

README

Personal information

Jianping DENG | jianpind@andrew.cmu.edu
Shan GAO | shangao@andrew.cmu.edu
Tan Tan (Zoe) KWAN | tantank@andrew.cmu.edu
Takuya FUNAHASHI | tfunahas@andrew.cmu.edu
Yunzhou NING | yunzhoun@andrew.cmu.edu

Introduction

This project was finished as the team project for the course Data-Focused Python at CMU. Basically we want to provide a news integrator with information from many channels such as BBC, ABC, Renmin etc. Moreover, we conducted analysis in many dimensions. In this README file, we aim to explain how to set up our program, including data source, data cleaning, data analysis and interface. For a quicker overview of this project, **you can see the 5-mins video to know what we have done: <https://youtu.be/t1Co5TQH79Y>**

Step 0: Setting up and some background

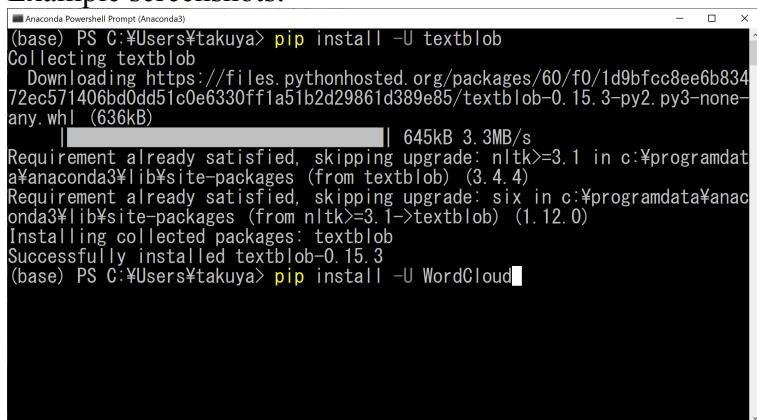
Please download all the files in this repository, unzip all the folders at where they are, make sure they are in the directory.



Install the following packages to set up the environment for functions within this program, such as emotional analysis, word cloud and LDA model.

```
$ pip install -U textblob
$ pip install -U WordCloud
$ pip install -U lda
```

Example screenshots:



A screenshot of an Anaconda PowerShell Prompt window. The command \$ pip install -U textblob is run, followed by \$ pip install -U WordCloud and \$ pip install -U lda. The output shows the download and installation process for each package, indicating that requirements were already satisfied or skipping upgrades for nltk and six. The prompt then changes to (base) PS C:\Users\takuya> pip install -U WordCloud.

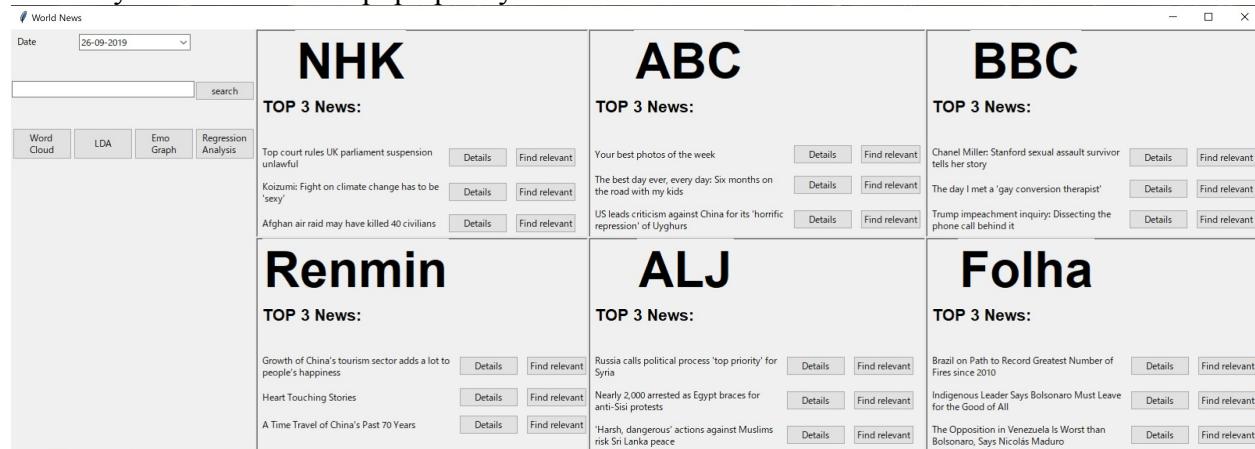
Step 1: Launching User Interface

Please keep in mind to include all files in the zip file in the same directory.

To run the graphic user interface, you need to launch the main file world_news.py file in Anaconda prompt:

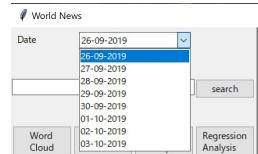
```
$ python world_news.py
```

You may see the **interface** pop up on your screen now:

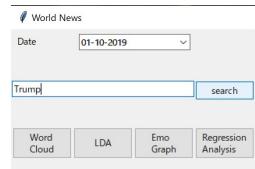


You can try different features out.

Feature 1: Date Selection to check the news content for each day (auto-refresh)



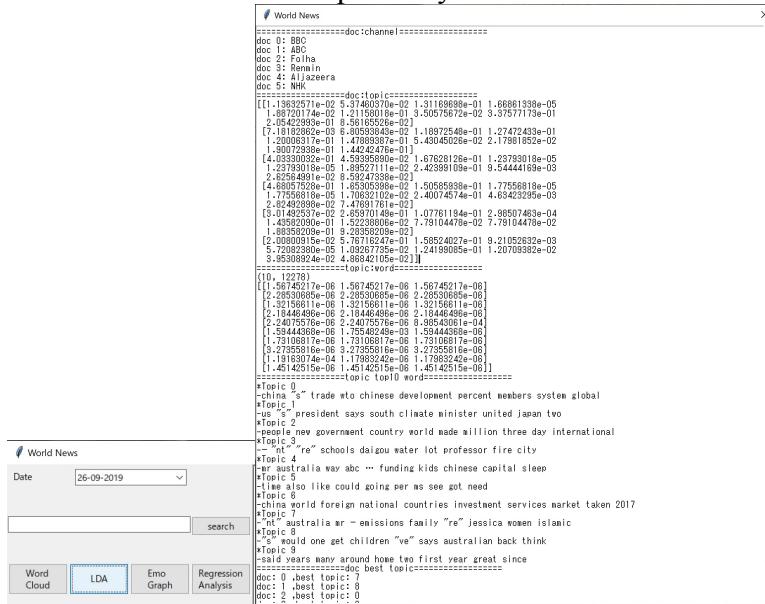
Feature 2: Search Bar to type in keywords and find relevant contents



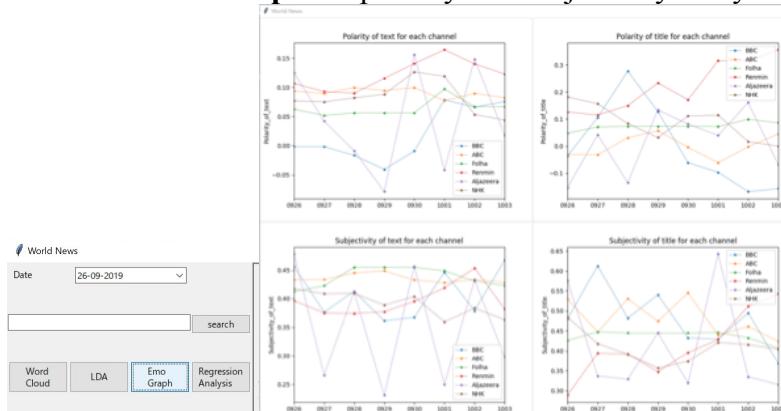
Feature 3: Word Cloud to check the text intensity of overall news content for all news sources (refreshed each day)



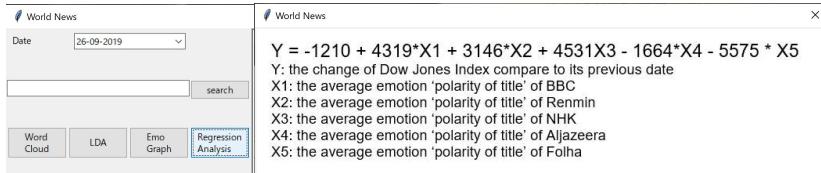
Feature 4: LDA for the topic analysis



Feature 5: Emo Graph for polarity and subjectivity analysis



Feature 6: Emotion Analysis for the detailed formula



Feature 7: See news details by pressing the **Details** button and **Find Relevant** news

The screenshot shows the 'ALJ' news section. It displays 'TOP 3 News:' with three items: 'Challenging debut for midnight marathon in Doha heat, humidity' (with 'Details' and 'Find relevant' buttons), 'Free Hong Kong' chants on 5 years of Umbrella Movement' (with 'Details' and 'Find relevant' buttons), and 'Ben Ali may be dead, but his regime is not' (with 'Details' and 'Find relevant' buttons). Below this, a detailed view of an article titled 'Top court rules UK parliament suspension unlawful' is shown. The details include: Link (https://www3.nhk.or.jp/nhkworld/en/news/20190924_53/), Date (20190926), Summary (Britain's top court says Prime Minister Boris Johnson's recent suspension of Parliament is unlawful and void. This is a major blow to the leader as he prepares for the country's planned exit from the European Union).

For the details, you can refer to the video: <https://youtu.be/t1Co5TQH79Y>

Zip file content

In the zip file, we have the following files. The checklist is below, and the details on how to use them is covered after the table, in sequence.

item	Document	Nature
1	world_news.py	The program of interface, also our main file
2	concat.xlsx	The merged data in excel file. Worked as the static database for our program
3	readData_vfinal.py	.py file
4	readExcel.py	.py file

5	linear_regression.py	.py file. Formula involved.
6	lda	A folder contains all the lda report in txt format
7	generate_word_cloud.py	.py file
8	WX20191007-131356@2x.png	png file to generate word cloud
9	find_relevant.py	.py file
10	handle_important_news.py	.py file
11	img	A folder containing all the pictures we need in the interface
12	important	A folder containing the result of handle_important_news.py, one of the data sources
13	analyze_channels.py	.py file
14	analyze_date.py	.py file
15	dowjones.csv	A csv file, one of the data sources
16	web scraping scripts	A folder. Each of us has done some web scraping from different news sources. The scripts are included in this folder.
17	news_eachsource.py	.py file
18	ReadMe.pdf	The instruction to use our product
19	topic_analysis.py	.py file to do conduct LDA model
20	PythonProject_DataStorage	A folder contains all data from the web scraping
21	YouTubeVideo	Web Link to Introduction YouTube Video
22	Python project final presentation.pptx	pptx file for the presentation

Other information: Web Scraping and Data Storage (You should follow this instruction to run our code)

a) Data Source

Details of our data source are shown in the following picture:

Source	Technology	Source from region
NHK	Webscraping	Asia(Japan)
Aljazeera	Webscraping	Middle East
Renmin	Webscraping	Asia(China)
ABC	Webscraping	Australia
BBC	Webscraping	Europe
Folha de S.Paulo	Webscraping	South America(Brazil)
Important News	Downloading	No specific region
*Dow Jones Industrial Average	Downloading	No specific region

b) Web Scraping and Data Storage

Web Scraping

The web scraping is conducted on a daily basis (not in real time) to acquire the news for each day from 26th Sep 2019, to 4th Oct 2019. During this process, two sources has not been scrapped every day, so the data for African News and the Wall Street Journal are not used in our analysis. Please refer to the folder “web scraping scripts” for all the scripts.

For example, open the folder web scraping scripts, and run the ‘ABC_clean.py’, the result will be saved in the folder ‘web scraping scripts’ as a .xlsx file, it should look like the following picture:

newsTitle	link	date	text	topics
Terrific Tigers take the flag as Gia	https://www.abc.net.au/news/2019-0	2019-09-28T20:29+1000	Richmond has swept to victory in the AFL grand final	['sport', 'australian-foot
Tigerland erupts as Richmond fa	https://www.abc.net.au/news/2019-0	2019-09-28T20:34+1000	Thousands of ecstatic Tigers fans have taken to the s	['sport', 'australian-foot
Baby born by C-section after mo	https://www.abc.net.au/news/2019-0	2019-09-28T19:02+1000	A baby has been delivered by cesarean section after	['domestic-violence', 'c
Frydenberg announces first majo	https://www.abc.net.au/news/2019-0	2019-09-28T17:54+1000	The first major look at Australia's savings system in 3	['superannuation', 'regu
Lucas Fowler's murder was 'randi	https://www.abc.net.au/news/2019-0	2019-09-28T11:23+1000	Two Canadian teens confessed to murdering three p	['law-crime-and-justice
Opinion: The cruel, creepy world	https://www.abc.net.au/news/2019-0	2019-09-28T10:04+1000	In spending any time probing the blather of bottom	['activism-and-lobbying
Japan stuns world number two Ir	https://www.abc.net.au/news/2019-0	2019-09-28T20:25+1000	Japan has stunned Ireland 19-12 in an extraordinary	['sport', 'rugby-union', '
Analysis: As the playlist begins, M	https://www.abc.net.au/news/2019-0	2019-09-28T11:38+1000	Before Scott Morrison got on the stage alongside D	['government-and-pol
Rape victim fights to keep 'absol	https://www.abc.net.au/news/2019-0	2019-09-28T13:54+1000	The only surviving victim of Tasmanian rapist and kil	['law-crime-and-justice
Analysis: How did CEO pay get to	https://www.abc.net.au/news/2019-0	2019-09-28T11:36+1000	Just as the dust was settling on the latest furore over	['business-economics-&
'You don't know what he has do	https://www.abc.net.au/news/2019-0	2019-09-28T09:55+1000	Jody Gore lived through years of violence at the han	['domestic-violence', 'w
When Sally became a mum, she	https://www.abc.net.au/news/2019-0	2019-09-27T09:29+1000	Having a baby is meant to be one of the most joyful	['parenting', 'depressio
Spider-Man won't be leaving the	https://www.abc.net.au/news/2019-0	2019-09-28T12:31+1000	The millions of Marvel fans around the world have b	['film-movies', 'film', 'ar
Parents' surprise grand final day	https://www.abc.net.au/news/2019-0	2019-09-28T17:46+1000	Police helped deliver a baby girl in the early hours o	['tottenham-3012', 'vic
'Buy lotto tickets': Sailing trio des	https://www.abc.net.au/news/2019-0	2019-09-28T13:59+1000	Three sailors have described their lucky rescue by a l	['disasters-and-acciden
'It was something normal to do':	https://www.abc.net.au/news/2019-0	2019-09-28T14:19+1000	You did not need a stopwatch or a measuring tape t	['sport', 'athletics', 'gata
How the introduction of predato	https://www.abc.net.au/news/2019-0	2019-09-28T12:50+1000	Until 2012, the animal inhabitants of Maria Island we	['mammals', 'environme
'Disenfranchised' millennials until	https://www.abc.net.au/news/2019-0	2019-09-28T12:03+1000	The future is looking increasingly bleak for Australiar	['business-economics-&
Scientists are building yeast 2.0. I	https://www.abc.net.au/news/science	2019-09-28T11:59+1000	ShareABC ScienceBrewer's yeast is a crucial ing	['business-economics-&
Two men arrived in Tasmania as	https://www.abc.net.au/news/2019-0	2019-09-28T10:37+1000	The stories of two Tasmanian officers added to the r	['history', 'community-&
Pakistan's Imran Khan warns of t	https://www.abc.net.au/news/2019-0	2019-09-28T16:15+1000	Pakistan Prime Minister Imran Khan has warned of a l	['unrest-conflict-and-w

Data Storage

We got the title, the date, the contents or the summary, the link for each piece of news. The category is also included for the ABC and NHK. **The result of web scraping is saved as .xlsx file for further integrating and analysis.** We name all data source files with its date and source name, as can be shown below. The folder is named as **PythonProject_DataStorage**.

 20190923_bbc.xlsx	2019-09-24 3:21 A...	Microsoft Excel W...	17 KB
 20190923aljazeera.xlsx	2019-09-23 6:00 PM	Microsoft Excel W...	13 KB
 20190923folha.xlsx	2019-09-23 6:00 PM	Microsoft Excel W...	30 KB
 20190923nhk.xlsx	2019-09-23 6:00 PM	Microsoft Excel W...	26 KB
 20190923renmin.xlsx	2019-10-04 5:37 PM	Microsoft Excel W...	42 KB
 20190924_bbc.xlsx	2019-09-24 3:22 A...	Microsoft Excel W...	16 KB
 20190924aljazeera.xlsx	2019-09-23 6:00 PM	Microsoft Excel W...	10 KB
 20190924folha.xlsx	2019-09-23 6:00 PM	Microsoft Excel W...	24 KB
 20190924nhk.xlsx	2019-09-23 6:00 PM	Microsoft Excel W...	18 KB
 20190924renmin.xlsx	2019-10-04 5:37 PM	Microsoft Excel W...	38 KB
 20190925_bbc.xlsx	2019-09-25 12:12 ...	Microsoft Excel W...	13 KB

c) Important News

For the data source ‘important news’, it contains more information other than web scraping, and it’s in JSON format. But the file of this source is too large, it takes too long to upload, download, and process as well. So we do not include the file here, the website to download this dataset is <https://webhose.io/free-datasets/popular-news-articles/>. **This program cannot run without the file, but the code and snapshot of results are shown below anyway:**

A	B	C	D	E
title	link	date	text	likes
Bears could	http://www	2017-02-0	Bears could	170
Mother ph	http://www	2017-02-0	Share this	510
Our impac	http://www	2017-01-3	What is ou	170
French giv	http://www	2017-02-0	Share 45 s	130
Bowling Gr	http://www	2017-02-0	February 3	200
Global foss	http://www	2017-02-0	Demand fo	160
Valentine H	http://www	2017-02-0	February 1	140
Duterte to	http://www	2017-01-3	Headlines	730
Trump Me	http://fort	2017-02-0	President D	160
Students le	http://www	2017-02-0	Students e	380
Will DeVos	http://www	2017-02-0	If Betsy De	530
The Bulldo	https://uga	2017-01-3	The Bulldo	170
2017 Natic	http://www	2017-02-0	32m ago •	880
Beirut's las	https://ww	2017-02-0	T ake a str	180
Identity Fra	http://www	2017-02-0	Feb 2 2017	340
Who was F	http://www	2017-02-0	"Frederick	110

The column ‘likes’ indicates how many people liked this on facebook.

Other information: Data Cleaning & Integration

a) **readData_Vfinal.py**

This file is used for data cleaning and data wrangling. Emotions are analyzed and stored(4 more columns); text are all also filtered (remove stopwords) and stored (2 more columns). There are a lot of effort here.

This file reads all the .xlsx files in the **PythonProject_DataStorage**, and generates the final results as another.xlsx file as **concat.xlsx**

See the result below:

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
	source	searchDate	title	link	date	text	categories	olarity_of_t	olarity_of_t	olarity_of_t	olarity_of_t	olarity_of_t	filtered_text	filtered_title			
0	ABC	20190928	Terrific Tig	https://www	20190928	Richmond has swept t	0.138029	0.1	0.459215	0.775	['richmond', 'terrific', 'tigers', 'take', 'flag', 'giants', 'grand', 'final']						
1	ABC	20190928	Tigerland e	https://www	20190928	Thousands of ecstatic	0.180882	0.433333	0.507353	0.8	['thousand', 'tigerland', 'erupts', 'richmond', 'fans', 'celebrate', 'i						
2	ABC	20190928	Baby born	https://www	20190928	A baby has been deliv	0.055655	0	0.353571	0.8	['baby', 'de baby', 'born', 'c-section', 'mother', 'critically', 'injur						
3	ABC	20190928	Frydenberg	https://www	20190928	The first major look at	0.065856	0.15625	0.503277	0.416667	['first', 'maj', 'frydenberg', 'announces', 'first', 'major', 'review', 'a						
4	ABC	20190928	Lucas Fowl	https://www	20190928	Two Canadian teens c	0.143304	-0.5	0.449702	0.5	['two', 'can', 'lucas', 'fowler', "s", 'murder', '"random", 'crime', 'c						
5	ABC	20190928	Opinion: T	https://www	20190928	In spending any time	-0.02692	-0.33333	0.482411	0.833333	['spending', 'opinion', 'cruel', 'creepy', 'world', 'think', "s", 'ok', '						
6	ABC	20190928	Japan stun	https://www	20190928	Japan has stunned Ire	0.17284	0	0.442372	0	['japan', 'st', 'japan', 'stuns', 'world', 'number', 'two', 'ireland', 'w						
7	ABC	20190928	Analysis: A	https://www	20190928	Before Scott Morrison	0.022638	0	0.416944	0.5	['scott', 'm', 'analysis', 'playlist', 'begins', 'morrison', 'discovers',						
8	ABC	20190928	Rape victim	https://www	20190928	The only surviving vict	0.017439	-0.09167	0.601025	0.55	['surviving', 'rape', 'victim', 'fights', 'keep', '"absolute", 'monster						
9	ABC	20190928	Analysis: H	https://www	20190928	Just as the dust was s	0.138891	-0.25	0.41217	0.5	['dust', 'set', 'analysis', 'ceo', 'pay', 'get', '500', 'times', 'wages', 'c						
10	ABC	20190928	'You don't	https://www	20190928	Jody Gore lived throu	0.026397	-0.875	0.450493	1	['jody', 'go', "you", "n't", 'know', 'done', 'brutal', 'truth', "ladies'						
11	ABC	20190928	When Sally	https://www	20190927	Having a baby is meat	0.107154	-0.5	0.531648	1	['baby', 'm', 'sally', 'became', 'mum', 'also', 'became', 'really', 'ai						
12	ABC	20190928	Spider-Man	https://www	20190928	The millions of Marvel	0.210347	0	0.494444	0	['millions', 'spider-man', 'wo', "n", 'leaving', 'marvel', 'univer						
13	ABC	20190928	Parents' su	https://www	20190928	Police helped deliver a	0.115895	0.25	0.548457	1	['police', 'h', 'parents', 'surprise', 'grand', 'final', 'day', 'baby', 'gi						
14	ABC	20190928	'Buy lotto	https://www	20190928	Three sailors have des	0.052186	0	0.559318	0	['three', 'sa', 'buy', 'lotto', 'tickets', 'sailing', 'trio', 'describe', 'sir						
15	ABC	20190928	'It was som	https://www	20190928	You did not need a st	0.150673	0.075	0.496946	0.325	['need', 'st', "it", 'something', 'normal', 'world', 'championships						
16	ABC	20190928	How the in	https://www	20190928	Until 2012, the animal	0.165617	0	0.444676	0	['2012', 'an', 'introduction', 'predators', 'rapidly', 'changed', 'isla						
17	ABC	20190928	Disenfranc	https://www	20190928	The future is looking i	0.201238	-0.5	0.514493	0.5	['future', 'ic', 'disenfranchised', 'millennials', 'unlikely', 'afford', '						
18	ABC	20190928	Scientists a	https://www	20190928	ShareABC ScienceSha	0.117965	0.3	0.495117	0.8	['shareabc', 'scientists', 'building', 'yeast', '2.0', "s", "s", 'exciting						
19	ABC	20190928	Two men a	https://www	20190928	The stories of two Tas	-0.00702	0	0.387587	0	['stories', 't', 'two', 'men', 'arrived', 'tasmania', 'convicts', 'almost						
20	ABC	20190928	Pakistan's	https://www	20190928	Pakistan Prime Minist	0.015167	0	0.503334	0	['pakistani', 'pakistan', "s", 'imran', 'khan', 'warns', '"bloodbath						
21	ABC	20190928	'We're goin	https://www	20190928	A Chinese journalist h	0.064794	-0.2	0.384221	0.2	['chinese', 'we', "re", 'going', 'see', 'trouble', 'hong', 'kong-b						
22	ABC	20190928	Wearing hi	https://www	20190927	Ten-year-old Sherick	0.076149	0.25	0.436304	0.333333	['ten-year', 'wearing', 'first', 'blade-runner', 'prosthetic', '10yo						
71	ABC	20190928	Teenager H	https://www	20190928	A teenage boy has dis	-0.01111	-0.4	0.386111	0.4	['teenage', 'teenager', 'killed', 'another', 'fighting', 'life', 'melbc						
72	ABC	20190928	Live: The R	https://www	20190928	In a replay of last year	0.153788	0.068182	0.495833	0.75	['replay', 'l', 'live', 'roosters', 'host', 'storm', 'blockbuster', 'preli						
94	ABC	20190928	Double wh	https://www	20190928	Visitor numbers have	0.007311	0	0.471621	0	['visitor', 'n', 'double', 'whammy', 'gold', 'coast', 'hinterland', 'tn						

b) **Data loading**

We can launch our system quickly by saving the integrated file statically. However, it creates a problem. The column 'filtered_text' and 'filtered_title' contains the list of words of each piece of news. But when we write it into file then read them back, the list becomes a string. To solve this problem, the modifyDf () function in '**readExcel.py**' is created to transform it back to list. See below:

```

2 # This function basically resolves a problem described below:
3 # We want to store our data locally and statically in a .xlsx file(otherwise, the process o
4 # But the data type of column filtered_text and filtered_title is list, if we write the fil
5 # This function change these two columns back to list
6 import re
7 import pandas as pd
8 def modifyDf(df):
9     ft_list = []
10    for line in df['filtered_text']:
11        df_line_list = re.findall(r'.*$', line);
12        for df_line in df_line_list:
13            ft_list.append(df_line.replace("", "").split(", "))
14    df['filtered_text'] = pd.Series(ft_list)
15    ft_list = []
16    for line in df['filtered_title']:
17        df_line_list = re.findall(r'.*$', line);
18        for df_line in df_line_list:
19            ft_list.append(df_line.replace("", "").split(", "))
20    df['filtered_title'] = pd.Series(ft_list)
21    return df
22 #testing code
23 if __name__ == '__main__':
24     df = pd.read_excel("concat.xlsx")
25     df = modifyDf(df)
26     print(type(df.iloc[0]['filtered_title']))# should display<class 'list'>

```

modifyDf()

n: readExcel ×

```

/usr/local/bin/python3.7 /Users/cloudyboat/Desktop/python/python_data/readExcel.py
<class 'list'>

Process finished with exit code 0

```

As the output shows, the column type of ‘filtered_text’ and ‘filtered_title’ is transformed back to list.

Other information: Linear Regression & Machine Learning

a) Linear Regression

We did a correlation analysis between the economic and sentiments of news. In the ‘linear_regression.py’, we use the change of Dow Jones Index to its previous date as y, and the average degree of sentiment for each channel as x1 to x5. You can run this file and the result will be shown below:

OLS Regression Results									
Dep. Variable:	difference	R-squared:	1.000						
Model:	OLS	Adj. R-squared:	nan						
Method:	Least Squares	F-statistic:	0.000						
Date:	Mon, 07 Oct 2019	Prob (F-statistic):	nan						
Time:	19:59:26	Log-Likelihood:	132.31						
No. Observations:	5	AIC:	-254.6						
Df Residuals:	0	BIC:	-256.6						
Df Model:	4								
Covariance Type:	nonrobust								
	coef	std err	t	P> t	[0.025	0.975]			
const	-1210.6542	inf	-0	nan	nan	nan			
BBC	4319.7468	inf	0	nan	nan	nan			
Renmin	3746.4688	inf	0	nan	nan	nan			
NHK	4531.6400	inf	0	nan	nan	nan			
Aljazeera	-1664.6700	inf	-0	nan	nan	nan			
Folha	-5517.7375	inf	-0	nan	nan	nan			
Omnibus:		nan	Durbin-Watson:		0.294				
Prob(Omnibus):		nan	Jarque-Bera (JB):		0.957				
Skew:		-1.056	Prob(JB):		0.620				
Kurtosis:		2.643	Cond. No.		92.4				

Warnings:

- [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
- [2] The input rank is higher than the number of observations.

This model is overfitted, and not robust enough, since we only used five sequences of data, but it still shows that there is a correlation between the stock market and emotions of news. The parameter is further explained below:

$$Y = -1210 + 4319*X1 + 3146*X2 + 4531X3 - 1664*X4 - 5575 * X5$$

Y: the change of Dow Jones Index compare to its previous date

X1: the average emotion ‘polarity of title’ of BBC

X2:the average emotion ‘polarity of title’ of Renmin

X3:the average emotion ‘polarity of title’ of NHK

X4:the average emotion ‘polarity of title’ of Aljazeera

X5:the average emotion ‘polarity of title’ of Folha

b) Topic analysis

‘topic_analysis.py’ is used to conduct the LDA model, you can run the file, the result report will be saved in txt file in the default folder. See the example of report below:

```

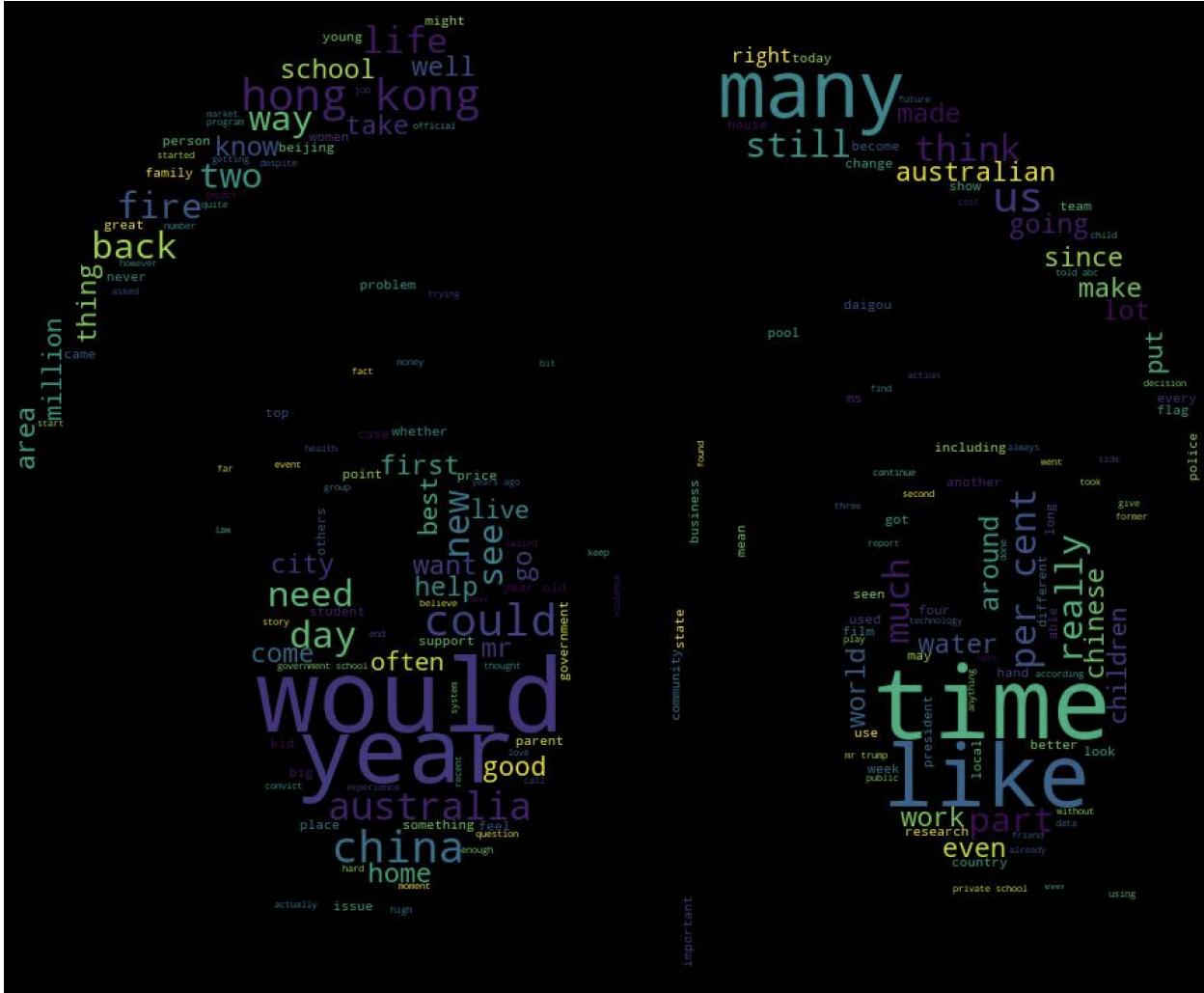
=====
doc:channel=====
doc 0: BBC
doc 1: ABC
doc 2: Folha
doc 3: Renmin
doc 4: Aljazeera
doc 5: NHK
=====
doc:topic=====
[[2.24215247e-04 2.24215247e-04 7.86995516e-02 8.31838565e-02
 2.13228700e-01 1.10089686e-01 2.10986547e-01 1.54932735e-01
 2.24215247e-04 1.48206278e-01]
 [1.25723371e-01 9.22608769e-02 1.07465358e-01 1.05085709e-01
 2.17348271e-02 1.13635598e-01 1.72158110e-01 1.06496652e-01
 9.12711115e-02 6.41683865e-02]
 [6.46871686e-03 1.06044539e-04 2.97985154e-02 4.67656416e-02
 1.49522800e-02 1.06044539e-04 1.86744433e-01 4.04029692e-02
 1.06044539e-04 6.74549311e-01]
 [4.38596491e-04 4.38596491e-04 4.38596491e-04 4.38596491e-04
 4.38596491e-04 4.38596491e-04 2.24122807e-01 4.38596491e-04
 4.38596491e-04 7.72368421e-01]
 [4.92610837e-04 4.92610837e-04 4.92610837e-04 4.92610837e-04
 4.92610837e-04 4.92610837e-04 1.38423645e-01 4.92610837e-04
 4.92610837e-04 8.57635468e-01]
 [5.10223048e-02 1.03263114e-05 9.48058653e-02 1.05235440e-01
 2.81608839e-01 8.09686080e-02 1.53665841e-01 1.69878150e-01
 1.45600991e-03 6.13486163e-02]]
=====
topic:word=====
(10, 13100)
[[1.51492198e-06 1.51492198e-06 1.51492198e-06]
 [2.21631206e-06 2.21631206e-06 2.21631206e-06]
 [1.60901046e-06 1.60901046e-06 1.60901046e-06]
 [1.60745861e-06 1.60745861e-06 6.44590902e-04]
 [2.50062516e-06 2.50062516e-06 2.50062516e-06]
 [1.57232704e-06 1.57232704e-06 1.57232704e-06]
 [9.85901607e-07 1.98166223e-04 9.85901607e-07]
 [2.30691643e-03 1.45533141e-04 1.44092219e-06]
 [2.23264121e-06 2.23264121e-06 2.23264121e-06]
 [2.07296849e-06 4.16666667e-04 2.07296849e-06]]
=====
topic top10 word=====
*Topic 0
-nt like - mr schools ms school really might good
*Topic 1
-- violence government daigou wallabies million many less says program
*Topic 2
--"re" get much kong australian "m" make know need left
*Topic 3
--would australia china told go new day three another young
*Topic 4
--mr police said "nt" anu staff hackers black car government
*Topic 5
--time per back even see way may work next got
*Topic 6
--"s" said one years world first hong could think us
*Topic 7
--people also "ye" two around life cent -- going home
*Topic 8
--often women ... planet funding place emissions cup lead points
*Topic 9
--says government president police fires minister already public law two
=====
doc best topic=====
doc: 0 ,best topic: 4
doc: 1 ,best topic: 6
doc: 2 ,best topic: 9
doc: 3 ,best topic: 9
doc: 4 ,best topic: 9
doc: 5 ,best topic: 4

```

Other information: Visualization

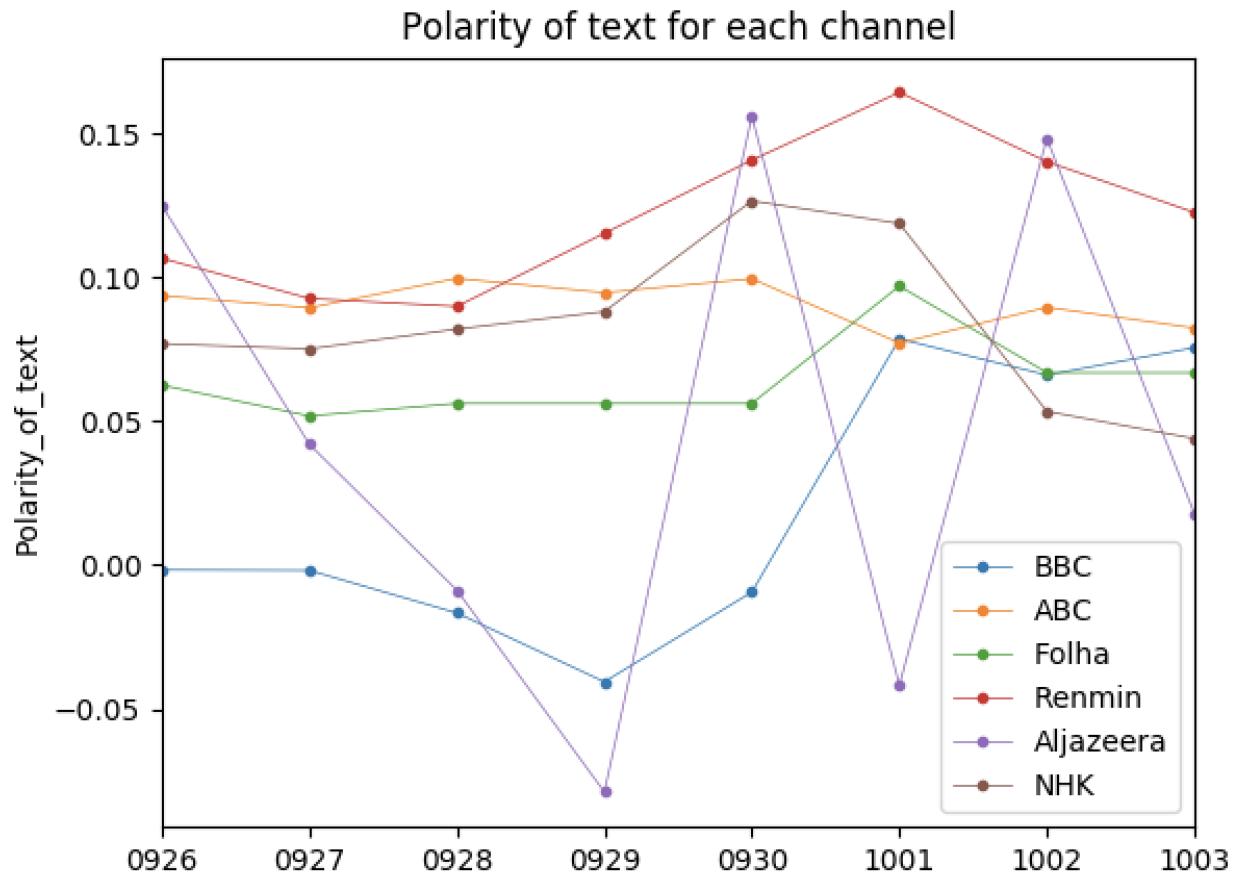
a) Word Cloud

‘generate_word_cloud.py’ is used to generate word cloud picture for each date. The results will be saved as ‘**2019MMDD.jpg**’ in the default directory, run the program and see the results below:



b) Line Charts of sentiments for each channel

‘analyze_channels.py’ is used to do the analysis in the channel dimension. The result is visualized and stored in the default folder in .png format, run the file and see the example result below:



Other information: Functions Providers

There are some files provide functions to other parts of our product without direct output, please check them individually.

They are ‘find_relevant.py’, ‘analyze_date.py’, ‘news_eachsource.py’