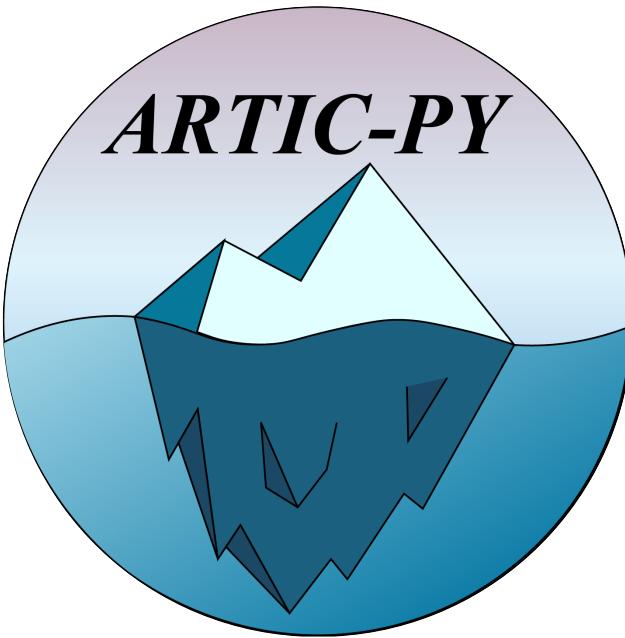
Analysis of PS2



Other Features



Moving Forward



Analysis of Problem Statement 1

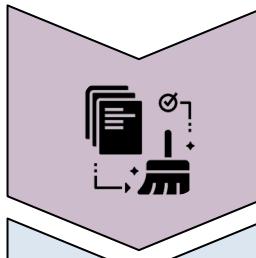
Analysis – Problem Statement 1

- Overview Procedure

Problem Statement 1

To analyze public sentiment from news articles to investigate how sentiment correlates to the crime index

01 Data Cleaning



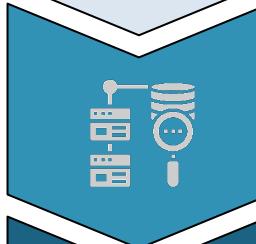
Raw data is pre-processed using Natural Language Processing Techniques to eliminate irrelevant details

02 Data Modification



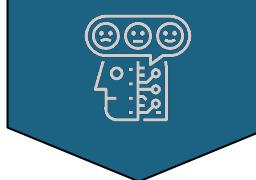
Tagging of the data according to its relevant countries

03 Data Querying



Querying the tagged data to extract a more focused dataset based on certain keyword(s)
– E.g. by Country

04 Sentiment Analysis



Analysis to determine the average polarity of a pre-processed data set



Analysis - Problem Statement 1

• 01 – Load, Clean & Visualise Module (Data Cleaning)

Step 1: Choose the Load, Clean & Visualise module

Step 2: Choose the Data Cleaning function

Expand and Collapse Module Description

Project Objective **Application Overview** **Analysis of PS1** **Analysis of PS2** **Other Features** **Moving Forward**



Analysis - Problem Statement 1

• 01 – Load, Clean & Visualise Module (**Data Cleaning**)

Load, Clean and Visualise Data

Module Description

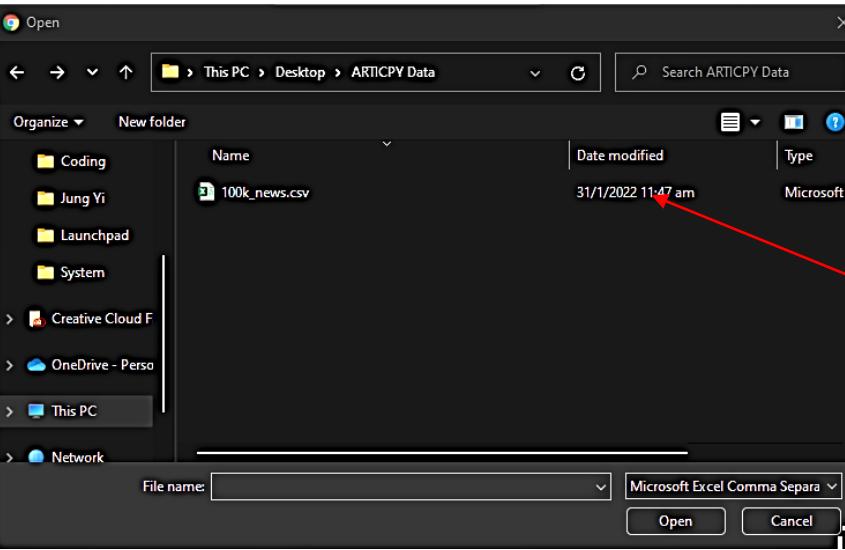
This module is used for the visualisation and cleaning of data used for NLP Analysis. Please ensure that sufficient storage (~1 GB) is available on your device when you run the app. Please ensure that sufficient storage (~1 GB) is available on your device when you run the app. Please ensure that sufficient storage (~1 GB) is available on your device when you run the app.

For the cleaning process, all non-ASCII characters will be removed, and all non-English characters will be converted to English.

Before proceeding, you will need to download the corpus needed to process your dataset. This will take up some space available so that you are able to download the corpus onto your system and run the analysis.

NTLK Data Detected

Begin Download



Step 3: Hit the button **browse file**

Step 4: Upload the file containing raw data – **100k_news.csv**

Upload Data

Origin of Data File

Local

Define the Data Input Format

CSV

Load CSV File

Drag and drop file here
Limit 1000GB per file • CSV

Browse files

Warning: Your Dataset file is not loaded.

File Uploading

- Option 1:** Upload from local machine
- Option 2:** Pull data from supported Cloud Service Providers (CSP)

Azure, Amazon, Google

Processing Mode

Choose the type of processing you want to apply to your dataset. You may choose between the three processes: Cleaning, Modification (Country Extraction) and Query.

Choose Data Processing Mode

Analysis - Problem Statement 1

• 01 – Load, Clean & Visualise Module (**Data Cleaning**)

Processing Mode

Choose the type of processing you want to apply to your dataset. You may choose between the three processes: Cleaning, Modification (Country Extraction) and Query.

Choose Data Processing Mode

Data Cleaning

Data Cleaning Mode Selected

Options

Save Outputs? ⓘ

Override Output Format?

Display Outputs? ⓘ

Data Points To Print

0 → 20

Display Advanced DataFrame Statistics? ⓘ

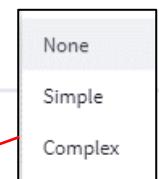
Select Preprocessing Pipelines

Complex

Tokenize Data? ⓘ

Extend List of Stopwords? ⓘ

Step 5: Check Save Outputs & Display Output



Step 6: Select Complex Pipeline

Data Scanning and Cleaning

Simple Pipeline

- Remove HTML tags
- Remove Diacritics
- Remove White spaces
- Remove cells with no content

Complex Pipeline

- Same as simple-
- Remove cells with NA value
 - Remove punctuation
 - Remove URLs
 - Lowercase all text
 - Remove stop words

Complex: Module to extend stop word list available

Data Cleaning and Visualisation

Ensure that you have successfully uploaded the required data and selected the correct column containing your data before clicking on the "Begin Analysis" button. The status of your file upload is displayed below for your reference.

File loaded.

Begin Analysis

Step 7: Hit begin analysis



Project Objective



Application Overview



Analysis of PS1



Analysis of PS2



Other Features



Moving Forward

Analysis - Problem Statement 1

• 01 – Load, Clean & Visualise Module (**Data Cleaning**)

ArticPy *

An app built to simplify and condense NLP tasks into one simple yet powerful Interface.

App Modules

Select the following available modules:

NLP Functions

Load, Clean and Visualise

Cleaned DataFrame

	CONTENT	CLEANED CONTENT
0	TOKYO -Back in 2004, Japan c	tokyo back japan came r envi
1	COLOMBO (AP) - Sri Lankan P	colombo ap sri lankan presid
2	MOSCOW (AP) An unidentified	moscow ap unidentified gunr
3	Norway's right-wing Progress	norway right wing progress p
4	ad with plans for a new tax or	ad plans new tax american te
5	Photo Credit: Yossi Zamir/Fla	photo credit yossi zamir flash
6	Everybody likes a little treat r	everybody likes little treat ive
7	A SmartyPants Vitamin Gold I	smartypants vitamin gold bo
8	Jan. 21 (UPI) -- Former footba	jan upi former football star cu
9	Disney calls it the future of th	disney calls future company c
10	Google Stadia, which launces	google stadia launched never
11	Scientists have identified the	scientists identified oldest kn
12	A record-setting spacewalker,	record setting spacewalker or
13	If you want to keep your licen	want keep licensing simple o
14	Barts Health NHS Trust has tu	barts health nhs trust turned
15	Fun fact: That's also how anir	fun fact also animals learn pu
16	INILAHCOM, Seoul - Perusahaan	inilahcom seoul perusahaan t
17	Published: 17:17 BST, 8 August	published bst august updated
18	Isobel Asher Hamilton, Busin	isobel asher hamilton business
19	Published: 16:43 BST, 8 August	published bst august updated

Quick View and Output

- Returns the top 20 rows of the cleaned data frame
- Download links for processed data frame in the form of CSV available for output

Download Data

[Download Raw Data](#)

[Download Cleaned Data](#)

Step 8: Hit “Download Cleaned data”

Analysis - Problem Statement 1

• 02 – Load, Clean & Visualise Module (**Data Modification**)

ArticPy *

An app built to simplify and condense NLP tasks into one simple yet powerful interface.

App Modules

Select the following available modules:

- NLP Functions
- Load, Clean and Visualise

Load, Clean and Visualise Data

Module Description +

Upload Data

Origin of Data File Define the Data Input Format

Local CSV

Load CSV File

Drag and drop file here Browse files
Limit 1000GB per file • CSV

100k_news.csv 433.3MB X

Choose Column where Data is Stored

CONTENT

Data Loaded from CONTENT!

Processing Mode

Choose the type of processing you want to apply to your dataset. You may choose between the three processes: Cleaning, Modification (Country Extraction) and Query.

Choose Data Processing Mode

Data Modification Data Modification Mode Selected

Step 9: Upload the Cleaned File

Step 10: Choose Data Modification

Analysis - Problem Statement 1

• 02 – Load, Clean & Visualise Module (**Data Modification**)

Data Modification Mode

Choose Mode
Country Extraction

Options

Save Outputs? ⓘ
 Override Output Format?
 Display Outputs? ⓘ
Data Points To Print
20
0
 Display Advanced DataFrame Statistics? ⓘ
 Generate a World Map Representation of the Countries Mentioned?

Step 11: Select the checkboxes to save and display the outputs

Step 12 Select the checkbox to visualize the data on a world map

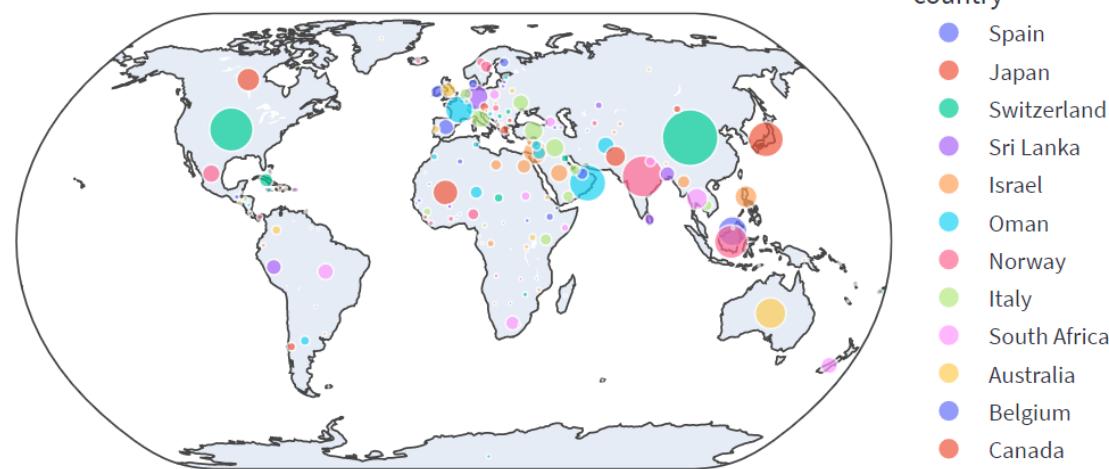
Country Name Mention Frequency

	country	count
0	Spain	1940
1	Japan	9148
2	Switzerland	1053
3	Sri Lanka	846
4	Israel	4258
5	Oman	9892
6	Norway	695
7	Italy	2591
8	South Africa	1500
9	Australia	7041
10	Belgium	916
11	Canada	4042
12	Cuba	1437
13	Germany	4794
14	Egypt	1566
15	France	5503
16	India	12590
17	Netherlands	1086
18	New Zealand	2098
19	Portugal	449

Analysis - Problem Statement 1

• 02 – Load, Clean & Visualise Module (**Data Modification**)

Country Name Mention Frequency



Step 1: Scanning & Extraction

- Scans each text in dataset
- Extracts country names mentioned within the text as lists

Step 2: Country Tagging

- Appends list of countries to the next column of data frame corresponding to the text

Purpose:

- Part of the data preparation process
- Allows for the tagging of the data according to country of relevance
- Important for next step

Libraries

pycountry
Geograpy3



Analysis - Problem Statement 1

• 03 – Load, Clean & Visualise Module (**Data Query**)

Load, Clean and Visualise Data

Module Description

Upload Data

Origin of Data File

Local

Define the Data Input Format

CSV

Load CSV File

 Drag and drop file here
Limit 1000GB per file • CSV

Browse files

 globe_data.csv 0.6GB

X

Choose Column where Data is Stored

COUNTRIES

Data Loaded from COUNTRIES!

Processing Mode

Choose the type of processing you want to apply to your dataset. You may choose between the three processes: Cleaning, Modification (Country Extraction) and Query.

Choose Data Processing Mode

Data Query

Data Query Mode Selected

Step 13: Upload the tagged data set

Step 14: Ensure that the tagged data set is chosen for querying – “Countries”

Options

Save Outputs? ⓘ

Override Output Format?

Display Outputs? ⓘ

Data Points To Print

20

0

?

1000

Display Advanced DataFrame Statistics? ⓘ

Query Must Match Exactly? ⓘ

Query

Singapore X Press Enter to extend list...

Alert: Words detected.

DataFrame Query

This module will allow you to query certain parts of your DataFrame to get documents which fulfills your criteria. You may choose between querying just one column at a time (one condition per query) or multiple columns at a time (multiple conditions per query).

File loaded.

Query Data

Step 16: Enter keyword for data query – E.g. Singapore

Step 17: Hit Query Data to initiate the query function





Analysis - Problem Statement 1

• 03 – Load, Clean & Visualise Module (Data Query)

Query Data

Query Successful!

Query Results

	CONTENT	CLEANED CONTENT	COUNTRIES
30	As a small city-state at the foot of p small city state foot peninsular tra	['India', 'Mali', 'Malaysia', 'Oman', 'S']	
35	Please refer to the attachment. ---- please refer attachment announce	['Singapore']	
36	March 2017 - Present Deputy Chief march present deputy chief execut	['India', 'Singapore']	
37	KUALA LUMPUR To spare the counl spare country expense worth p mil	['Malaysia', 'Philippines', 'Singapore']	
38	Damage to the portside is visible a: damage portside visible missile de	['Afghanistan', 'Mexico', 'Singapore']	
39	On Monday morning, the world ha monday morning world one geopc	['Brazil', 'Bhutan', 'China', 'India', 'O']	
40	A five-year-old Spanish boy died in five year old spanish boy hospital l	['Singapore']	
41	Published August 28, 2017, 6:15 PM august ben one championship phc	['United Arab Emirates', 'China', 'Jo']	
42	(Aug 28): Bonds of a Dalian Wanda bond group unit fell chinese medi	['China', 'Hong Kong', 'Singapore']	
43	Inter Milan and Juventus both cam inter milan came behind maintain	['Singapore']	
44	PETALING JAYA (Aug 28): Eco World world development group high str	['Austria', 'India', 'Malaysia', 'Singap']	
56	KUALA LUMPUR (Jan 2): The FBM K jan fell midday break first trading c	['China', 'Malaysia', 'Singapore']	
72	A Malaysia Ringgit note is seen in t note seen illustration photo june p	['China', 'Malaysia', 'Singapore']	
85	Nearly half of 100 countries evalua nearly half country may good fit be	['Australia', 'Bhutan', 'China', 'Cook']	
87	When liberals talk about their heal liberal talk health care utopia score	['Canada', 'Germany', 'France', 'Sing']	
99	MANILA, July 31 (Xinhua) -- China a manila july china association south	['China', 'Indonesia', 'Cambodia', 'N']	
105	- Advertisement - An article on The advertisement article economist fa	['United Kingdom', 'Singapore']	
118	If you could trace the exact momer could trace exact moment australi	['Australia', 'China', 'Singapore']	
121	HRjobs: Asia's only regional recruit regional recruitment job post vaca	['Singapore']	
136	Amid desperate attempts by the ne amid desperate attempt medium c	['Singapore', 'United States']	

Libraries



Importance:

- Data set includes articles from around the world
- Insights from analysis of unfiltered dataset not useful for collective intelligence and analysis process
- Data Modification and Data Query allows for a much more specific analysis based on:
 - Country
 - Events
 - Topics
 - Personnel of Interest

Save Query

Download Queried Data

Step 18: Download the queried focused data set



Project Objective



Application Overview



Analysis of PS1



Analysis of PS2



Other Features



Moving Forward

Analysis - Problem Statement 1

• 04 – NLP Toolkit (Analyse Sentiment)

NLP Toolkit

Module Description +

Upload Data

Origin of Data File ? Define the Data Input Format

Local CSV

Load CSV File

Drag and drop file here
Limit 1000GB per file • CSV

Singapore.csv 68.8MB

Browse files X

Step 19: Upload the previously queried focused data set

Choose Column where Data is Stored

CLEANED CONTENT

Data Loaded from CLEANED CONTENT!

NLP Operations

Select the NLP functions which you wish to execute.

Select the NLP Operation to execute

Analyse Sentiment

Analyse Sentiment Selected

Analysis - Problem Statement 1

• 04 – NLP Toolkit (Analyse Sentiment)

Sentiment Analysis

For this module, both the VADER and TextBlob models will be used to analyse the sentiment of the text you upload.

Options

Choose the Backend Engine Used to Conduct Sentiment Analysis

VADER

- Save Outputs?
- Override Output Format?
- Display Outputs?

Data points

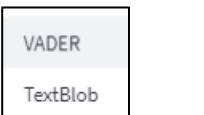
20

Choose Colour of Marker to Display



Display Advanced DataFrame Statistics? ?

Start Analysis



Step 20 Select the checkboxes to save and display the outputs

Sentiment DataFrame

	CONTENT	CLEANED CONTENT	COUNTRIES	VADER SENTIMENT
0	SINGAPORE: Prime Minister L prime minister lee leave two	['Singapore']	SINGAPORE Prime Mir	
1	As a small city-state at the foot of the small city state foot peninsul	['India', 'Mali', 'Malaysia', 'Omn	As a small city state at	
2	Please refer to the attachment please refer attachment ann	['Singapore']	Please refer to the atta	
3	March 2017 - Present Deputy march present deputy chief e	['India', 'Singapore']	March Present Deputy	
4	KUALA LUMPUR To spare the spare country expense worth	['Malaysia', 'Philippines', 'Sing	KUALA LUMPUR To spe	
5	Damage to the portside is vis damage portside visible miss	['Afghanistan', 'Mexico', 'Singa	Damage to the portsid	
6	On Monday morning, the wor monday morning world one	['Brazil', 'Bhutan', 'China', 'Inc	On Monday morning t	
7	A five-year-old Spanish boy d five year old spanish boy hos	['Singapore']	A five year old Spanish	
8	Published August 28, 2017, 6: august ben one championshi	['United Arab Emirates', 'Chin	Published August PM I	
9	(Aug 28): Bonds of a Dalian W bond group unit fell chinesi	['China', 'Hong Kong', 'Singap	Aug Bonds of a Dalian	
10	Inter Milan and Juventus botl inter milan came behind mai	['Singapore']	Inter Milan and Juvent	
11	PETALING JAYA (Aug 28): Eco world development group hij	['Austria', 'India', 'Malaysia', 'S	PETALING JAYA Aug Ec	
12	KUALA LUMPUR (Jan 2): The jan fell midday break first tra	['China', 'Malaysia', 'Singapor	KUALA LUMPUR Jan Ti	
13	A Malaysia Rincoit note is see note seen illustration photo in	['China', 'Malaysia', 'Singapor	A Malaysia Rincoit not	



Open Source Library

(VADER) - Valence Aware Dictionary and sEntiment Reasoner

(Recommended)

- Lexicon sentiment analysis (*created by experts*) tool. Able to accommodate slangs and emoticons (**Informal text** sentiments in social media)
- Score is between: -1 (most negative) and +1 (most positive)

TextBlob – Formal Text (Complementary)

Lexicon based sentiment analysis. (*created by experts*)

Sentiment Analysis Calculation Process

- Assigns scores to each word based on a pre-defined dictionary of positive and negative words. Average is used to calculate the overall sentiment.
- Polarity:** -1 (negative) to +1 (positive)
- Subjectivity:** 0 (objective) to 1 (subjective)



Analysis - Problem Statement 1

• 04 – NLP Toolkit (Analyse Sentiment)

Sentiment Analysis – Data frame view

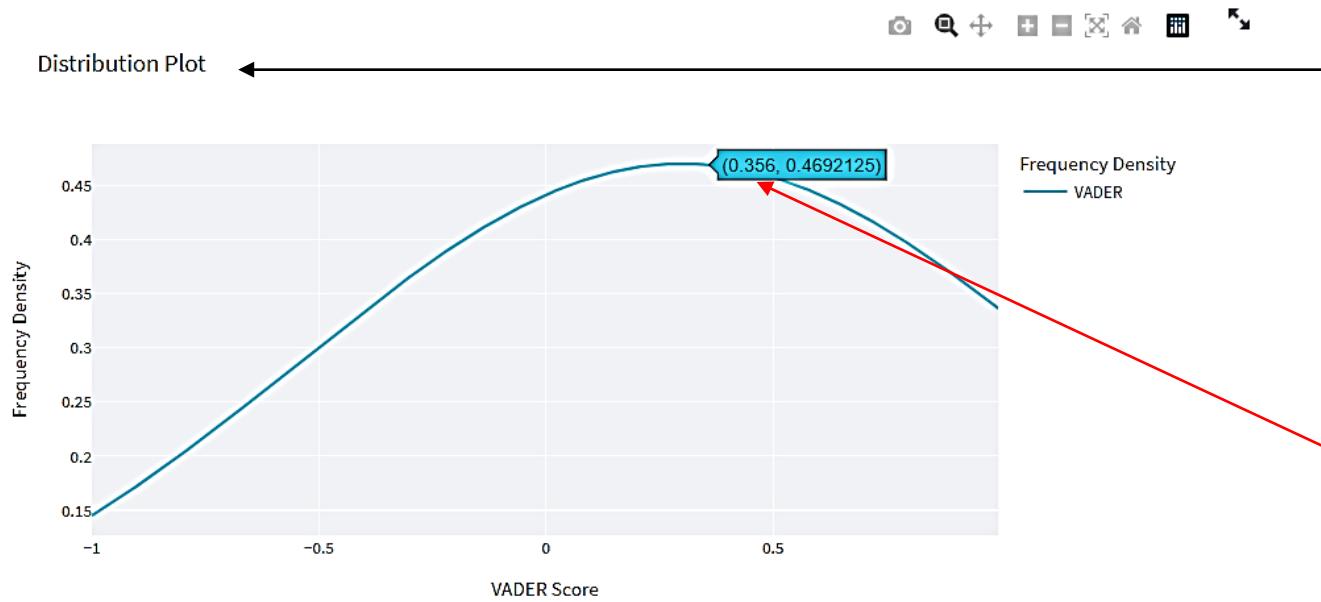
	CONTENT	CLEANED CONTENT	COUNTRIES	VADER SENTIMENT TEXT	VADER SENTIMENT	VADER SCORE	VADER OVERALL SCORE	VADER OVERALL SENTIMENT
0	SINGAPORE: Prime Minister Lee	prime minister lee leave two w	['Singapore']	SINGAPORE Prime Minister Lee	Negative	-0.4767	-0.4767	Negative
1	As a small city-state at the foot	small city state foot peninsular	['India', 'Mali', 'Malaysia', 'Omar']	As a small city state at the foot	Positive	0.9994	0.9998	Positive
2	Please refer to the attachment	please refer attachment annou	['Singapore']	Please refer to the attachment	Positive	0.7096	0.7430	Positive
3	March 2017 - Present Deputy C	march present deputy chief ex	['India', 'Singapore']	March Present Deputy Chief Ex	Positive	0.7783	0.7783	Positive
4	KUALA LUMPUR To spare the o	spare country expense worth p	['Malaysia', 'Philippines', 'Singa']	KUALA LUMPUR To spare the o	Positive	0.9738	0.9883	Positive
5	Damage to the portside is visib	damage portside visible missil	['Afghanistan', 'Mexico', 'Singap']	Damage to the portside is visib	Negative	-0.9854	-0.9862	Negative
6	On Monday morning, the world	monday morning world one ge	['Brazil', 'Bhutan', 'China', 'Indi']	On Monday morning the world	Negative	-0.875	-0.9749	Negative
7	A five-year-old Spanish boy die	five year old spanish boy hospi	['Singapore']	A five year old Spanish boy die	Positive	0.1027	-0.5423	Negative
8	Published August 28, 2017, 6:1	august ben one championship	['United Arab Emirates', 'China']	Published August PM Ben Askr	Positive	0.9991	0.9993	Positive
9	(Aug 28): Bonds of a Dalian Wa	bond group unit fell chinese m	['China', 'Hong Kong', 'Singapo']	Aug Bonds of a Dalian Wanda C	Negative	-0.5106	-0.9485	Negative
10	Inter Milan and Juventus both	inter milan came behind maint	['Singapore']	Inter Milan and Juventus both	Positive	0.0772	0.1280	Positive
11	PETALING JAYA (Aug 28): Eco W	world development group high	['Austria', 'India', 'Malaysia', 'Sir']	PETALING JAYA Aug Eco World	Positive	0.9933	0.9930	Positive
12	KUALA LUMPUR (Jan 2): The FBM	jan fell midday break first tradi	['China', 'Malaysia', 'Singapore']	KUALA LUMPUR Jan The FBM F	Negative	-0.7184	-0.7717	Negative
13	A Malaysia Ringgit note is seen	note seen illustration photo jur	['China', 'Malaysia', 'Singapore']	A Malaysia Ringgit note is seen	Negative	-0.6124	-0.2263	Negative
14	Nearly half of 100 countries ev	nearly half country may good f	['Australia', 'Bhutan', 'China', 'C']	Nearly half of countries evalua	Negative	-0.9687	-0.9861	Negative
15	When liberals talk about their l	liberal talk health care utopia s	['Canada', 'Germany', 'France', 'I']	When liberals talk about their l	Positive	0.9998	0.9999	Positive
16	MANILA, July 31 (Xinhua) -- Chi	manila july china association s	['China', 'Indonesia', 'Cambodia']	MANILA July Xinhua China and	Positive	0.9902	0.9939	Positive
17	- Advertisement - An article on	advertisement article economi	['United Kingdom', 'Singapore']	Advertisement An article on Th	Positive	0.6214	-0.8313	Negative
18	If you could trace the exact mo	could trace exact moment aust	['Australia', 'China', 'Singapore']	If you could trace the exact mo	Positive	0.9747	-0.0936	Negative
19	HRjobs: Asia's only regional re	regional recruitment job post v	['Singapore']	HRjobs Asias only regional reci	Positive	0.995	0.9955	Positive



Analysis - Problem Statement 1

• 04 – NLP Toolkit (Analyse Sentiment)

VADER Score



Data Visualisation – Normal Distribution

- Normally distributes the sentiment score in the previous data frame.
- The vertex (peak) of the distribution shows the average sentiment score of the entire data set
- In this case, the Singapore data set.

Download Data

[Download Sentiment Score Data](#)

Step 22 Hit to download the Sentiment Score data for Singapore data set

Graphs

[Download Plot](#)



Project Objective



Application Overview



Analysis of PS1



Analysis of PS2



Other Features



Moving Forward

Analysis - Problem Statement 1

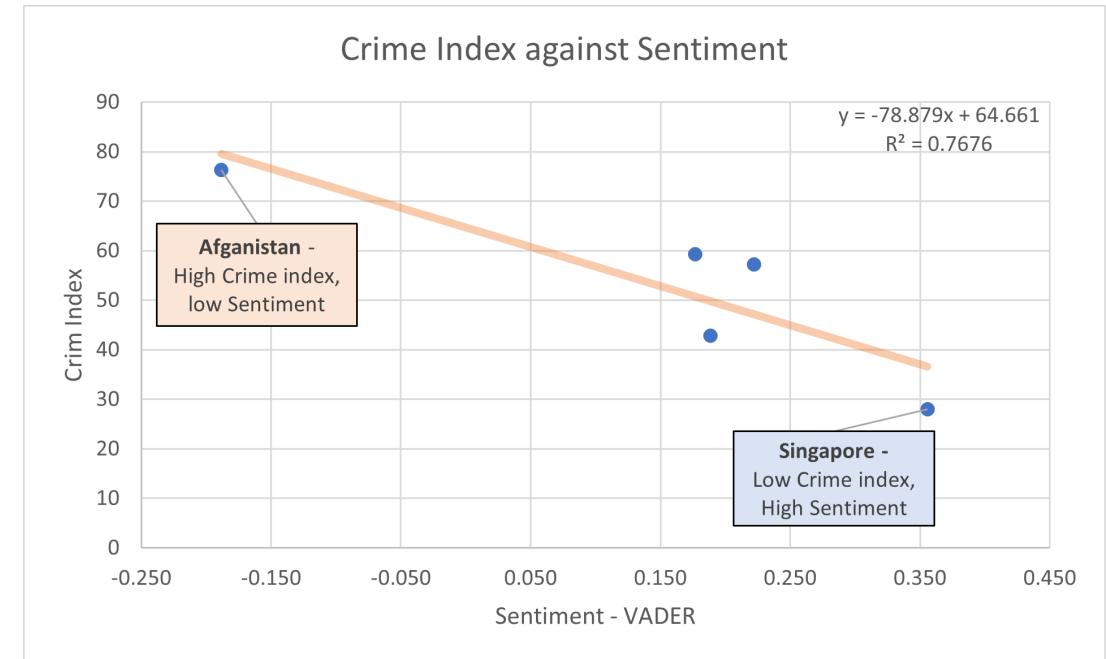
- Insights into Problem Statement 1

Problem Statement 1

Analyse public sentiment through news articles to investigate whether public sentiment correlate to the crime Index

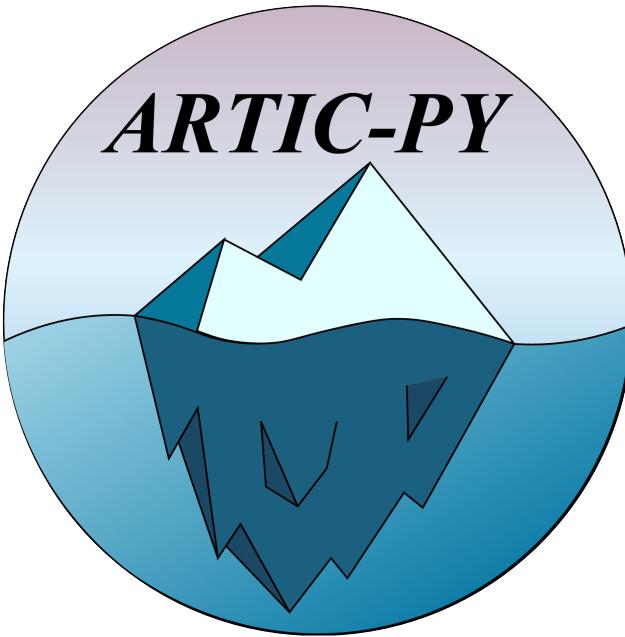
Insights

- Based on the graph, there is an **inverse correlation between crime rate and public sentiment**
- It is also observed that while New Zealand has a lower sentiment than Zimbabwe, it has a lower crime index than Zimbabwe.
- This shows that public sentiment is not the only factor affecting crime rate.
- Malaysia and Zimbabwe are similar in crime index. Correspondingly they have similar public sentiments.
- Observations validates the model used



Countries	Crime Index	Sentiment x100
Singapore	28.0	35.6
New Zealand	42.9	18.9
Malaysia	57.3	22.2
Zimbabwe	59.3	17.7
Afghanistan	76.3	-18.8

Range of selected countries across the crime index for comparison



Analysis of Problem Statement 2

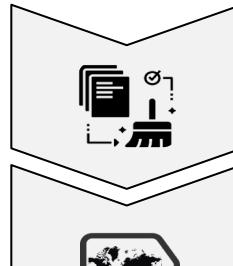
Analysis - Problem Statement 2

- Overview Procedure

Problem Statement 2

To analyze public sentiments from news articles to identify topics or issues within society.

01 Data Cleaning



Raw data is pre-processed using Natural Language Processing Techniques to eliminate irrelevant details

02 Data Modification



Tagging of the data according to its relevant countries

03 Data Querying



Querying the tagged data to extract a more focused dataset based on certain keyword(s) –
E.g. by Country

04 Word Cloud



Generating a word cloud with the most frequent words for visualisation

05 Additional Data Cleaning



Additional Data Cleaning to add more stop words to be cleaned from the data set, before re-generating the word cloud





Analysis - Problem Statement 2

• 04 – NLP Toolkit (Word Cloud)

NLP Toolkit

Module Description

An app built to simplify and condense NLP tasks into one simple yet powerful Interface.

App Modules
Select the following available modules:

NLP Functions

NLP Toolkit

Upload Data

Origin of Data File: Local | Define the Data Input Format: CSV

Load CSV File:

- Drag and drop file here (Limit 1000GB per file • CSV)
- Singapore.csv 68.6MB

Choose Column where Data is Stored:

- Topic Modelling
- Topic Classification
- Analyse Sentiment
- Word Cloud**
- Named Entity Recognition
- POS Tagging
- Summarise

Word Cloud

Word Cloud Selected

Step 1: Upload the Singapore file

Step 2: Choose the Word Cloud function

Word Cloud Generation

This module takes in a long list of documents and converts it into a WordCloud representation of all the documents.

Project Objective Application Overview Analysis of PS1 **Analysis of PS2** Other Features Moving Forward

Analysis - Problem Statement 2

• 04 – NLP Toolkit (Word Cloud)

Word Cloud Generation

This module takes in a long list of documents and converts it into a WordCloud representation of all the documents.

Note that the documents should not be tokenized, but it should be cleaned and lemmatized to avoid double-counting words.

Step 3: Check box to enable saving of output

Save Outputs? (?)

Key in the maximum number of words to display
200

Key in the contour width of your WordCloud
3

Key in the Width of the WordCloud image generated
800

Key in the Height of the WordCloud image generated
400

Step 4: Hit to generate word cloud

Step 5: Visually inspect the word cloud

Wordcloud For Text Inputted

Download Image

Download Word Cloud

Customising - Word Cloud Using the widgets

1. Maximum number of words displayed
2. Contour Width (density of words)
3. Width & Height of the word cloud generated

Quick view of the word cloud generated

Insights:
Some irrelevant words are still included after the cleaning such as

- Said, u, may, will say

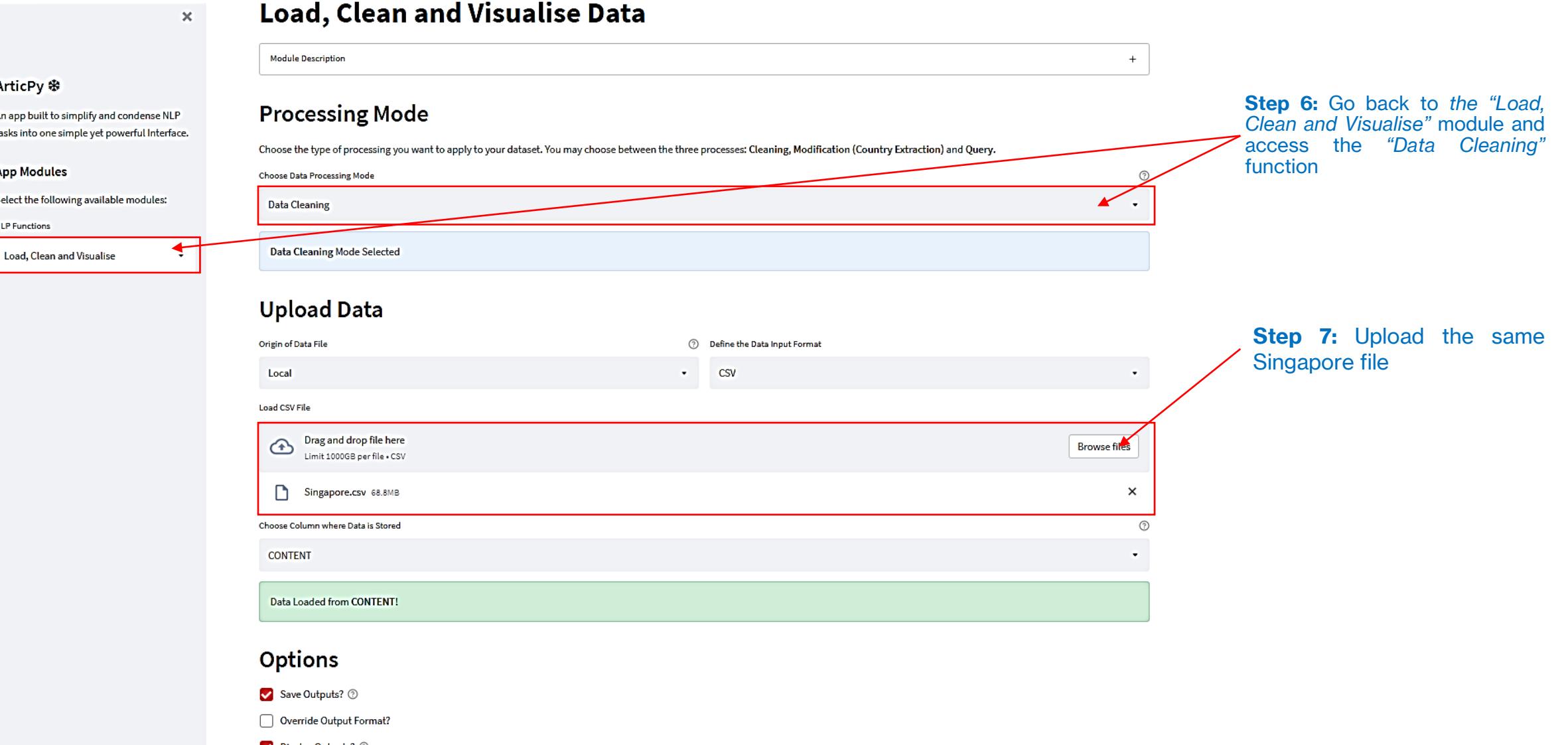
Analysis - Problem Statement 2

- 05 – Load, Clean and Visualise (**Data Cleaning**)

Load, Clean and Visualise Data

Step 6: Go back to the “Load, Clean and Visualise” module and access the “Data Cleaning” function

Step 7: Upload the same Singapore file



ArcticPy *
An app built to simplify and condense NLP tasks into one simple yet powerful Interface.

App Modules
Select the following available modules:

NLP Functions

Load, Clean and Visualise

Processing Mode
Choose the type of processing you want to apply to your dataset. You may choose between the three processes: Cleaning, Modification (Country Extraction) and Query.

Choose Data Processing Mode
Data Cleaning

Data Cleaning Mode Selected

Upload Data

Origin of Data File
Local Define the Data Input Format
CSV

Load CSV File
Drag and drop file here Limit 1000GB per file • CSV
Browse files
Singapore.csv 68.8MB

Choose Column where Data is Stored
CONTENT

Data Loaded from CONTENT!

Options

Save Outputs? ⓘ
 Override Output Format?
 Display Outputs? ⓘ

Analysis - Problem Statement 2

• 05 – Load, Clean and Visualise (**Data Cleaning**)

ArticPy *

An app built to simplify and condense NLP tasks into one simple yet powerful Interface.

App Modules

Select the following available modules:

NLP Functions

Load, Clean and Visualise

Data Loaded from CONTENT!

Options

Save Outputs? ⓘ

Override Output Format?

Display Outputs? ⓘ

Step 8: Select the checkboxes to save and display the outputs

Data Points To Print
20

Display Advanced DataFrame Statistics? ⓘ

Select Preprocessing Pipelines
Complex

Tokenize Data? ⓘ

Extend List of Stopwords? ⓘ

Step 9: Check the box to extend the list of Stop words

Keyword List
Said X said X say X u X may X will X s X Press Enter to extend list...

Step 10: Check the box to extend the list of Stop words

☞ Data Cleaning and Visualisation

Ensure that you have successfully uploaded the required data and selected the correct column containing your data before clicking on the "Begin Analysis" button. The status of your file upload is displayed below for your reference.

File loaded.

Begin Analysis

Step 11: Hit to begin the additional data cleaning process & Save the Output

Analysis - Problem Statement 2

• 04 – NLP Toolkit (Word Cloud)

NLP Toolkit

Module Description

Upload Data

Origin of Data File

Define the Data Input Format

Local

CSV

Load CSV File

Drag and drop file here
Limit 1000GB per file • CSV

 Singapore_stopword.csv 86.0MB

Choose Column where Data is Stored

CLEANED CONTENT

Data Loaded from CLEANED CONTENT!

NLP Operations

Select the NLP functions which you wish to execute.

Select the NLP Operation to execute

Word Cloud

Word Cloud Selected

Step 12: Upload the Cleaned Singapore dataset

Browse files

Step 13: Ensure CLEANED CONTENT column is chosen for the analysis

Step 14: Output and save the data



Libraries wordcloud

NLP Toolkit - Word Cloud

Description: Visual Analytic Tool

Purpose: Create a word cloud with the most frequent words for easy visualisation

Benefit:

Produces easy to grasp images,
Casual and visually appealing way to visualise data

Limitations: Lack of context of the word



Saved Output - Singapore



Project Objective



Application Overview



Analysis of PS1



Analysis of PS2



Other Features



Moving Forward

Analysis & Summary

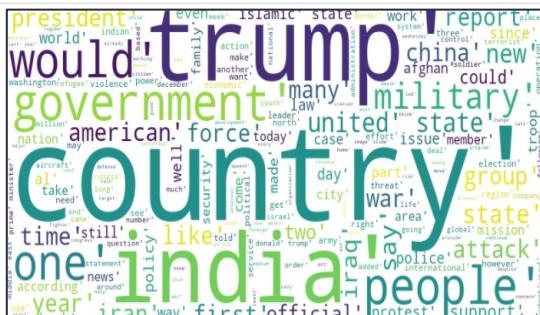
- Insights into Problem Statement 2 & Summary of Problem Statements 1 & 2

Problem Statement 2

To analyze public sentiment through news articles to identify topics/issues that affect public sentiment

Afghanistan

1. Lower Sentiment due to current events.
2. US exit from Afghanistan causing instability in the country
3. Afghans lost of trust with their government,



Project Objective



Application Overview



Analysis of PS1

Conclusion

- Analysis shows - public sentiment correlates to a country's crime index
- **Important for SPF to regularly monitor public sentiment as part of horizon sensing**
- The word cloud can be used to identify emerging issues of concern and to enhance current positive sentiments
- Facilitate SPF's collective intelligence and decision-making process.
- ARTIC-PY can be used to complement ground sensing efforts by SPF.



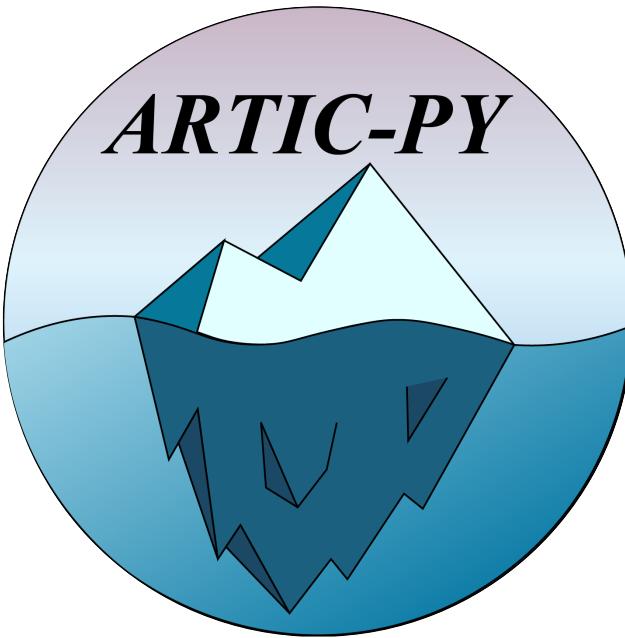
Analysis of PS2



Other Features



Moving Forward



Other Features

Other Features – Text Summarizer

Summarization of Text

For this function, you are able to upload a piece of document or multiple pieces of documents in a CSV file to create a summary for the documents of interest.

However, do note that this module takes a long time to process a long piece of text, hence, it may be better to process your data in smaller batches to speed up your workflow.

In an effort to enhance this module to provide users with meaningful summaries of their document, we have implemented two modes of summarization in this function, namely Basic and Advanced Mode.

Summary Complexity

Basic Mode uses the spaCy package to distill your documents into the specified number of sentences. No machine learning model was used to produce a unique summary of the text.

Advanced Mode uses the PyTorch and Huggingface Transformers library to produce summaries using Google's T5 Model.

Choose Mode

Step 1: Choose Advanced

Options

Choose the minimum and maximum number of words to summarise to below. If you are an advanced user, you may choose to modify the number of input tensors for the model. If you do not wish to modify the setting, a default value of 512 will be used for your summary.

If your system has a GPU, you may wish to install the GPU (CUDA) enabled version of PyTorch. If so, click on the expander below to install the correct version of PyTorch and to check if your GPU is enabled.

Step 2: Install the required CUDA packages imply by hitting a button in the widget

Step 3: Check to save and display outputs

Step 4: Hit to summarise text



Basic

Text Extractive Summarization –
Weighing sentences using the an open-source library, concatenating them to form a summary

Advanced

Text Abstractive Summarisation –
Generating new sentences using a Machine Learning algorithm to create a summary

Customizing – Advanced Summary

Using the widgets

1. Minimum number of words to summarize
2. Maximum number of words to summarize
3. Number of vectors to consider



Other Features – Text Summarizer

Summarised Text

	RAW	CLEANED	SUMMARISED
0	TOKYO -Back in 2004, Japan came back japan came r environmental i	Shinjiro Koizumi is the son of forme	
1	COLOMBO (AP) - Sri Lankan Preside sri president said thursday become Sri Lankan president says he is the		
2	MOSCOW (AP) An unidentified gun moscow unidentified gunman fire t the FSB says the assailant was acti		
3	Norway's right-wing Progress Party norway right wing progress party g the progress party pulls out of the g		
4	ad with plans for a new tax on Ame ad plan new tax american tech gear threat escalates an ongoing row be		
5	Photo Credit: Yossi Zamir/Flash90 photo credit flash ministerial comm the ministerial committee on privat		
6	Everybody likes a little treat now at everybody like little treat recently! not all kids love sweets, but it's a go		
7	A SmartyPants Vitamin Gold Box, E vitamin gold box security system m this week's best deals include a bac		
8	Jan. 21 (UPI) -- Former football star jan former football star current mir former football star and current mi		
9	Disney calls it the future of the corr call future company plus entertain Disney Plus is the entertainment gi		
10	Google Stadia, which launched in N stadium november company big ju google Stadia is a cloud-gaming sul		
11	Scientists have identified the oldes scientist known impact crater earth scientists have identified the oldest		
12	A record-setting spacewalker, one c record setting one two woman spa astronauts will be inducted into the		
13	If you want to keep your licensi s want keep simple sure appeal base the latest version of the software sl		
14	Barts Health NHS Trust has turned bart health trust turned help estate the three-year agreement will see C		
15	Fun fact: That's also how animals le fun fact also animal learn put cat fr if you put a cat in front of two door:		
16	INILAHCOM, Seoul - Perusahaan tel electronics dan chip artificial intellig Baidu KUNLUN adalah proyek yang		
17	Published: 17:17 BST, 8 August 2018 august august crown prince bin ove foreign minister: "Canada made a t		
18	Isobel Asher Hamilton, Business In: asher hamilton business insider u c source YouTube YouTube has uploa		
19	Published: 16:43 BST, 8 August 2018 august august columbia c south ca Zakaryia abdin, 19, pleaded guilty t		

Type	Text	Word Count
Original	TOKYO -Back in 2004, Japan came up with the "3R" environmental initiative to reduce, reuse and recycle under then Prime steps, if 100 people take just one step forward and make our efforts collectively, we can achieve a zero-carbon future."	764 words
Basic	" Experts have also said there appears to be very little political will in Japan to abandon toxic coal plants owing progress has been slow partly because of the gulf between how Japan and the world see energy and environment issues.	127 words
Advanced	Shinjiro Koizumi is the son of former prime minister Junichiro Koizumi. the 38-year-old wants the country to reclaim global leadership but he says he is eager to look at how his country might be able to wean itself off coal and nuclear power entirely.	72 words

Purpose of Summarizer Function

- To be used at the start when the raw data is first received
- Saves on reading time as it reduces the number of words by providing a summary
- Limitations: Might miss out on some relevant information

Download Summarised Data

Download Summarised Data



Project Objective



Application Overview



Analysis of PS1



Analysis of PS2



Other Features



Moving Forward

Other Features – Topic Modelling

Topic Modelling

Ensure that your data is lemmatized and properly cleaned; data should not be tokenized for this step. Use the Load, Clean and Visualise module to clean and lemmatize your data if you have not done so already.

Topic Modelling Model Selection

Choose Model to use

- Latent Semantic Indexing (selected)
- Latent Semantic Indexing Selected
- Short Explanation on Models used

Step 1: Choose LSI

Options

Save Outputs?
 Override Output Format?
 Display Outputs? ⓘ

Step 2: Check to save and display outputs

Display Advanced DataFrame Statistics? ⓘ

Data points: 20
 Number of Topics to Generate: 10
 Maximum Number of Features to Extract: 5000
 Iterations of Model Training (Epochs): 10
 Generate LSI Plot?
 Generate Word Representation of LSI Plot?
 Choose Colour of Marker to Display: Blue

Step 3: Hit to start modelling



Latent Semantic Indexing (Recommended)

1. Able to identify correlation
2. Accounts multiple meanings for a word/phrase Reduces complexity & resource required

Non-Negative Matrix Factorisation (NMF)

1. Unsupervised learning to factorise the vectors within the data
2. Works best with shorter texts such as tweets or titles

Latent Dirichlet Allocation (LDA) Model

1. to Identify abstract probability distributions/groupings in data set
2. When you are unsure of what the topic are supposed to be

Other Features – Topic Modelling

Model Data

Topic 1			Topic 2			Topic 5			Topic 6		
0	word	weight	0	market	weight	0	case	weight	0	million	weight
1	said	0.3398	1	year	0.2758	1	north	0.1831	1	target	0.2489
2	case	0.2262	2	company	0.1780	2	virus	0.1801	2	met	0.2392
3	year	0.2065	3	million	0.1614	3	confirmed	0.1612	3	china	0.2253
4	new	0.1869	4	business	0.1266	4	trump	0.1468	4	year	0.1442
5	virus	0.1741	5	growth	0.1021	5	answer	0.1289	5	quarter	0.1325
6	country	0.1632	6	global	0.0911	6	read	0.1270	6	net	0.1071
7	people	0.1588	7	data	0.0904	7	state	0.1212	7	company	0.0873
8	chinese	0.1426	8	time	0.0875	8	kim	0.1046	8	financial	0.0826
9	health	0.1278	9	investment	0.0852	9	patient	0.1006	9	revenue	0.0764

Topic 3			Topic 4			Topic 7			Topic 8		
0	word	weight	0	china	weight	0	target	weight	0	year	weight
1	market	0.4801	1	north	0.2528	1	met	0.6658	1	wear	0.3712
2	china	0.1920	2	state	0.2118	2	said	0.1593	2	million	0.3326
3	case	0.1590	3	market	0.1844	3	market	0.1448	3	child	0.2939
4	virus	0.1472	4	trump	0.1664	4	wear	0.1664	4	sale	0.1319
5	wear	0.1419	5	united	0.1615	5	woman	0.0434	5	north	0.1306
6	global	0.1047	6	answer	0.1309	6	analysis	0.0404	6	corresponding	0.1281
7	million	0.0996	7	confirmed	0.1192	7	country	0.0397	7	table	0.1229
8	confirmed	0.0924	8	president	0.1192	8	table	0.0376	8	chart	0.1228
9	analysis	0.0910	9	nuclear	0.1186	9	graph	0.0359	9	graph	0.1181

Topic 9			Topic 10								
0	word	weight	0	macaw	weight	0	wear	weight	0	year	weight
1	wear	0.2992	1	actor	0.1662	1	scott	0.1630	1	child	0.1626
2	university	0.1662	2	scott	0.1630	2	virus	0.1626	2	child	0.1509
3	people	0.1630	3	virus	0.1626	3	people	0.1509	3	year	0.1440
4	china	0.1626	4	people	0.1440	4	know	0.1440	4	time	0.1411
5	child	0.1509	5	know	0.1411	5	market	0.1411	5	world	0.1411
6	year	0.1440	6	market	0.1411	6	yeah	0.1395	6	quot	0.1395
7	time	0.1411	7	yeah	0.1395	7	tax	0.1395	7	like	0.1359
8	world	0.1411	8	tax	0.1395	8	like	0.1359	8	quot	0.1237
9	quot	0.1395	9	like	0.1359	9	like	0.1237	9	like	0.1120

LSI Topic DataFrame

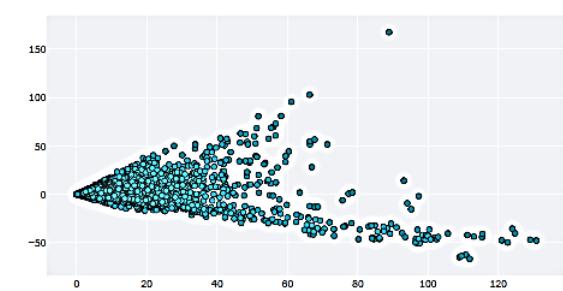
	word_0	word_1	word_2	word_3	word_4	word_5	word_6	word_7
topic_0	china	said	case	year	new	virus	country	people
topic_1	market	year	company	million	business	growth	global	data
topic_2	market	china	case	virus	wear	global	million	confirm
topic_3	china	north	state	market	trump	united	answer	preside
topic_4	case	north	virus	confirmed	trump	answer	read	state
topic_5	million	target	met	china	year	quarter	net	compar
topic_6	target	met	said	market	wear	woman	analysis	country
topic_7	year	wear	million	child	sale	north	corresponding	table
topic_8	wear	university	people	china	child	year	time	world
topic_9	macaw	actor	scott	virus	people	know	market	yeah

LSI(SVD) Scatterplot

Note that the following visualisation will use a Topic count of 2, overriding previous inputs above, as you are only able to visualise data on 2 axis. However the full analysis result of the above operation will be saved in the dataset you provided and be available for download later on in the app.

The main aim of the scatterplot is to show the similarity between topics, which is measured by the distance between markers as shown in the following diagram. The diagram contained within the expander is the same as the marker diagram, just that the markers are all replaced by the topic words the markers actually represent.

Scatter Plot



Topic Modelling

1. Insights: Some of the topic groupings do not make sense.
2. Working on increasing the accuracy
3. Can be used to compliment word cloud in identifying issues in the community
4. Identify topics > Query the words > Evaluate Sentiment score





Position of Speech (POS) Tagging & Named Entity Recognition (NER)

Libraries
displacy **spaCy**  pandas

POS Tagging: Tagging words corresponding with a particular part of speech depending on the definition of word and its context (*E.g. noun, verb*)

Benefits - POS

1. POS tag assigned to the next word depends on the previous word (accommodates words with multiple meanings in different contexts)
2. Search for grammatical or lexical patterns

POS Tagging: Tagging words corresponding with a particular part of speech depending on the definition of word and its context (*E.g. noun, verb*)

Benefits – NER

1. Suited to provide a quick overview for large quantities of text.
2. Tags allow users to understand the subject and theme of the text
3. Users can group text based on relevance easily
4. Improve relevance of search (*especially through querying*)

Other Features – NER & POS

- Named Entity Recognition (NER)

Note that this module takes a long time to process a long piece of text. If you intend to process large chunks of text, prepare to wait for hours for the NER Tagging process to finish.  RUNNING... 

In the meantime, it may be better to process your data in smaller batches to speed up your workflow.

Options

Due to limits imposed on the visualisation engine and to avoid cluttering of the page with outputs, you will only be able to visualise the NER outputs for a single document. However, you will still be able to download a text/html file containing the outputs for you to save onto your disks.

NLP Models

Select one model to use for your NLP Processing. Choose en_core_web_sm for a model that is optimised for efficiency or en_core_web_lg for a model that is optimised for accuracy.

Select spaCy model

en_core_web_sm en_core_web_lg

Step 1: Choose the "lg" accuracy model

Display Outputs?

Choose Number of Data Points to Display
20 0 1000

Step 2: Choose the "lg" accuracy model

Visualise One Data Point?

You are conducting NER on the entire dataset. Only the final DataFrame is printed. NER output will be automatically saved.

Display Advanced DataFrame Statistics? 

Save Outputs? 

Override Output Format?

Step 3: Hit to start modelling

Conduct Named Entity Recognition

Accuracy Model:

En_core_web_lg

- Requires more resources and processing time.
- More Accurate

Efficiency Model:

En_core_web_sm

- Requires less resources and time
- Allows for a quick view of the processed result

Other Features – NER & POS

- Named Entity Recognition (NER) - Output

DisplaCy Rendering

If rendering is not clean, choose to save files generated and download the rendering in HTML format.

```
<body style="font-size: 16px; font-family: -apple-system, BlinkMacSystemFont, 'Segoe UI', Helvetica, Arial, sans-serif, 'Apple Color Emoji', 'Segoe UI Emoji', 'Segoe UI Symbol';">
```

SINGAPORE GPE :

Prime Minister Lee Hsien Loong PERSON will be on leave for two weeks DATE from Saturday DATE (Dec 21 DATE) to Jan 3 DATE , according to a statement from the Prime Ministers Office ORG on Friday DATE .

During his absence, Deputy Prime Minister and Finance Minister Heng Swee Keat PERSON will be Acting Prime Minister from Dec 25 to Jan 3 DATE .

Advertisement Advertisement "

I will be on leave for the rest of the year DATE , " Mr Lee PERSON said in a Facebook ORG post on Friday DATE .

He added he would use his break to understand a mathematics problem called the Collatz LOC conjecture. "

Also plan to catch up on my other readings, spend time with family, and perhaps go #jalanjalan (if the rain stays away)," he added.

NER DataFrame

	CONTENT	NER	COMPILED_LABELS
0	SINGAPORE: Prime Minister Lee Hsi	[('SINGAPORE', 'GPE'), ('Lee Hsien L', 'PERSON'), ('GPE', 'LOC'), ('ORG', 'DATE')]	
1	As a small city-state at the foot of p	[('Malaysia', 'GPE'), ('Singapore', 'G', 'PERSON'), ('EVENT', 'DATE'), ('PERSON', 'MONEY')]	
2	Please refer to the attachment. -----	[('Company', 'ORG'), ('Company', 'C', 'PERSON'), ('DATE', 'FAC'), ('WORK_OF', 'WORK_OF')]	
3	March 2017 - Present Deputy Chief	[('March 2017 - Present', 'DATE'), ('C', 'GPE'), ('ORG', 'DATE')]	
4	KUALA LUMPUR To spare the count	[('KUALA', 'GPE'), ('the Philippine O', 'PERSON'), ('EVENT', 'DATE'), ('PERSON', 'ORDINA')]	
5	Damage to the portside is visible as	[('Guided', 'ORG'), ('USS John S. Mc', 'PERSON'), ('EVENT', 'PRODUCT'), ('DATE', 'PERSON')]	
6	On Monday morning, the world has	[('Monday', 'DATE'), ('morning', 'TIME'), ('TIM', 'PERSON'), ('TIME', 'CARDINAL')]	
7	A five-year-old Spanish boy died in	[('five-year-old', 'DATE'), ('Spanish', 'PERSON'), ('DATE', 'PERSON'), ('TIME', 'GPE'), ('NO')]	
8	Published August 28, 2017, 6:15 PM	[('August 28, 2017', 'DATE'), ('6:15 P', 'TIME'), ('EVENT', 'PRODUCT'), ('PERSON', 'DA')]	
9	(Aug 28): Bonds of a Dalian Wanda	[('Aug 28', 'DATE'), ('Dalian Wanda C', 'PERSON'), ('PRODUCT', 'PERSON'), ('DATE', 'MON')]	
10	Inter Milan and Juventus both cam	[('Inter Milan', 'ORG'), ('Juventus', 'C', 'PERSON'), ('DATE', 'MONEY'), ('ORDIN')]	
11	PETALING JAYA (Aug 28): Eco World	[('PETALING JAYA', 'PERSON'), ('Aug 28', 'DATE'), ('PERSON', 'PERCENT')]	
12	KUALA LUMPUR (Jan 2): The FBM K	[('KUALA LUMPUR', 'GPE'), ('FBM KL', 'DATE'), ('PERCENT', 'PERSON'), ('ORD')]	
13	A Malaysia Ringgit note is seen in th	[('Malaysia', 'GPE'), ('June 1, 2017', 'DATE'), ('PERSON', 'ORDINAL'), ('MON')]	
14	Nearly half of 100 countries evalua	[('Nearly half', 'CARDINAL'), ('100', 'NUMBER'), ('DATE', 'PERSON'), ('ORDINAL'), ('MON')]	
15	When liberals talk about their healt	[('Frances', 'PERSON'), ('1', 'CARDINAL'), ('PRODUCT', 'PERSON'), ('PERCENT', 'PERSON')]	
16	MANILA, July 31 (Xinhua) -- China a	[('MANILA', 'ORG'), ('July 31', 'DATE'), ('PERSON', 'CARDINAL'), ('LAW')]	
17	- Advertisement - An article on The	[('Singapore', 'ORG'), ('Singapore', 'PERSON'), ('ORDINAL'), ('WOR')]	
18	If you could trace the exact momen	[('Australia', 'GPE'), ('China', 'GPE'), ('EVENT', 'PERSON'), ('DATE', 'LANGUAGE')]	
19	HRjobs: Asia's only regional recruit	[('Asia', 'LOC'), ('100%', 'PERCENT'), ('DATE', 'PERCENT'), ('MONEY', 'PERSON')]	

Other Features – NER & POS

- Position of Speech (POS) - Output

ArticPy *

An app built to simplify and condense NLP tasks into one simple yet powerful Interface.

App Modules

Select the following available modules:

NLP Functions

NLP Toolkit

NLP Models

Select one model to use for your NLP Processing. Choose en_core_web_sm for a model that is optimised for efficiency or en_core_web_lg for a model that is optimised for accuracy.

Select spaCy model

- en_core_web_sm
- en_core_web_lg

Step 1: Choose the "lg" accuracy model

Options

- Display Outputs?

Choose Number of Data Points to Display

20
0

POS DataFrame

	CONTENT	POS	COMPILED_LABELS
0	SINGAPORE: Prime Minister Lee Hsi	['(SINGAPORE', 'PROPN'), (';', 'PUNC', 'DET', 'PART', 'ADJ', 'PROPN', 'PRON']	
1	As a small city-state at the foot of p	['(As', 'SCONJ'), ('a', 'DET'), ('small', 'DET', 'ADJ', 'SYM', 'PART', 'X', 'PROPN']	
2	Please refer to the attachment. -----	['(Please', 'INTJ'), ('refer', 'VERB'), ('-----', 'DET', 'PART', 'SYM', 'X', 'PROPN', 'NOUN']	
3	March 2017 - Present Deputy Chief	['(March', 'PROPN'), ('2017', 'NUM'), ('SYM', 'PROPN', 'SCONJ', 'NUM', 'VE')]	
4	KUALA LUMPUR To spare the count	['(KUALA', 'PROPN'), ('LUMPUR', 'PRON', 'DET', 'PART', 'ADJ', 'SPACE', 'PROPN']	
5	Damage to the portside is visible a	['(Damage', 'NOUN'), ('to', 'ADP'), ('the', 'DET', 'ADJ', 'SYM', 'PART', 'X', 'SPACE']	
6	On Monday morning, the world had	['(On', 'ADP'), ('Monday', 'PROPN'), ('had', 'DET', 'ADJ', 'PART', 'SPACE', 'PROPN']	
7	A five-year-old Spanish boy died in	['(A', 'DET'), ('five', 'NUM'), ('die', 'VERB'), ('in', 'DET', 'ADJ', 'PART', 'PROPN', 'PRON']	
8	Published August 28, 2017, 6:15 PM	['(Published', 'PROPN'), ('August', 'F', 'DET', 'ADJ', 'PART', 'SYM', 'X', 'PROPN']	
9	(Aug 28): Bonds of a Dalian Wanda	['(Aug', 'PUNCT'), ('Bonds', 'PROPN'), ('of', 'DET', 'ADJ', 'PART', 'SYM', 'SPACE', 'PROPN']	
10	Inter Milan and Juventus both cam	['(Inter', 'PROPN'), ('Milan', 'PROPN'), ('both', 'DET', 'ADJ', 'PART', 'SYM', 'ADJ', 'PROPN']	
11	PETALING JAYA (Aug 28): Eco World	['(PETALING', 'PROPN'), ('JAYA', 'PRON', 'PR', 'DET', 'ADJ', 'PART', 'X', 'SPACE', 'PROPN']	
12	KUALA LUMPUR (Jan 2): The FBM K	['(KUALA', 'PROPN'), ('LUMPUR', 'PRON', 'The', 'DET', 'ADJ', 'PART', 'PROPN', 'PRON']	
13	A Malaysia Ringgit note is seen in t	['(A', 'DET'), ('Malaysia', 'PROPN'), ('Ringgit', 'PROPN'), ('note', 'DET', 'ADJ', 'SYM', 'SPACE', 'PROPN']	
14	Nearly half of 100 countries evalua	['(Nearly', 'ADV'), ('half', 'NOUN'), ('of', 'DET', 'PART', 'ADJ', 'SYM', 'SPACE', 'COUNTRIES']	
15	When liberals talk about their healt	['(When', 'ADV'), ('liberals', 'NOUN'), ('talk', 'DET', 'PART', 'ADJ', 'SYM', 'SPACE', 'HEALTH']	
16	MANILA, July 31 (Xinhua) -- China a	['(MANILA', 'PROPN'), ('China', 'PROPN'), ('a', 'PUNCT'), ('--', 'DET', 'ADJ', 'PART', 'PROPN', 'PRON']	
17	- Advertisement - An article on The	['(-', 'PUNCT'), ('Advertisement', 'PROPN'), ('The', 'DET', 'ADJ', 'PART', 'PROPN', 'PRON']	
18	If you could trace the exact momen	['(If', 'SCONJ'), ('you', 'PRON'), ('could', 'DET', 'ADJ', 'PART', 'SYM', 'SPACE', 'MOMENT']	
19	HRjobs: Asia's only regional recruit	['(HRjobs', 'NOUN'), ('only', 'PUNCT'), ('Asia', 'DET', 'ADJ', 'PART', 'SYM', 'SPACE', 'RECRUIT']	

Renders are not shown due to the sheer size of the image. Kindly save the render in HTML format below to view it.

Step 3: Hit to start modelling

- Save Outputs? ②
- Override Output Format?

Start POS Tagging

1. Purpose
2. Insights: Some of the topic groupings do not make sense.
3. Working on increasing the accuracy
4. Can be used to compliment word cloud in identifying issues in the community
5. Identify topics > Query the words > Evaluate Sentiment score

Download Data

- [Download POS Data](#)
- [Download Rendering Data](#)

Step 4: Hit to start download the output data





Moving Forward

- ARTICPY in SPF & Possible Extensions

Other SPF Uses

1. Analyze the sentiment of data from News Articles and Social Media platforms to track public satisfaction of the SPF.
 - Identify areas of commendation and improvement
2. Gather Intelligence by querying data from News Articles and Social Media Platforms for Persons / Organisations of Interest.
3. Morale Sensing within PNSmen through Sentiment Analysis
4. Reduces the amount of time required to process Surveys and Feedback
 - Efficiently identifies topics/aspect that deserves commendation or improvement

Possible Extensions

1. Implementation of a database within ARTIC-PY's app. Allows users to pull and push data to the data base. This could increase the efficiency of the querying process
(Implementation of a database require financial resources)
2. Multi Language support -translate non-English text to English text *(Unable to implement due to time constraints)*
3. Sentiment analysis model used is not specifically trained for the detection of security related threats to Singapore.
(Accuracy largely limited by cleaning process.)

Thank you

