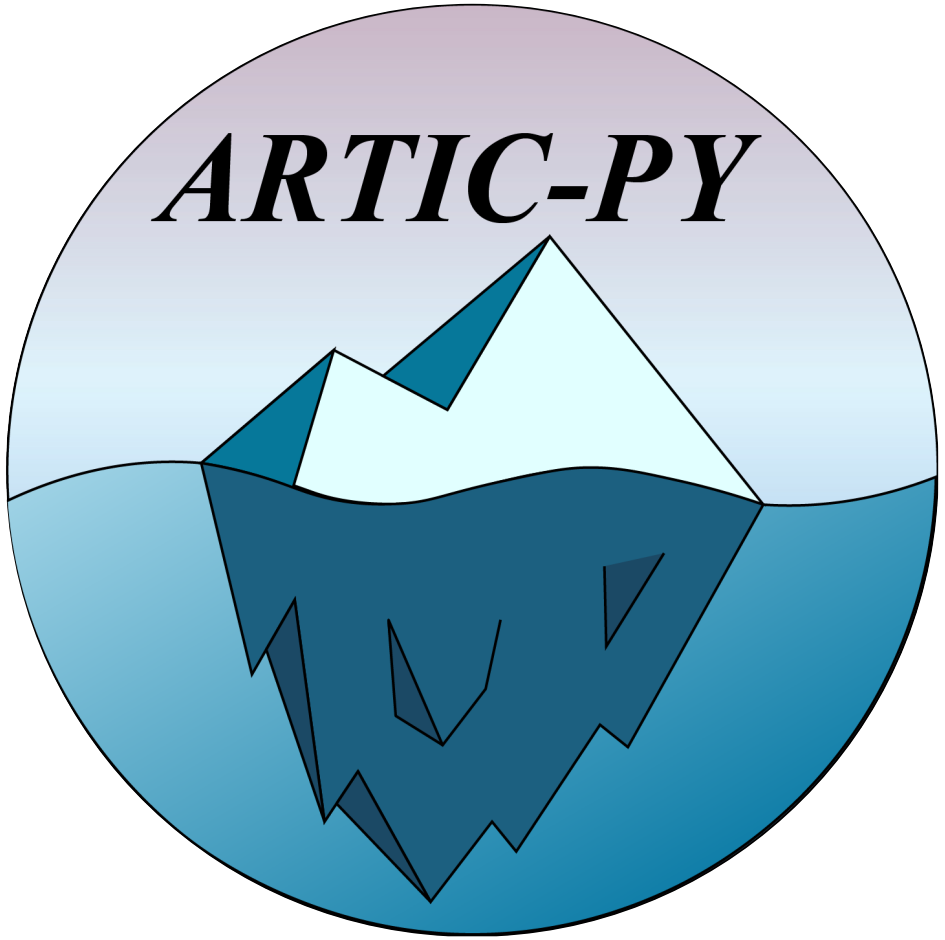
Coding Challenge 2021
Topic 2: News Analysis

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

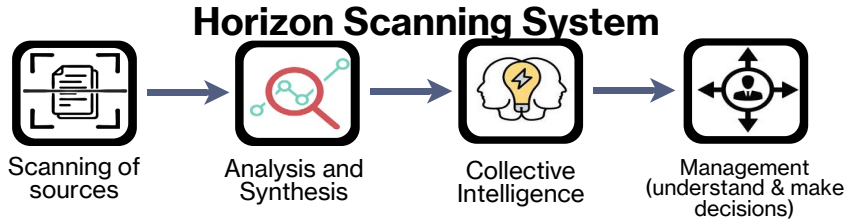
To analyse public sentiments from news articles to identify topics or issues within society.

Importance of understanding Public Sentiment

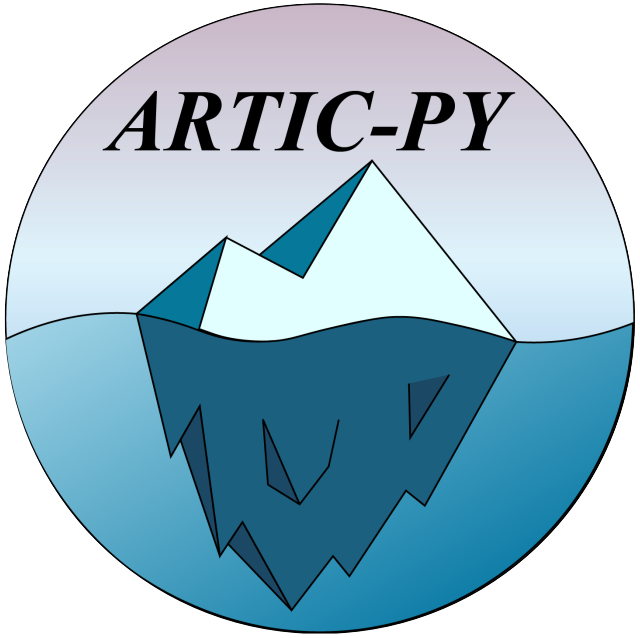
- Ground Sensing and surveys are used by officers most to understand sentiment
  - Communities that distrust police/ government often will not participate in such interactions. Concerns harder to detect

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

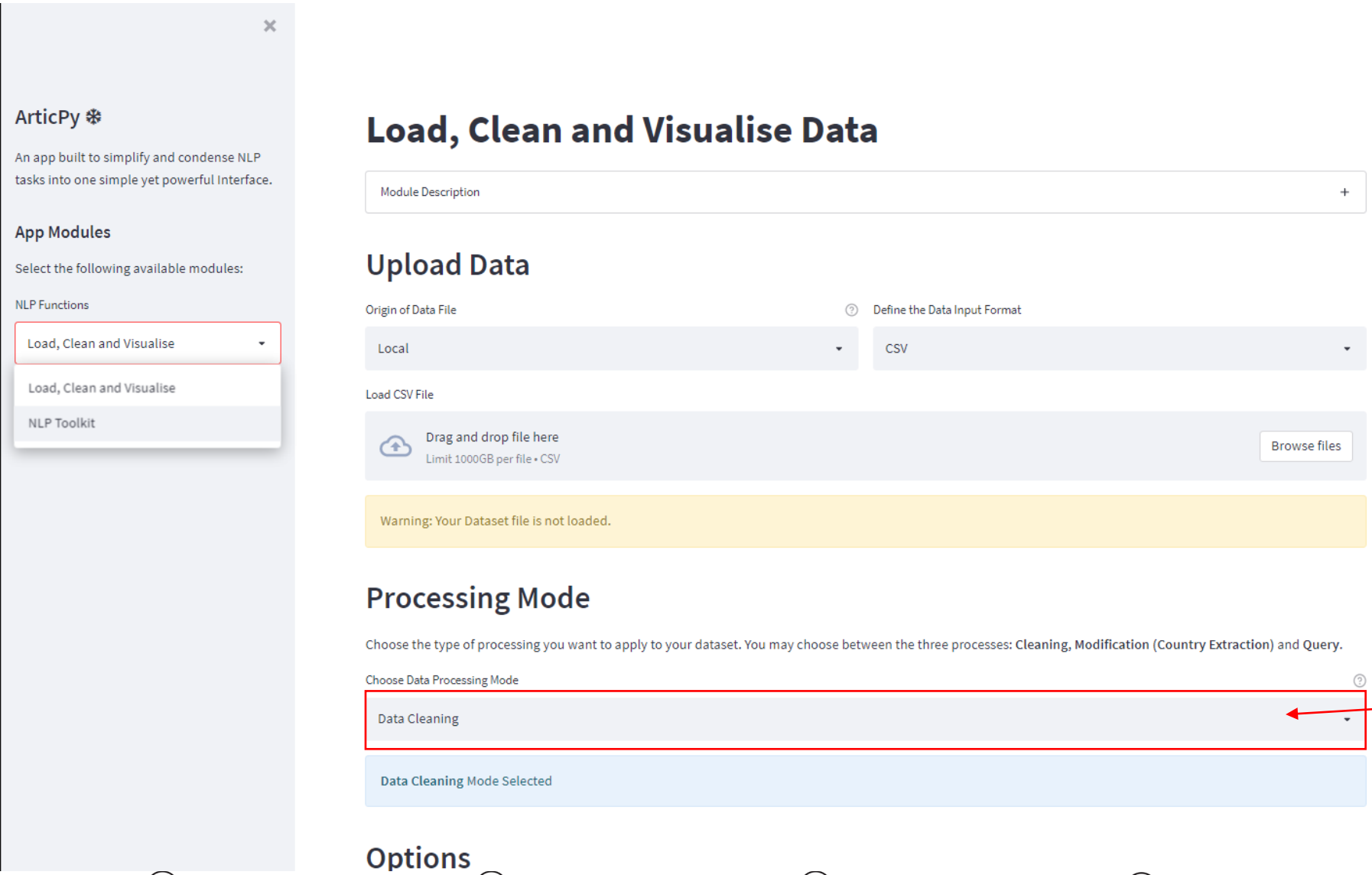
1. Very convenient to initiate the app

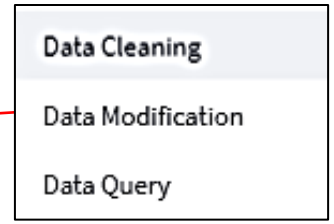
Limitations

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





ARTIC-PY

Program Language: Python3

Software Platform: Docker

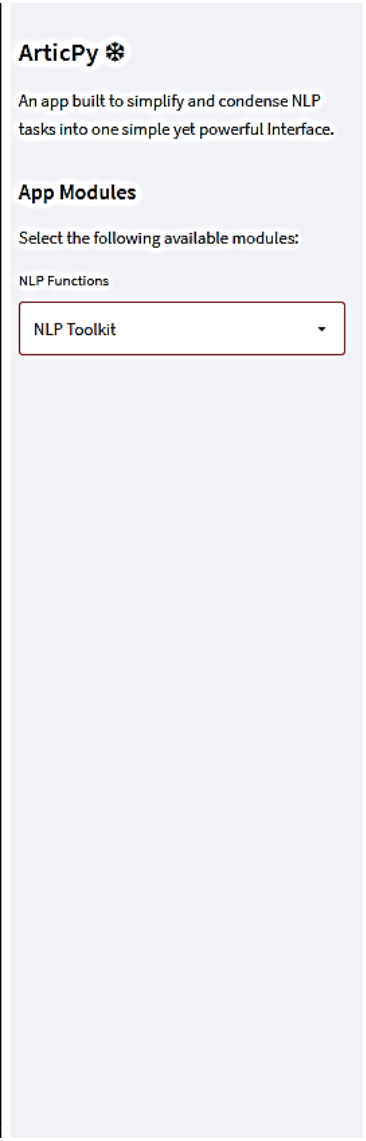
App UI: Streamlit

Module Functions

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

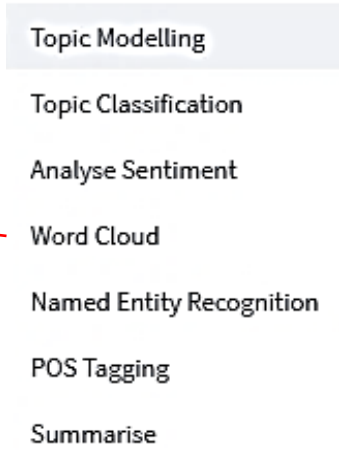
App Overview

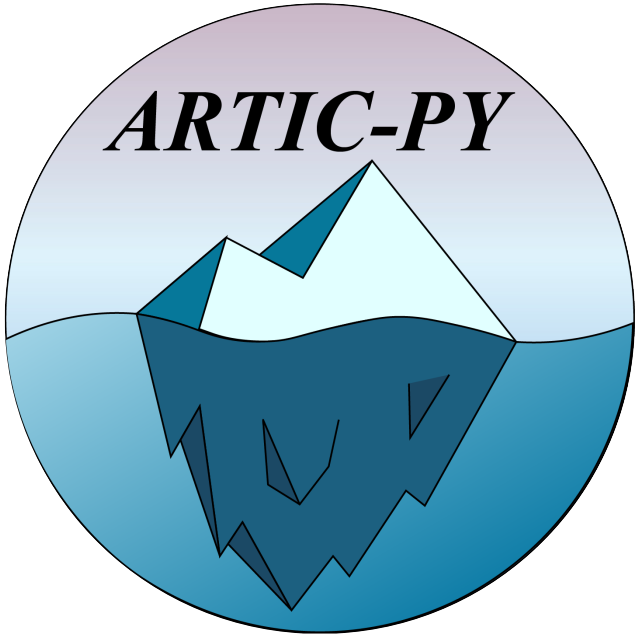
- Natural Language Processing Toolkit Module



Module Functions

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





Analysis of Problem Statement 1

Analysis – Problem Statement 1

- Overview Procedure

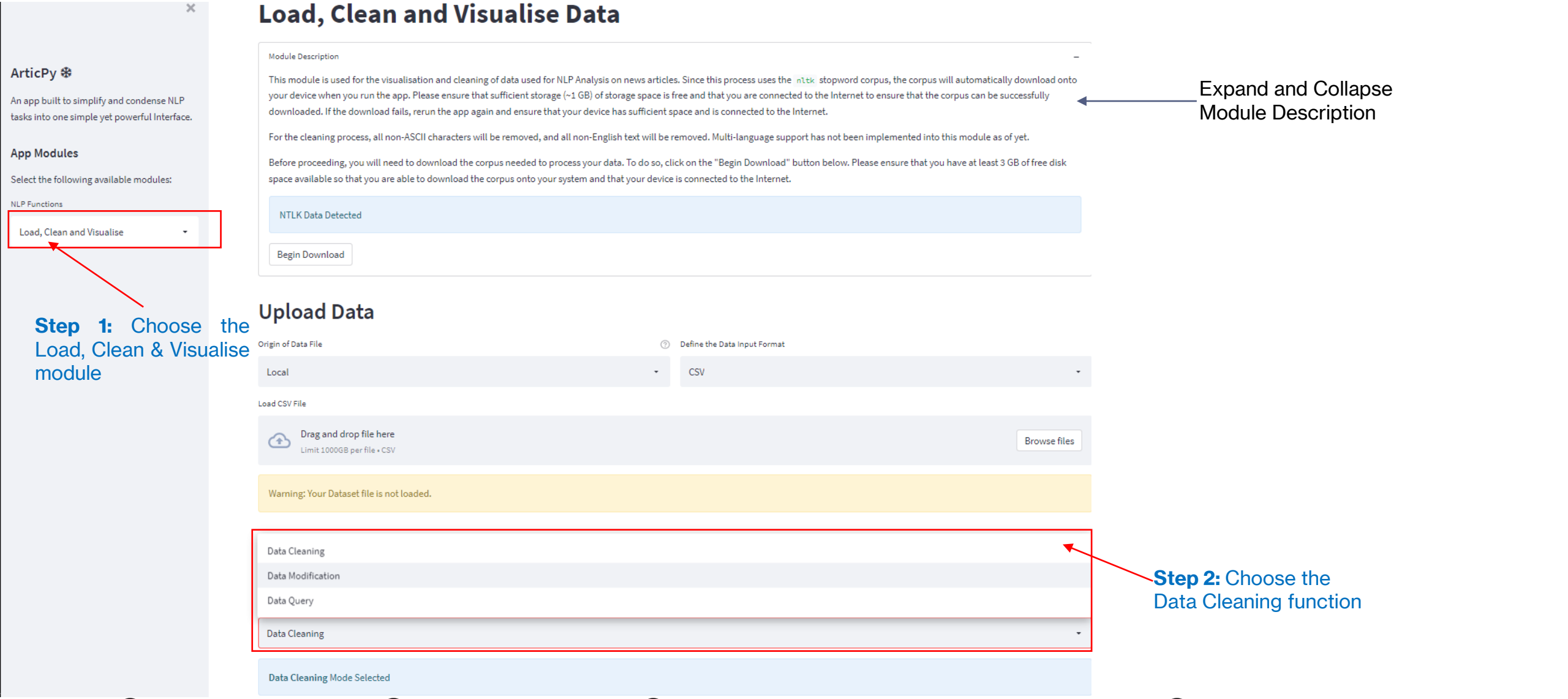
Problem Statement 1

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)**



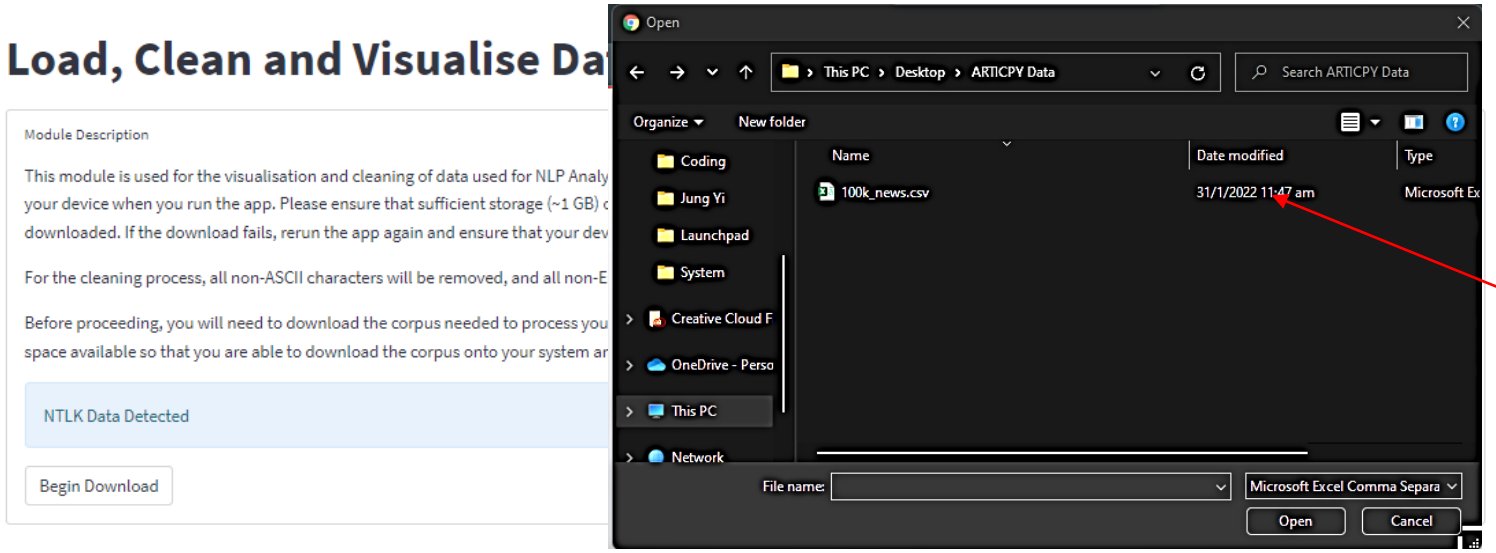
Expand and Collapse Module Description

Step 1: Choose the Load, Clean & Visualise module

Step 2: Choose the Data Cleaning function

Analysis - Problem Statement 1

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



Step 3: Hit the button browse file

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

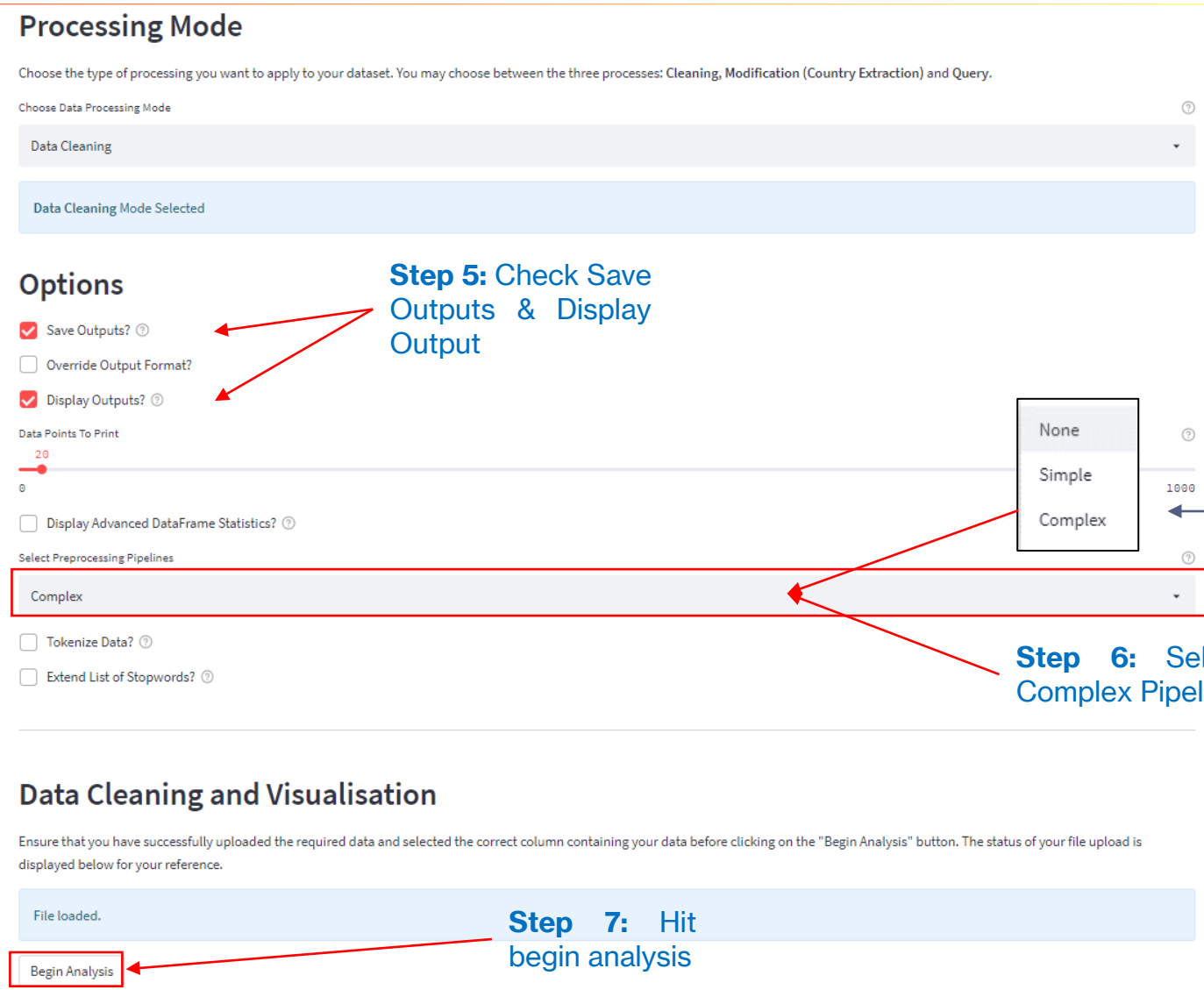
File Uploading

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

Azure, Amazon, Google

Analysis - Problem Statement 1

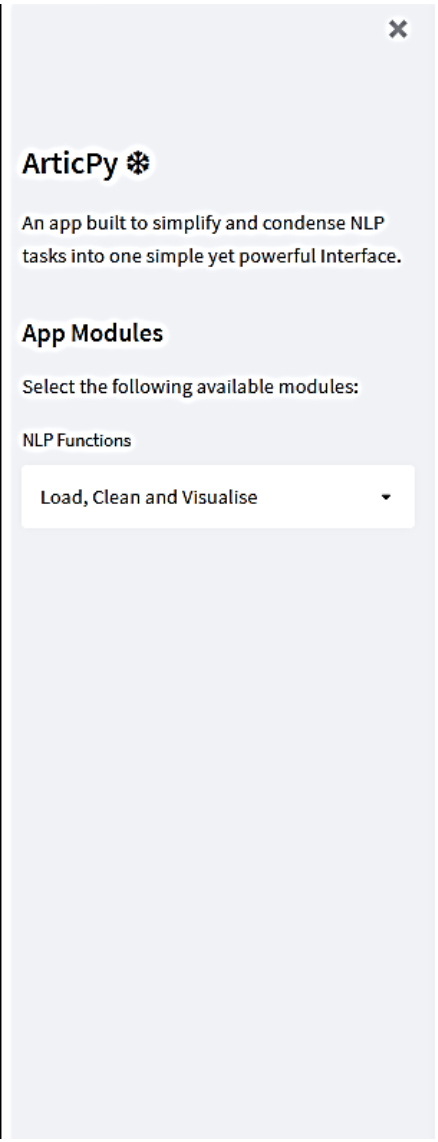
- 01 – Load, Clean & Visualise Module **(Data Cleaning)**



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	

Analysis - Problem Statement 1

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



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 unidentifie	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 box
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 launche	google stadia launched nover
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 op
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 - Perusaha	inilahcom seoul perusahaan t
17	Published: 17:17 BST, 8 Augus	published bst august updated
18	Isobel Asher Hamilton, Busin	isobel asher hamilton busine
19	Published: 16:43 BST, 8 Augus	published bst august updated

Download Data

Download Raw Data

Download Cleaned Data

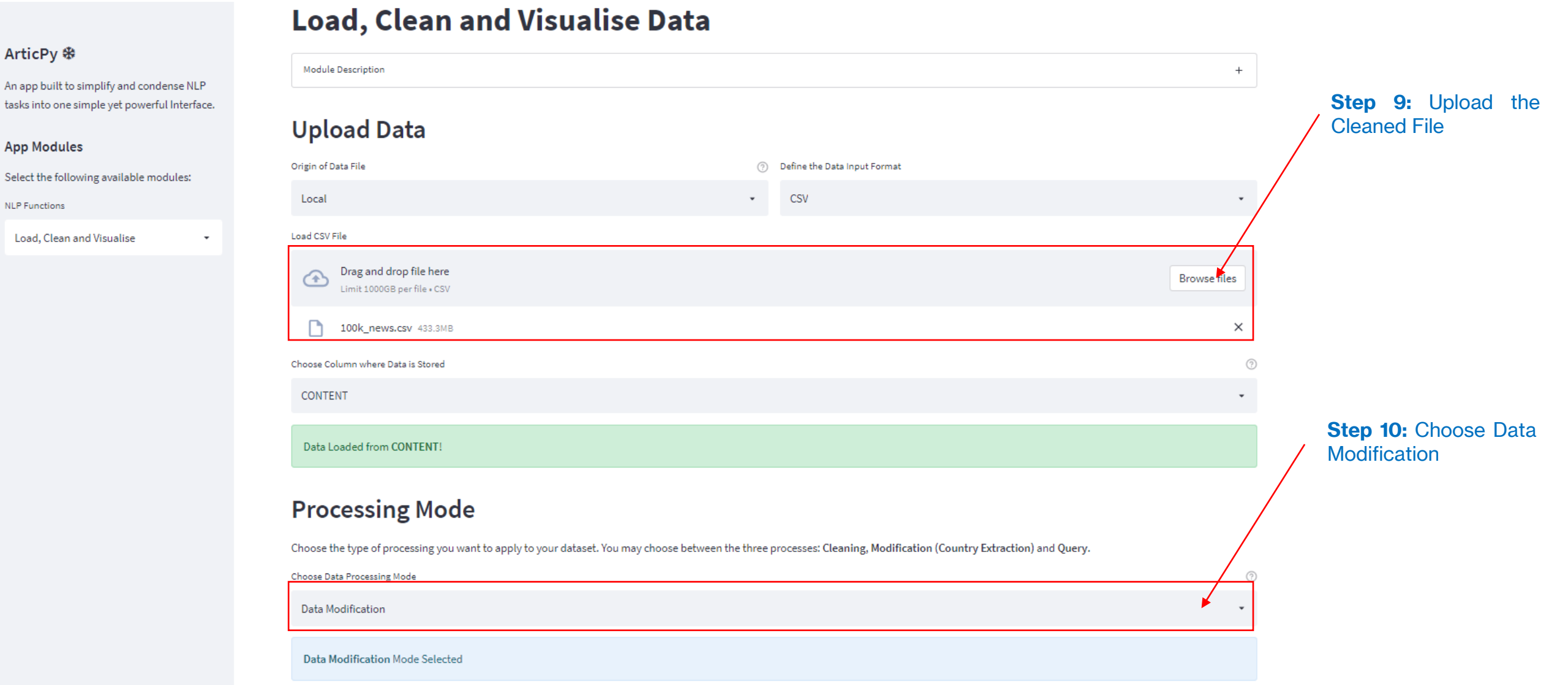
Step 8: Hit "Download Cleaned data"

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

Analysis - Problem Statement 1

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

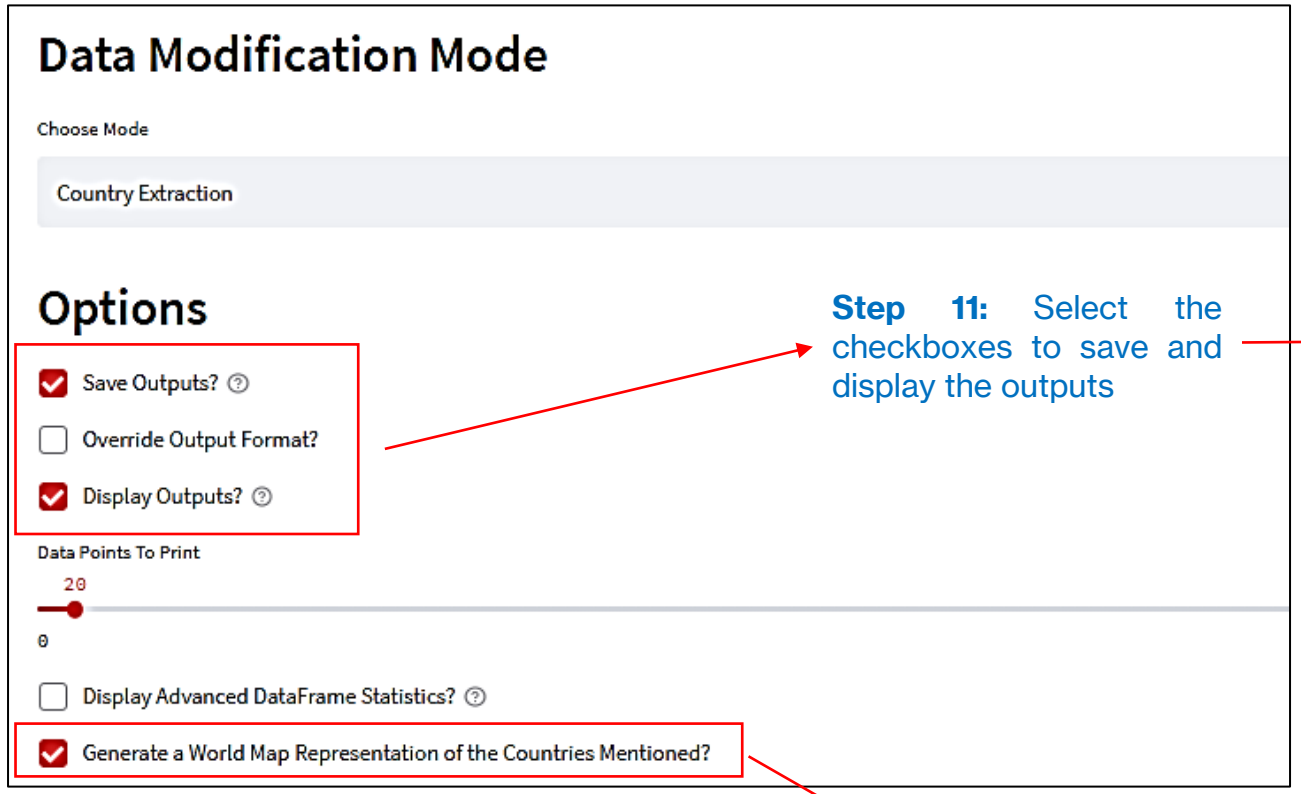


Step 9: Upload the Cleaned File

Step 10: Choose Data Modification

Analysis - Problem Statement 1

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



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

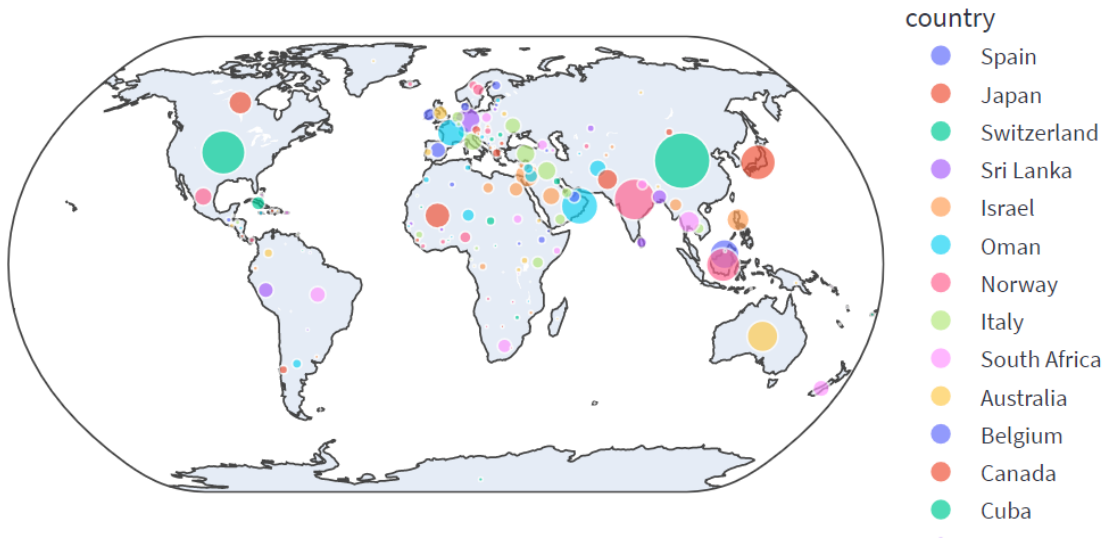
- 02 – Load, Clean & Visualise Module **(Data Modification)**

Libraries

pycountry
Georgapy3

pandas

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

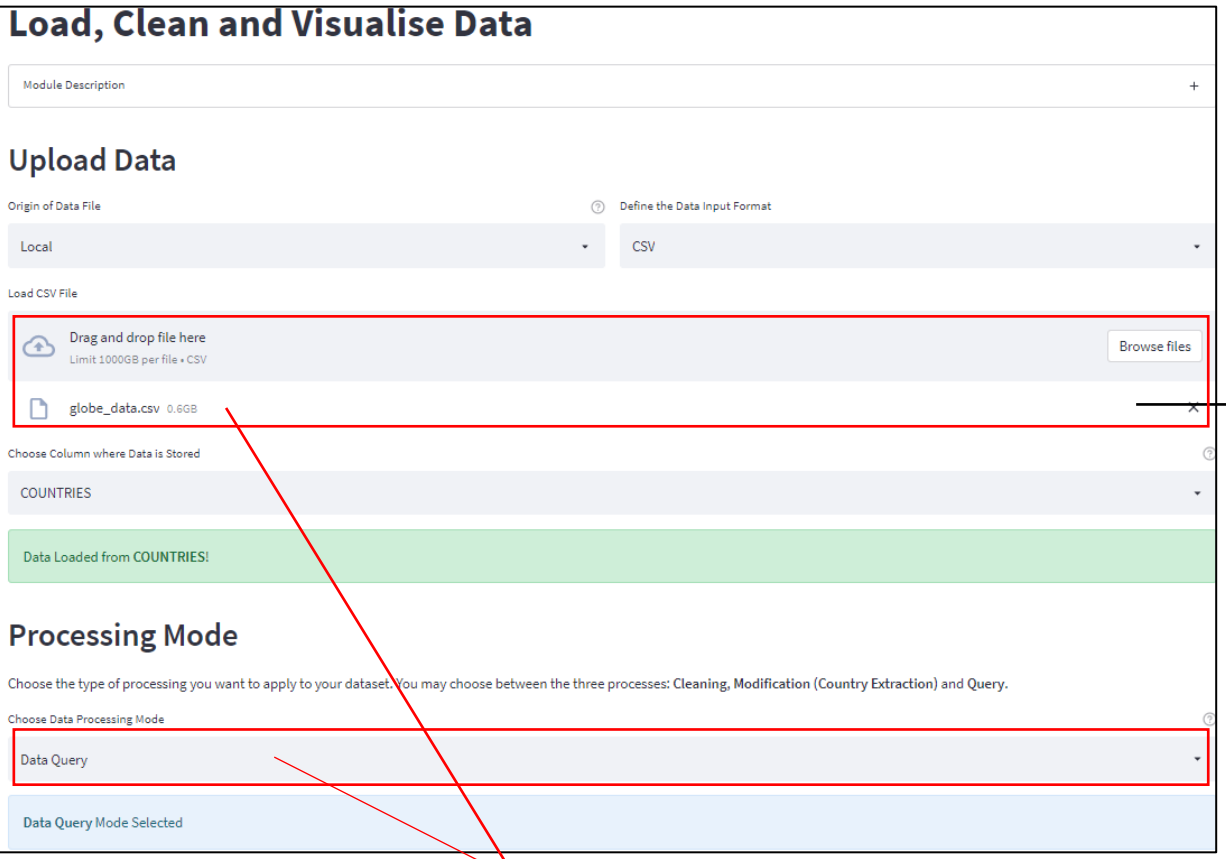
- Appends list of countries to the next column of data frame corresponding to the text

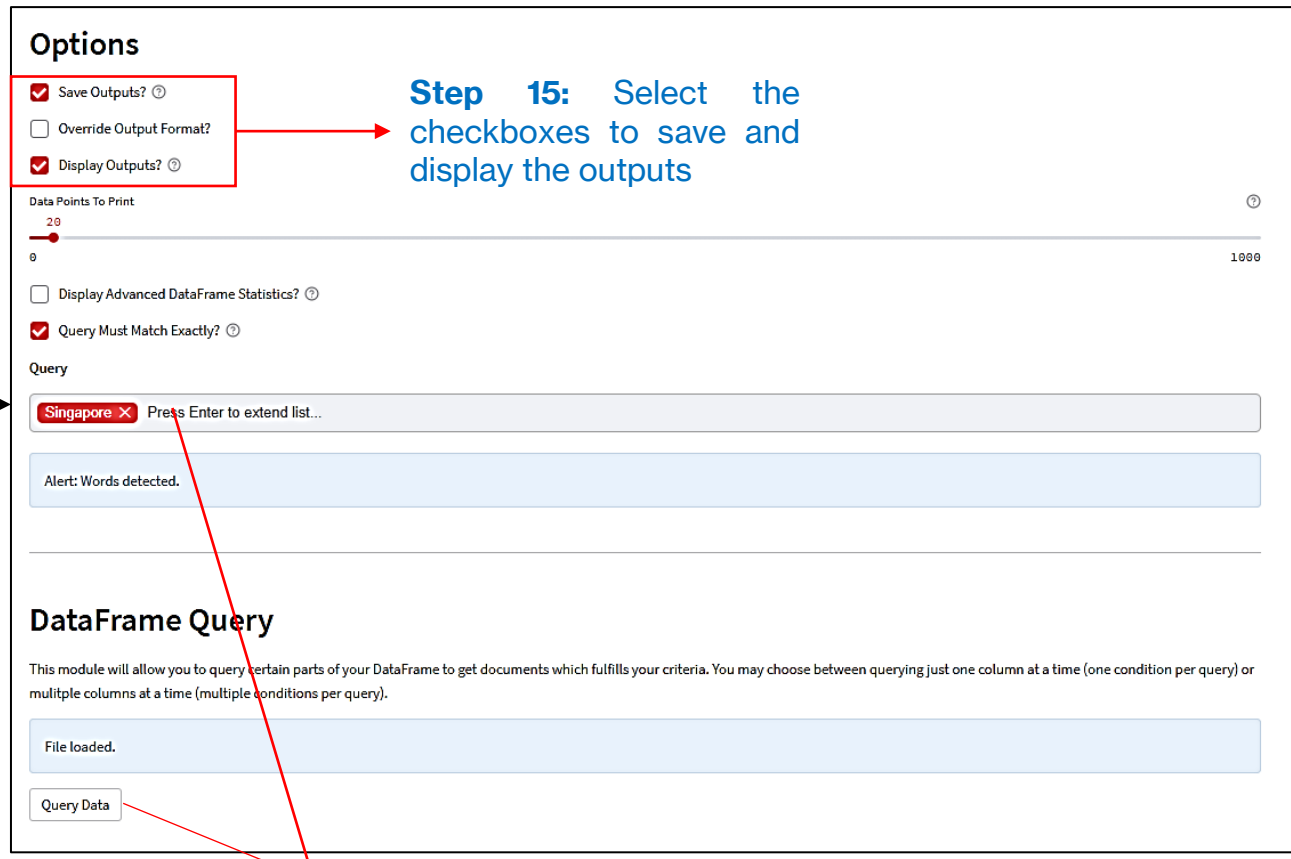
Purpose:

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

Analysis - Problem Statement 1

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





Step 15: Select the checkboxes to save and display the outputs

Step 13: Upload the tagged data set
Step 14: Ensure that the tagged data set is chosen for querying – "Countries)

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 trai	['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 count	spare country expense worth p mil	['Malaysia', 'Philippines', 'Singapor
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 mediu	['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 th	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 healt	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', 'M
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 australia	['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']

Save Query

Download Queried Data

Step 18: Download the queried focused data set

Libraries

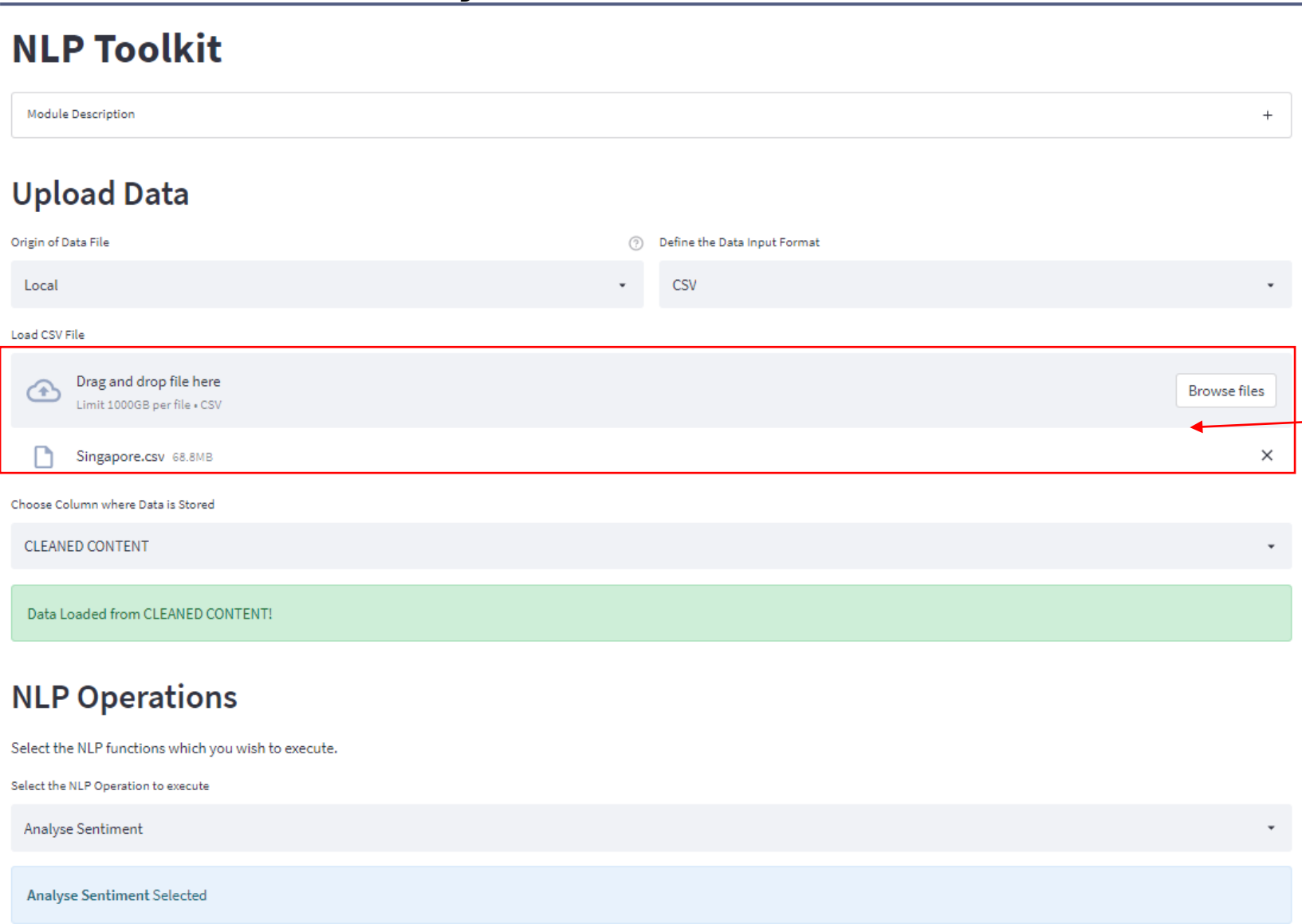
pandas

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

Analysis - Problem Statement 1

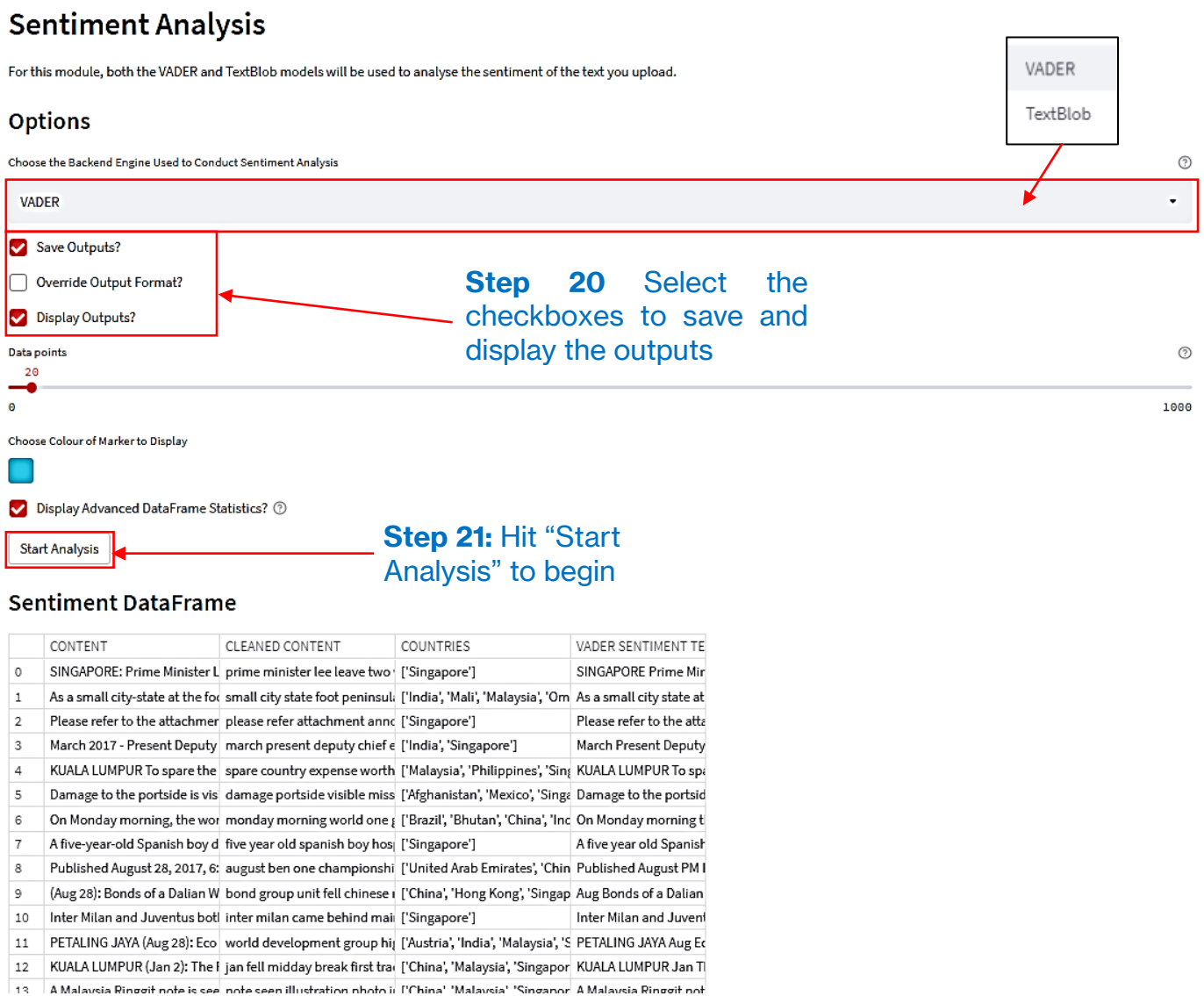
- 04 – NLP Toolkit **(Analyse Sentiment)**



Step 19: Upload the previously queried focused data set

Analysis - Problem Statement 1

- 04 – NLP Toolkit **(Analyse Sentiment)**



Libraries

spaCy, NLTK, plotly, TextBlob, kaleido

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)**

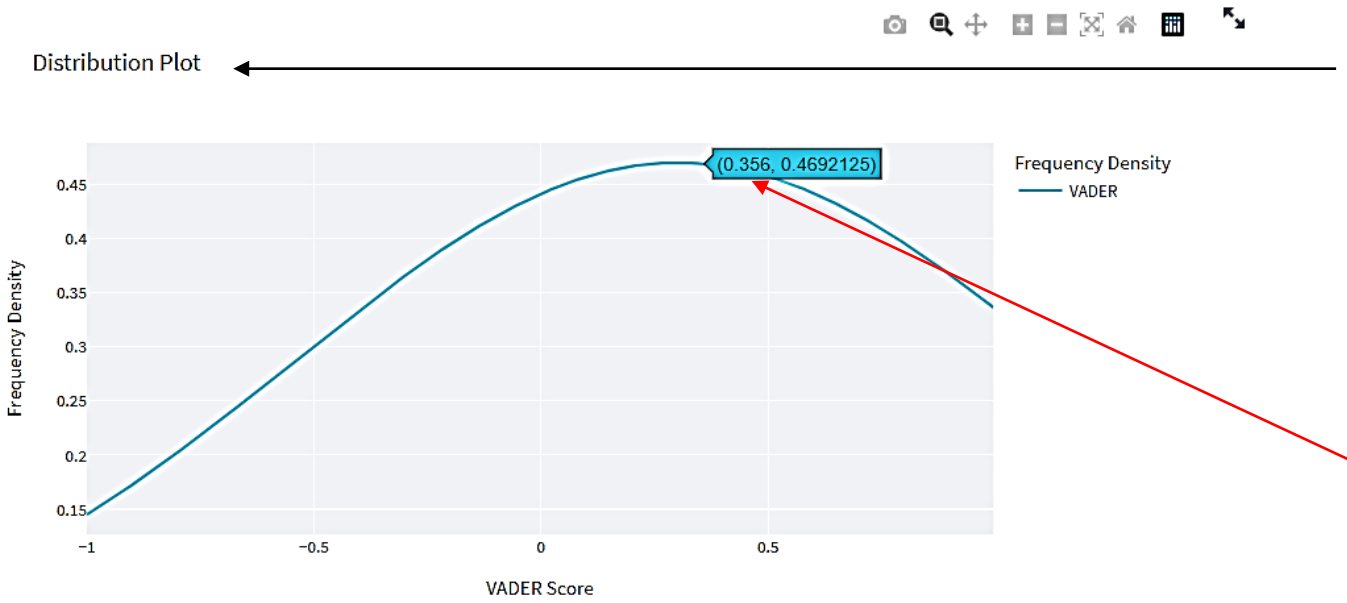
Sentiment Analysis – Data frame view

	CONTENT	CLEANED CONTENT	COUNTRIES	VADER SENTIMENT TEXT	VADER SENTIMENT	VADER SCORE	VADER OVERALL SCORE	VADER OVERALL SENTIMENT
0	SINGAPORE: Prime Minister Le	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', 'Omai	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 c	spare country expense worth p	['Malaysia', 'Philippines', 'Singa	KUALA LUMPUR To spare the c	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 worl	monday morning world one ge	['Brazil', 'Bhutan', 'China', 'India	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 FE	jan fell midday break first tradi	['China', 'Malaysia', 'Singapore'	KUALA LUMPUR Jan The FBM K	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 eva	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 I	liberal talk health care utopia s	['Canada', 'Germany', 'France', '	When liberals talk about their I	Positive	0.9998	0.9999	Positive
16	MANILA, July 31 (Xinhua) -- Chi	manila july china association s	['China', 'Indonesia', 'Cambodi	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 recr	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.

VADER Score:0.356

Download Data

Download Sentiment Score Data

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

Graphs

Download Plot

Analysis - Problem Statement 1

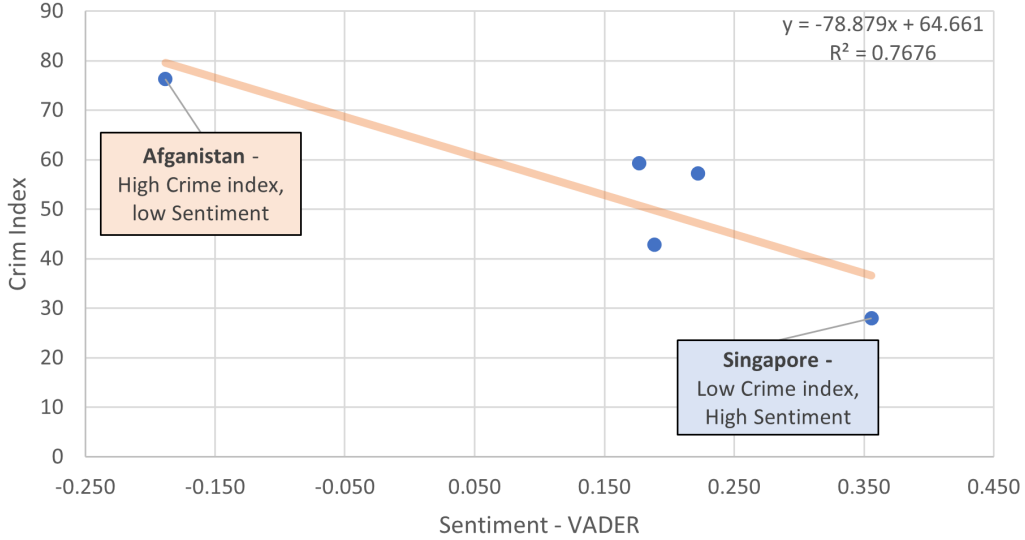
- Insights into Problem Statement 1

Problem Statement 1

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

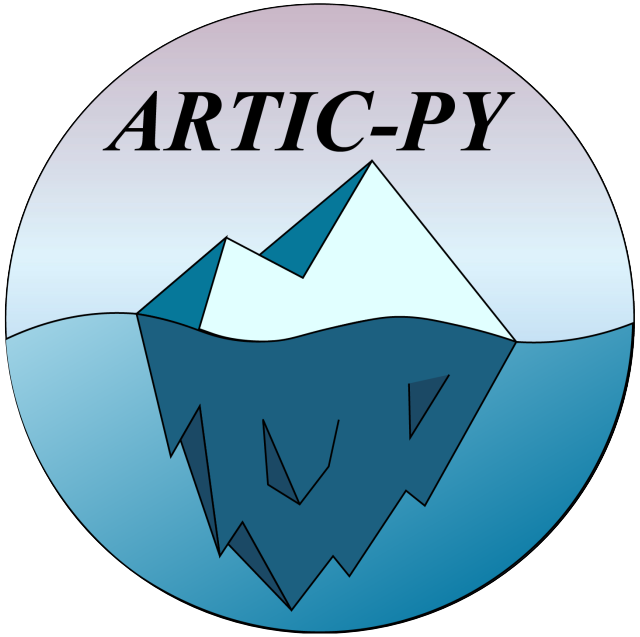
Insights

- 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

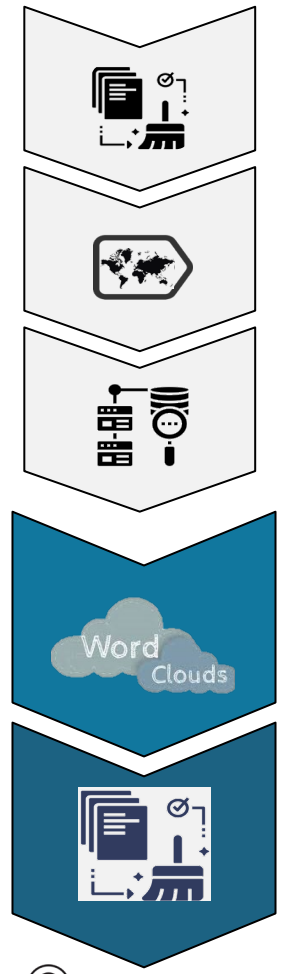
Analysis - Problem Statement 2

- Overview Procedure

Problem Statement 2

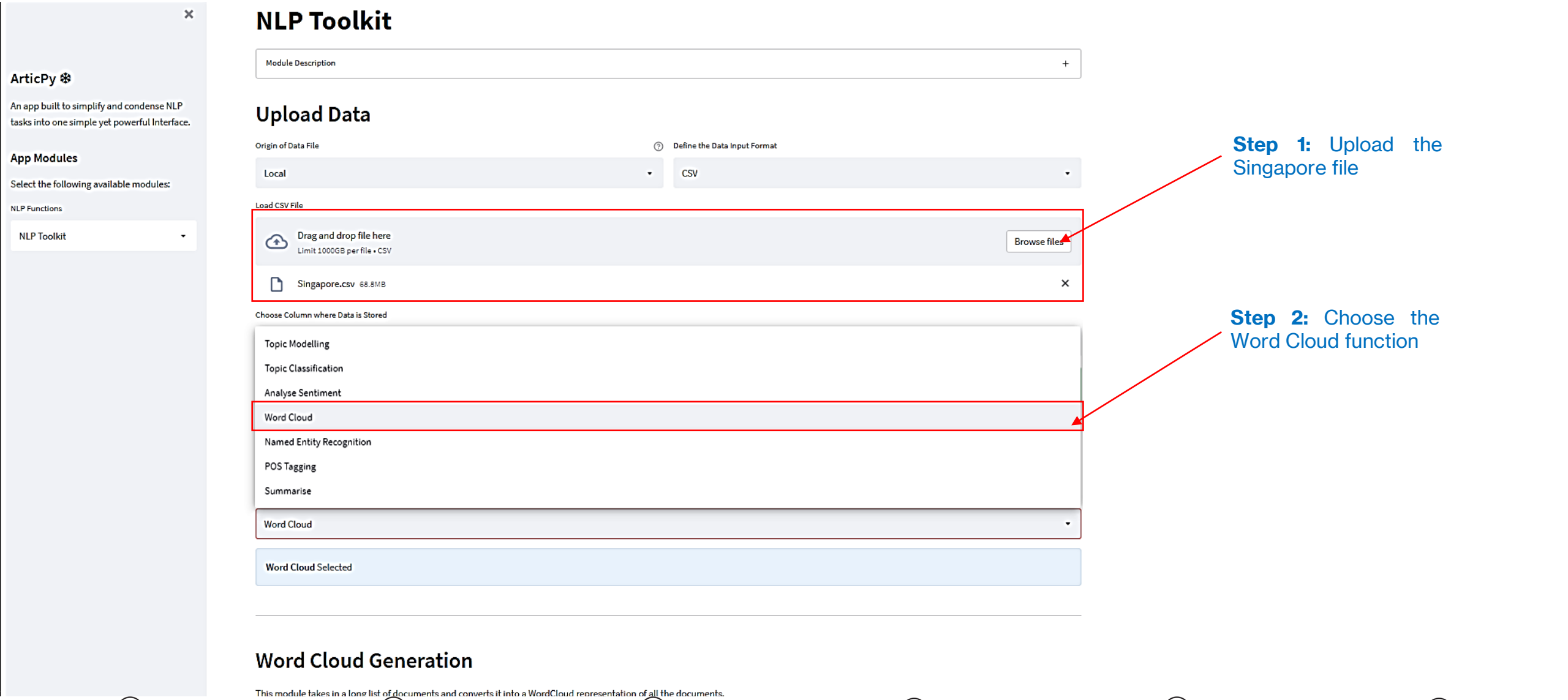
To analyze public sentiments from news articles to identify topics or issues within society.

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)**

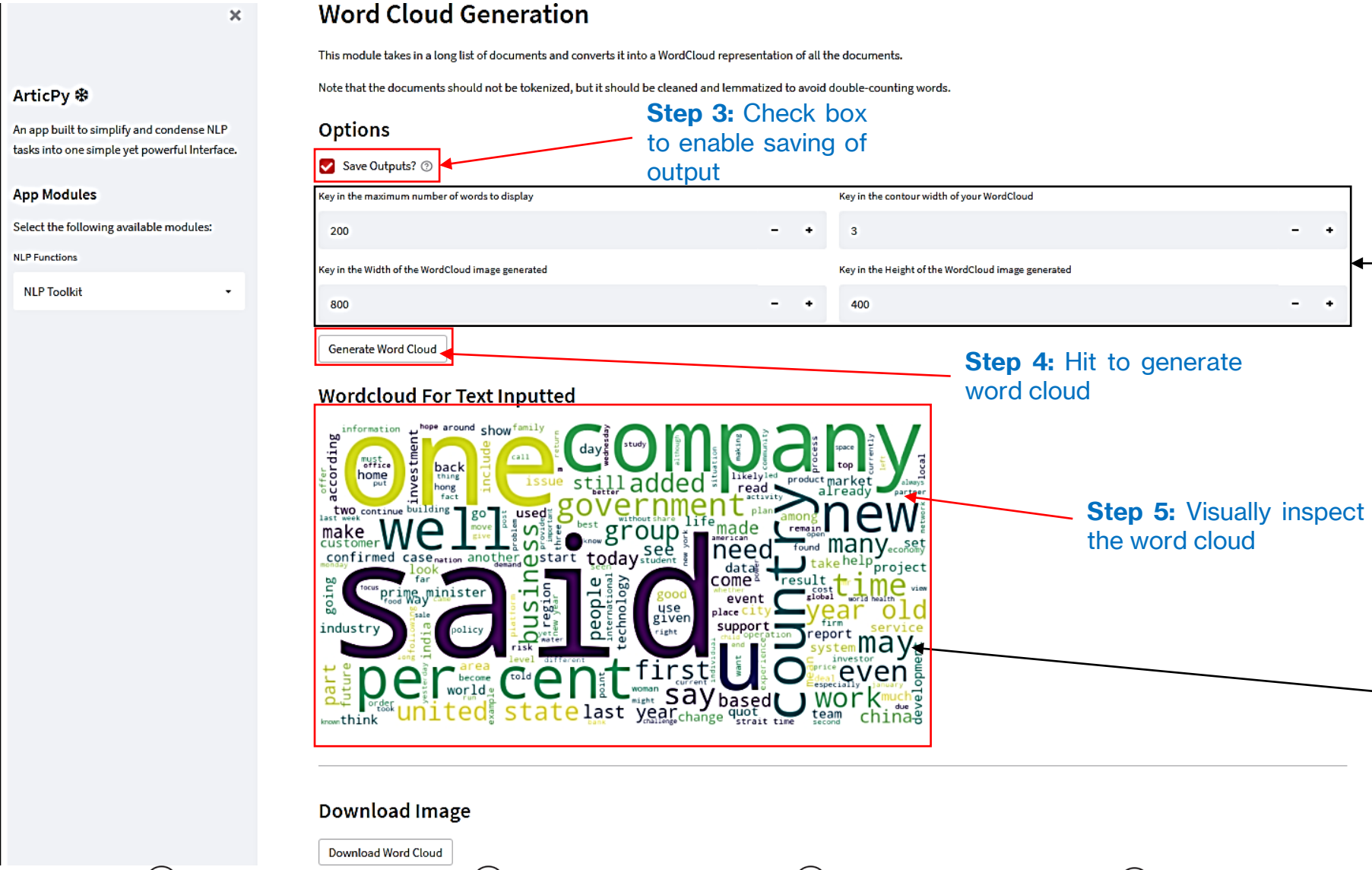


Step 1: Upload the Singapore file

Step 2: Choose the Word Cloud function

Analysis - Problem Statement 2

- 04 – NLP Toolkit **(Word Cloud)**



Step 3: Check box to enable saving of output

Step 4: Hit to generate word cloud

Step 5: Visually inspect the 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

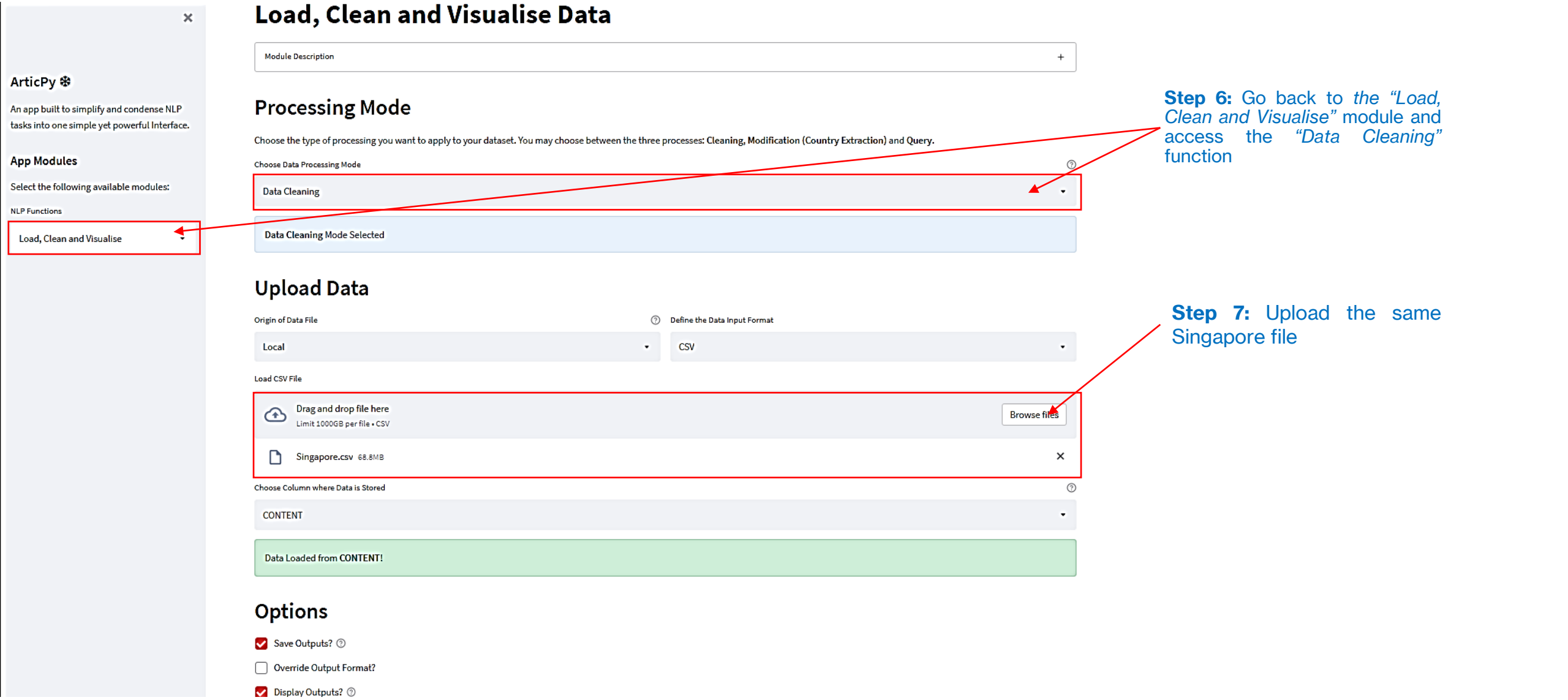
Insights:

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)**

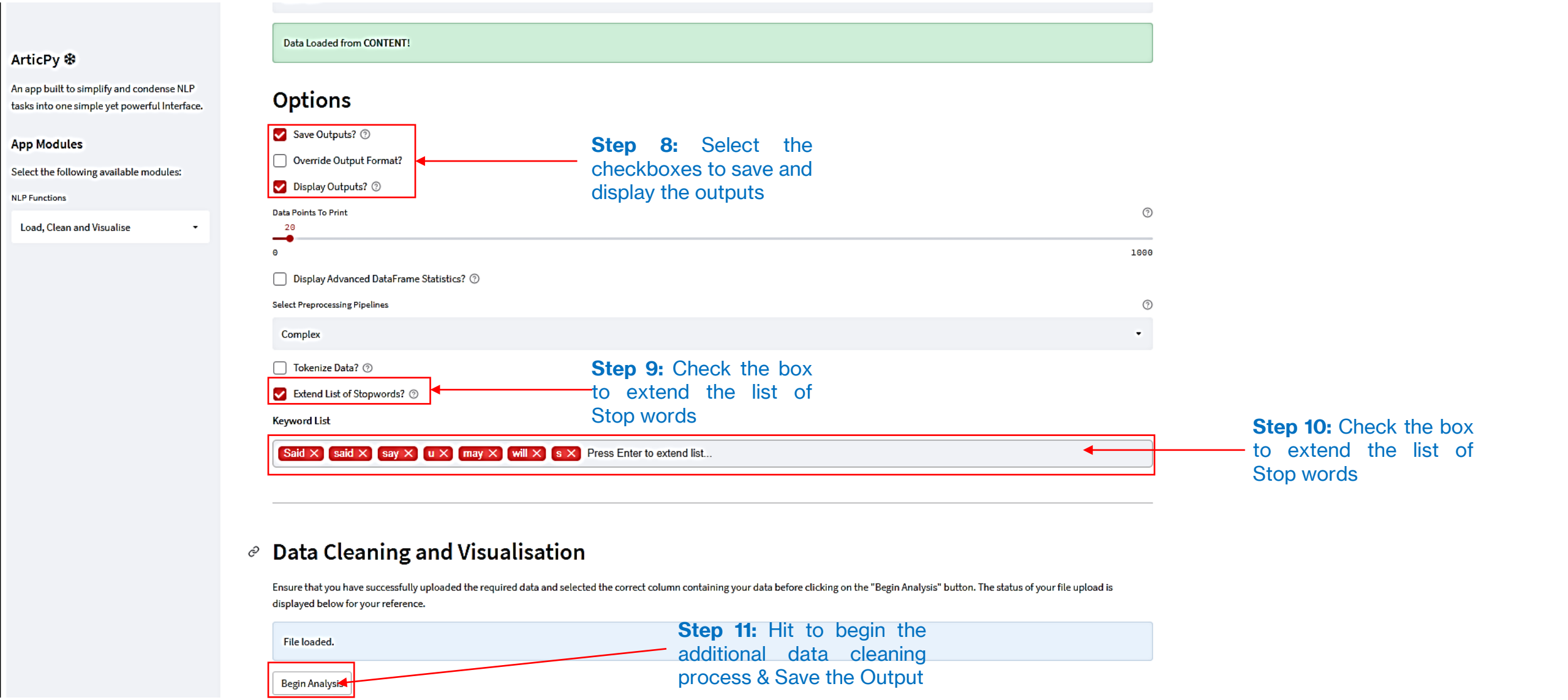


Step 6: Go back to *the "Load, Clean and Visualise"* module and access the *"Data Cleaning"* function

Step 7: Upload the same Singapore file

Analysis - Problem Statement 2

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



Analysis - Problem Statement 2

- 04 – NLP Toolkit **(Word Cloud)**

ARTIC-PY

NLP Toolkit

Module Description +

Upload Data

Origin of Data File

Local

Define the Data Input Format

CSV

Step 12: Upload the Cleaned Singapore dataset

Load CSV File

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

Browse files

Singapore_stopword.csv 86.0MB

Choose Column where Data is Stored

CLEANED CONTENT

Data Loaded from CLEANED CONTENT!

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

NLP Operations

Select the NLP functions which you wish to execute.

Select the NLP Operation to execute

Word Cloud

Step 14: Output and save the data

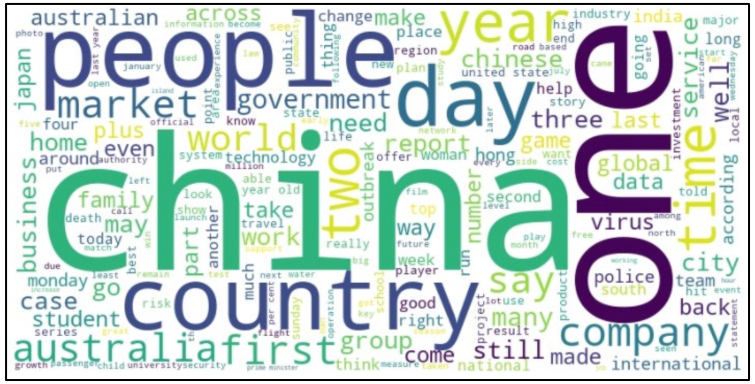
Word Cloud Selected

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

Analysis & Summary

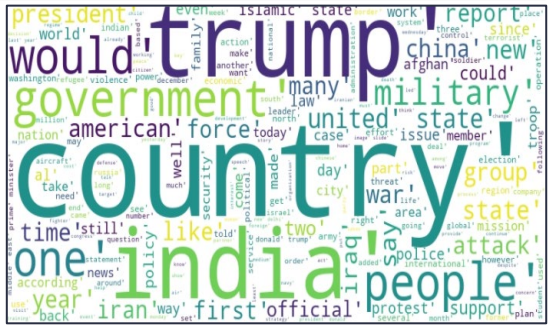
- Insights into Problem Statement 2 & Summary of Problem Statements 1 & 2

Problem Statement 2

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

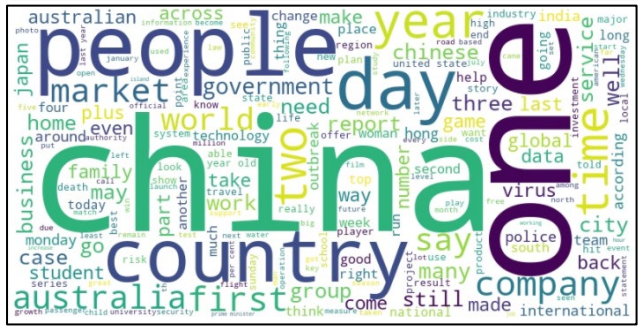
Afghanistan

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,



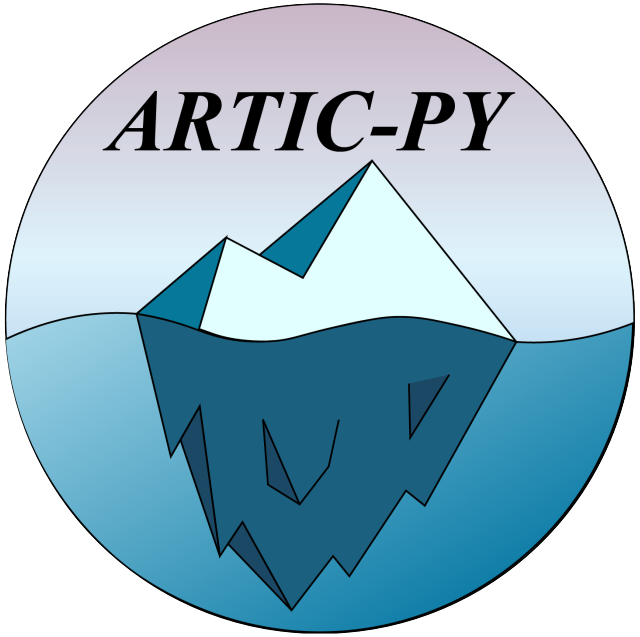
Singapore

- Higher sentiment from a united community *"one, people, country"*, as Singapore emerge stronger from the pandemic
- Worried about the economic and political relationship between China and US.
- Worried about the virus and cases



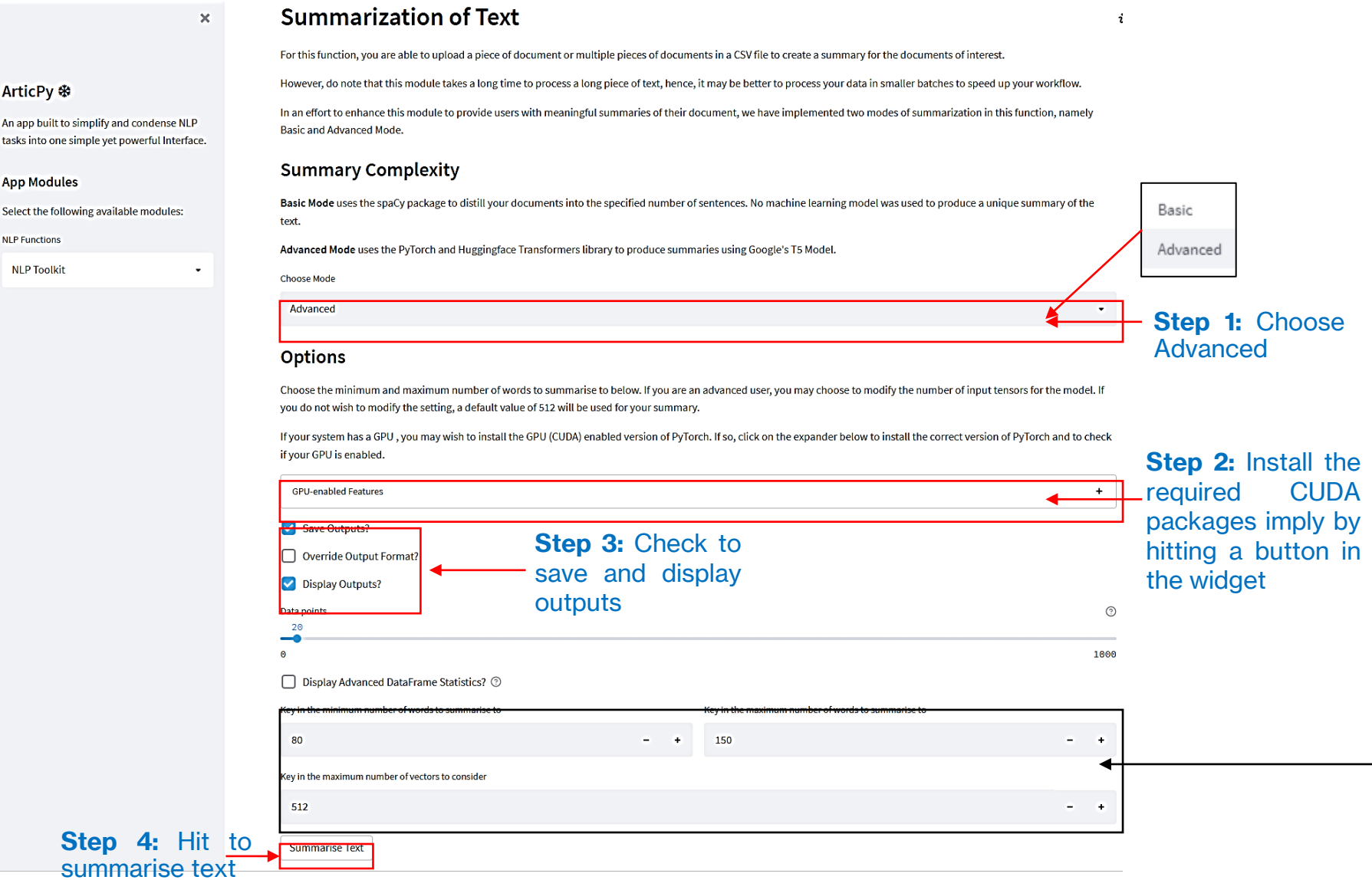
Conclusion

- 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.



Other Features

Other Features – Text Summarizer



Libraries

Text Summariser

spaCy

PyTorch

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

	ï»¿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 gunr	moscow unidentified gunman fire t	the FSB says the assailant was actir
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 giar	threat escalates an ongoing row be
5	Photo Credit: Yossi Zamir/Flash90 T	photo credit flash ministerial comn	the ministerial committee on privat
6	Everybody likes a little treat now ar	everybody like little treat recently l	not all kids love sweets, but it's a g
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 com	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 licensing s	want keep simple sure appeal base	the latest version of the software su
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 doors
16	INILAHCOM, Seoul - Perusahaan tel	electronics dan chip artificial intell	Baidu KUNLUN adalah proyek yang
17	Published: 17:17 BST, 8 August 2018	august august crown prince bin ove	foreign minister: "Canada made a b
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

Download Summarised Data

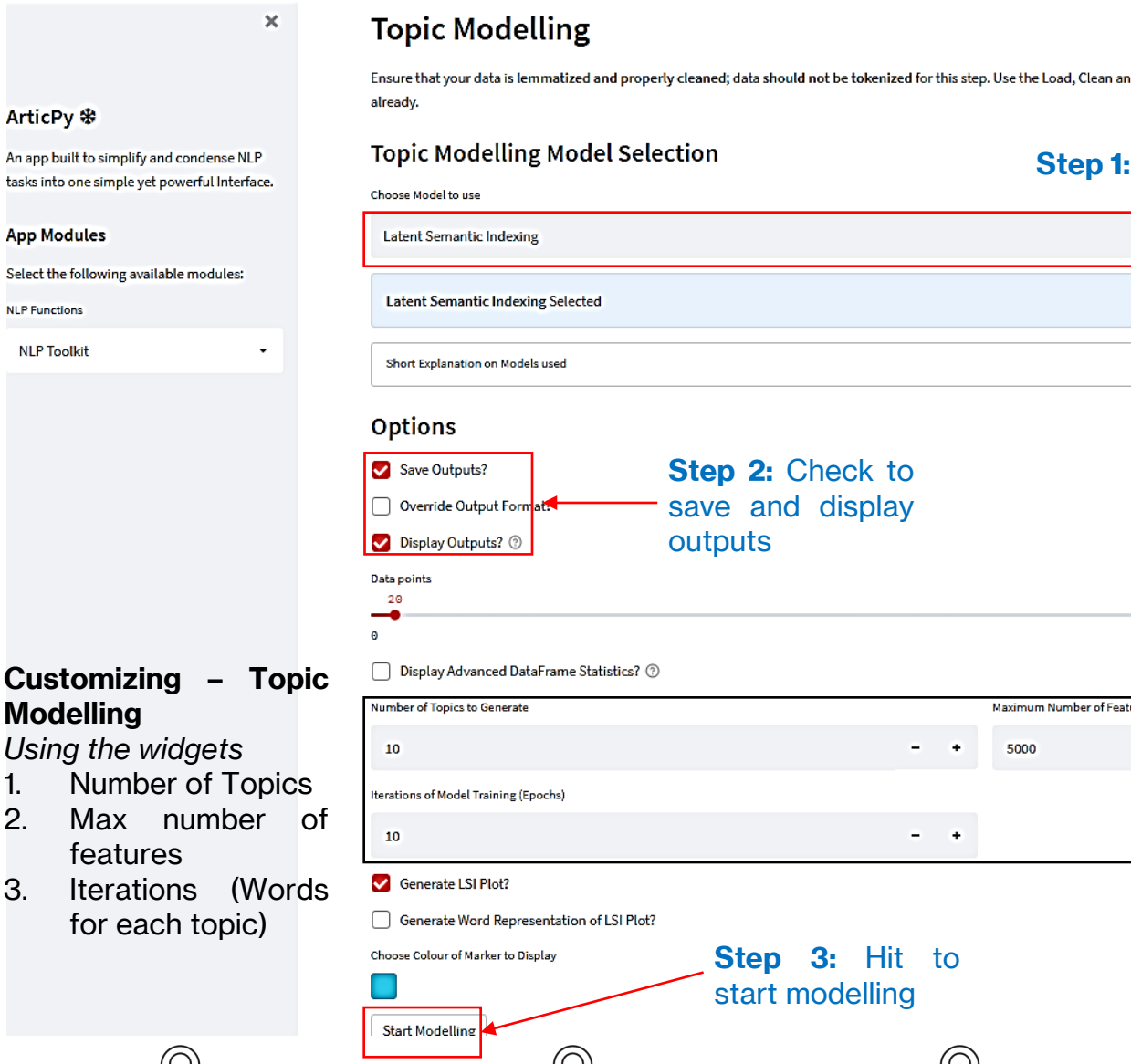
Download Summarised Data

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

Other Features – Topic Modelling



Customizing – Topic Modelling

Using the widgets

1. Number of Topics
2. Max number of features
3. Iterations (Words for each topic)

Libraries

scikit learn, pandas, NLTK

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

	word	weight
0	china	0.3897
1	said	0.3398
2	case	0.2262
3	year	0.2065
4	new	0.1869
5	virus	0.1741
6	country	0.1632
7	people	0.1588
8	chinese	0.1426
9	health	0.1278

Topic 2

	word	weight
0	market	0.3164
1	year	0.2758
2	company	0.1780
3	million	0.1614
4	business	0.1266
5	growth	0.1021
6	global	0.0911
7	data	0.0904
8	time	0.0875
9	investment	0.0852

Topic 5

	word	weight
0	case	0.4103
1	north	0.1831
2	virus	0.1801
3	confirmed	0.1612
4	trump	0.1468
5	answer	0.1289
6	read	0.1270
7	state	0.1212
8	kim	0.1046
9	patient	0.1006

Topic 6

	word	weight
0	million	0.2672
1	target	0.2489
2	met	0.2392
3	china	0.2253
4	year	0.1442
5	quarter	0.1325
6	net	0.1071
7	company	0.0873
8	financial	0.0826
9	revenue	0.0764

Topic 3

	word	weight
0	market	0.4801
1	china	0.1920
2	case	0.1590
3	virus	0.1472
4	wear	0.1419
5	global	0.1047
6	million	0.0996
7	confirmed	0.0924
8	analysis	0.0910
9	growth	0.0814

Topic 4

	word	weight
0	china	0.4675
1	north	0.2528
2	state	0.2118
3	market	0.1844
4	trump	0.1664
5	united	0.1615
6	answer	0.1309
7	president	0.1192
8	nuclear	0.1186
9	kim	0.1145

Topic 7

	word	weight
0	target	0.6658
1	met	0.6493
2	said	0.1593
3	market	0.1448
4	wear	0.0856
5	woman	0.0434
6	analysis	0.0404
7	country	0.0397
8	table	0.0376
9	child	0.0358

Topic 8

	word	weight
0	year	0.3712
1	wear	0.3326
2	million	0.2939
3	child	0.1319
4	sale	0.1306
5	north	0.1281
6	corresponding	0.1229
7	table	0.1228
8	chart	0.1181
9	graph	0.1164

Topic 9

	word	weight
0	wear	0.2992
1	university	0.1662
2	people	0.1630
3	china	0.1626
4	child	0.1509
5	year	0.1440
6	time	0.1411
7	world	0.1395
8	quot	0.1237
9	like	0.1120

Topic 10

	word	weight
0	macaw	0.3762
1	actor	0.3264
2	scott	0.2620
3	virus	0.2103
4	people	0.1607
5	know	0.1383
6	market	0.1306
7	yeah	0.1044
8	tax	0.1040
9	like	0.1025

LSI Topic DataFrame

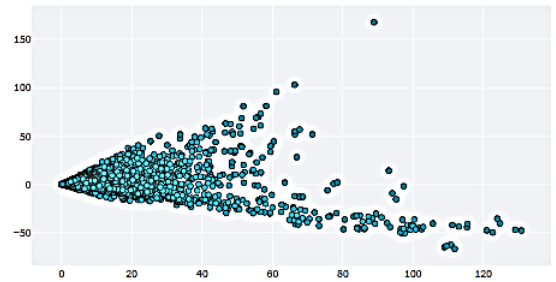
	word_0	word_1	word_2	word_3	word_4	word_5	word_6	word_7
topic_0	china	said	case	year	new	virus	country	people
topic_1	market	year	company	million	business	growth	global	data
topic_2	market	china	case	virus	wear	global	million	confirm
topic_3	china	north	state	market	trump	united	answer	preside
topic_4	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

Other Features – NER & POS

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

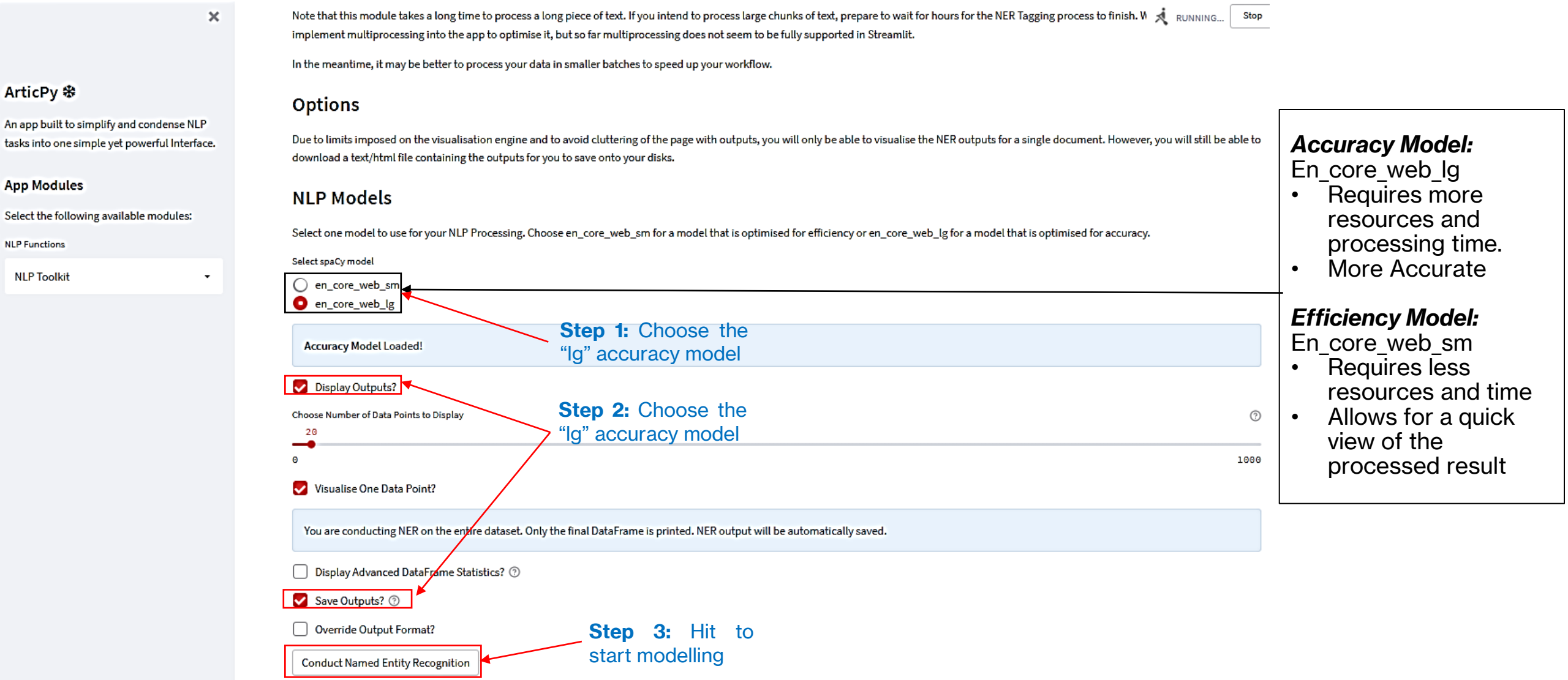
POS Tagging: Tagging words corresponding with a particular part of speech depending on the definition of word and its context *(E.g. noun, verb)*

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)



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',
```

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. "

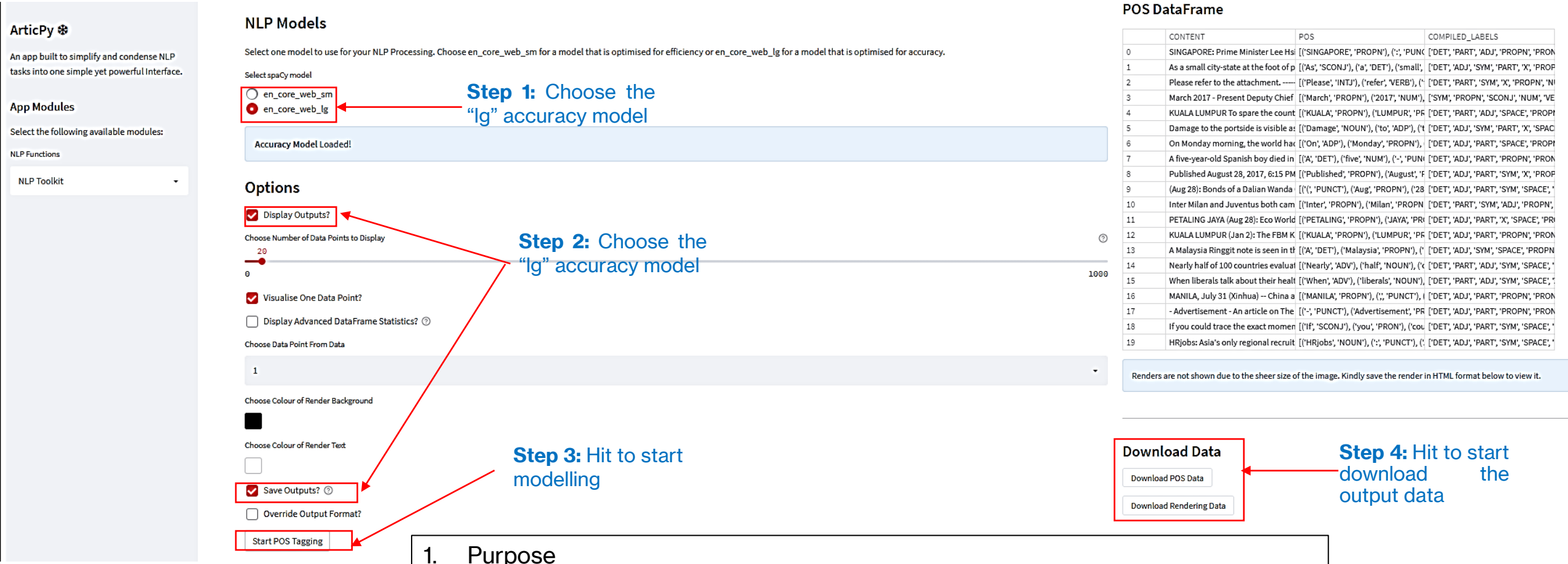
Also plan to catch up on my other readings, spend time with family, and perhaps go #jalanjalan (if the rain stays away)," he added.

NER DataFrame

	CONTENT	NER	COMPILED_LABELS
0	SINGAPORE: Prime Minister Lee Hsi	[('SINGAPORE', 'GPE'), ('Lee Hsien L	['DATE', 'PERSON', 'GPE', 'LOC', 'ORG
1	As a small city-state at the foot of p	[('Malaysia', 'GPE'), ('Singapore', 'GI	['EVENT', 'DATE', 'PERSON', 'MONEY
2	Please refer to the attachment. -----	[('Company', 'ORG'), ('Company', 'C	['PERSON', 'DATE', 'FAC', '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	['EVENT', 'DATE', 'PERSON', 'ORDINA
5	Damage to the portside is visible as	[('Guided', 'ORG'), ('USS John S. Mc	['EVENT', 'PRODUCT', 'DATE', 'PERSO
6	On Monday morning, the world hac	[('Monday', 'DATE'), ('morning', 'TIM	['DATE', 'PERSON', 'TIME', 'CARDINA
7	A five-year-old Spanish boy died in	[('five-year-old', 'DATE'), ('Spanish',	['DATE', 'PERSON', 'TIME', 'GPE', 'NO
8	Published August 28, 2017, 6:15 PM	[('August 28, 2017', 'DATE'), ('6:15 P	['EVENT', 'PRODUCT', 'PERSON', 'DA
9	(Aug 28): Bonds of a Dalian Wanda	[('Aug 28', 'DATE'), ('Dalian Wanda C	['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	['EVENT', 'DATE', 'PERSON', 'PERCEN
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 evaluat	[('Nearly half', 'CARDINAL'), ('100', '	['DATE', 'PERSON', 'ORDINAL', 'MON
15	When liberals talk about their healt	[('Frances', 'PERSON'), ('1', 'CARDIN	['PRODUCT', 'PERSON', 'PERCENT', '
16	MANILA, July 31 (Xinhua) -- China a	[('MANILA', 'ORG'), ('July 31', 'DATE'	['DATE', 'PERSON', 'CARDINAL', 'LAW
17	- Advertisement - An article on The	[('Singapores', 'ORG'), ('Singapores	['DATE', 'PERSON', 'ORDINAL', 'WOR
18	If you could trace the exact momen	[('Australia', 'GPE'), ('China', 'GPE'),	['EVENT', 'PERSON', 'DATE', 'LANGU
19	HRjobs: Asia's only regional recruit	[('Asia', 'LOC'), ('100%', 'PERCENT'),	['DATE', 'PERCENT', 'MONEY', 'PERS

Other Features – NER & POS

- Position of Speech (POS) - Output



Step 1: Choose the "lg" accuracy model

Step 2: Choose the "lg" accuracy model

Step 3: Hit to start modelling

Step 4: Hit to start download the output data

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

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