

# **EVALUATING COMMUNITY BUILDING EFFECTIVENESS OF TRANSPORTATION INVESTMENTS: KNOWLEDGE TRANSFER WEBINAR SERIES**

**WEBINAR II: SOCIAL MEDIA DATA DOWNLOAD  
AND ANALYSIS FOR TRANSPORTATION PROJECTS  
PART 2: DEMO**

*Presented by*

**Samiul Hasan, Assistant Professor**

**Naveen Eluru, Professor**

**Jiechao Zhang, PhD Student**

**Civil, Environmental, and Construction Engineering  
University of Central Florida**

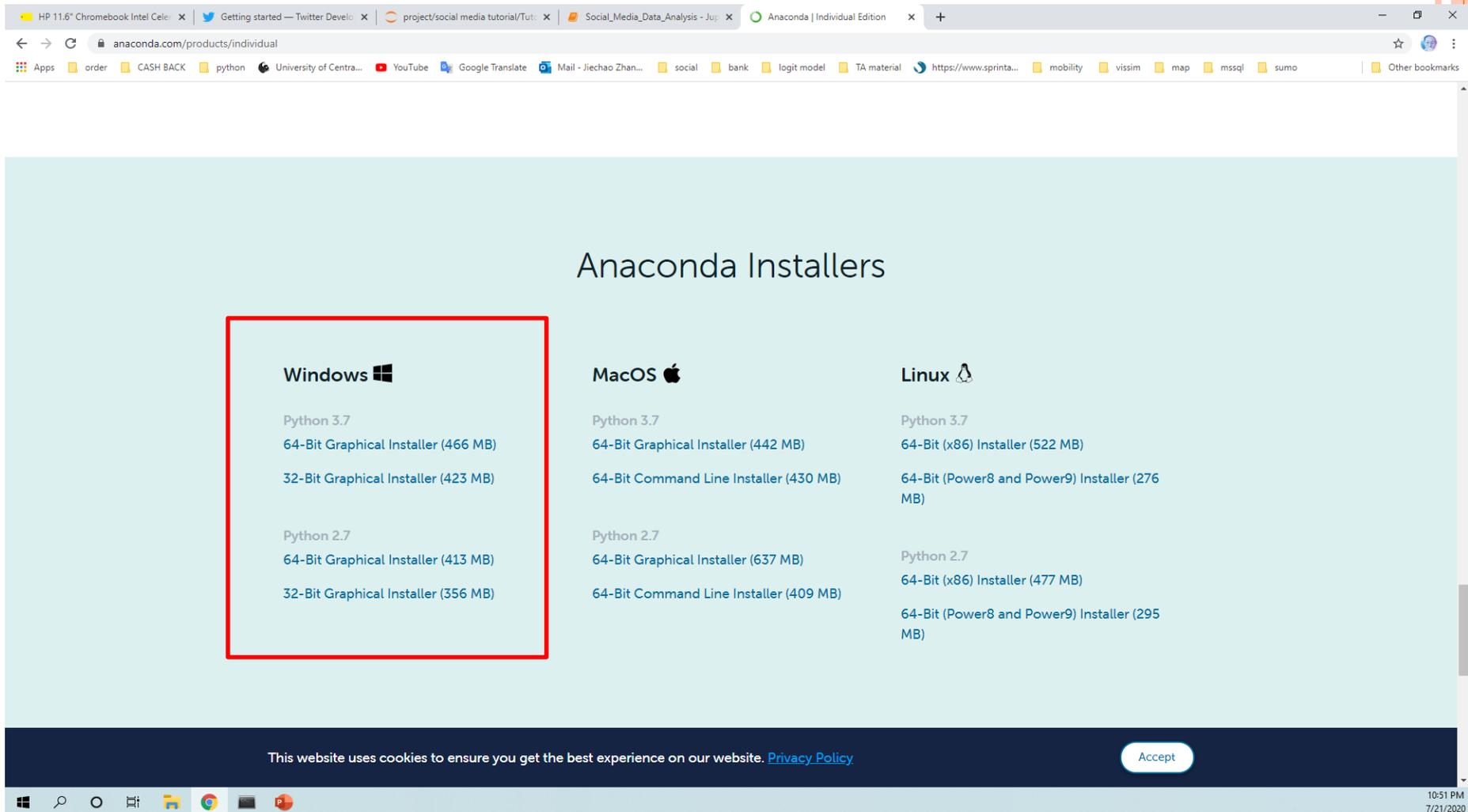
# OUTLINE

- Demo 1: Software installation and data downloading
- Demo 2: Sentiment Analysis and Visualization
- Demo 3: Topic Analysis and Visualization

# **Demo 1: Software Installation and Data Downloading**

# DATA COLLECTION – DOWNLOAD ANACONDA

Download Link: <https://www.anaconda.com/products/individual>



The screenshot shows a web browser window displaying the Anaconda website. The browser's address bar shows the URL [anaconda.com/products/individual](https://www.anaconda.com/products/individual). The page content is titled "Anaconda Installers" and is organized into three columns: Windows, MacOS, and Linux. Each column lists installers for Python 3.7 and Python 2.7, with options for graphical and command-line installation. The Windows section is highlighted with a red rectangular box.

Operating System	Python Version	Installer Type	Size	
Windows	Python 3.7	64-Bit Graphical Installer	466 MB	
		32-Bit Graphical Installer	423 MB	
	Python 2.7	64-Bit Graphical Installer	413 MB	
		32-Bit Graphical Installer	356 MB	
	MacOS	Python 3.7	64-Bit Graphical Installer	442 MB
			64-Bit Command Line Installer	430 MB
Python 2.7		64-Bit Graphical Installer	637 MB	
		64-Bit Command Line Installer	409 MB	
Linux	Python 3.7	64-Bit (x86) Installer	522 MB	
		64-Bit (Power8 and Power9) Installer	276 MB	
	Python 2.7	64-Bit (x86) Installer	477 MB	
		64-Bit (Power8 and Power9) Installer	295 MB	

This website uses cookies to ensure you get the best experience on our website. [Privacy Policy](#) Accept

# DATA COLLECTION – OPEN ANACONDA NAVIGATOR

Anaconda Navigator

File Help

ANACONDA NAVIGATOR

Sign in to Anaconda Cloud

Home

Applications on base (root)

Channels

Refresh

Environments

Learning

Community



CMD.exe Prompt  
0.1.1

Run a cmd.exe terminal with your current environment from Navigator activated

Launch



JupyterLab  
2.1.4

An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.

Launch



Jupyter Notebook  
6.0.3

Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.

Launch



Powershell Prompt  
0.0.1

Run a Powershell terminal with your current environment from Navigator activated

Launch



Glueviz  
0.15.2

Multidimensional data visualization across files. Explore relationships within and among related datasets.

Install



Orange 3  
3.26.0

Component based data mining framework. Data visualization and data analysis for novice and expert. Interactive workflows with a large toolbox.

Install



Qt Console  
4.7.5

PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.

Install



RStudio  
1.1.456

A set of integrated tools designed to help you be more productive with R. Includes R essentials and notebooks.

Install



Spyder  
4.1.3

Scientific Python Development Environment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features

Install

Documentation

Developer Blog



Type here to search



10:54 PM  
7/21/2020



# DATA COLLECTION – INSTALL PACKAGES

Celer x | Getting started — Twitter Develo x | project/social media tutorial/Tut: x | Social\_Media\_Data\_Analysis - Jup: x | +

it:8888/notebooks/project/social%20media%20tutorial/Tutorial/Tutorial/Social\_Media\_Data\_Analysis.ipynb

SH BACK python University of Centra... YouTube Google Translate Mail - Jiechao Zhan... social bank logit model TA material https://www.sprinta... mobility vissim map mssql sumo

jupyter Social\_Media\_Data\_Analysis Last Checkpoint: Last Thursday at 1:37 PM (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

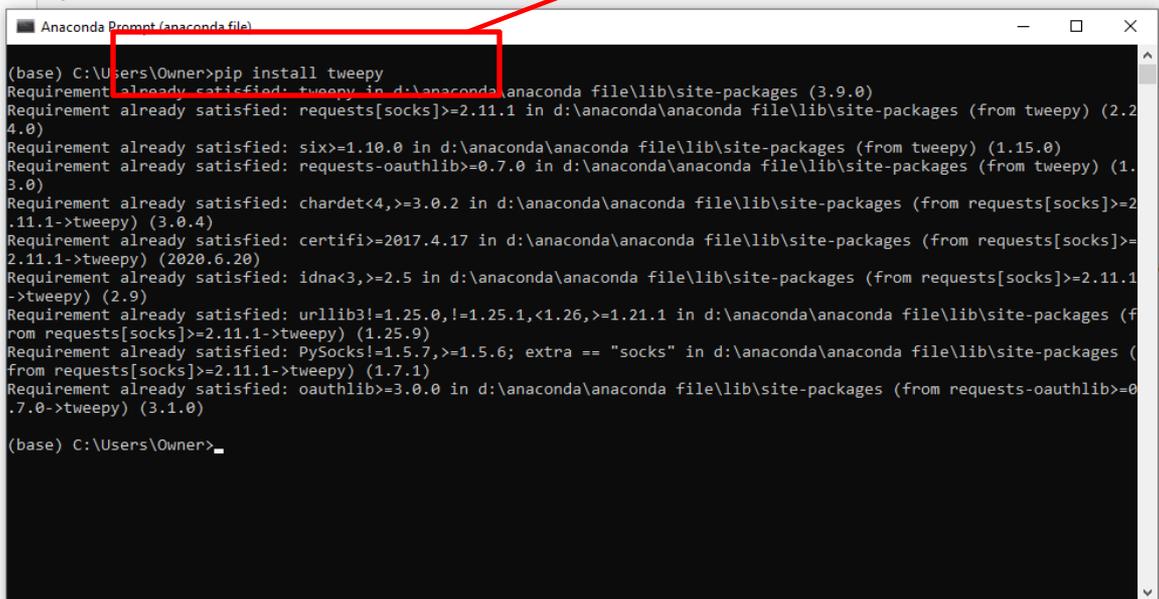
## 1 Data Collection (user accounts)

In [2]:

```
#!/usr/bin/env python
# encoding: utf-8

import tweepy # https://github.com/tweepy/tweepy
import csv, pdb
import time
```

**pip install tweepy**



```
(base) C:\Users\Owner>pip install tweepy
Requirement already satisfied: tweepy in d:\anaconda\anaconda file\lib\site-packages (3.9.0)
Requirement already satisfied: requests[socks]>=2.11.1 in d:\anaconda\anaconda file\lib\site-packages (from tweepy) (2.24.0)
Requirement already satisfied: six>=1.10.0 in d:\anaconda\anaconda file\lib\site-packages (from tweepy) (1.15.0)
Requirement already satisfied: requests-oauthlib>=0.7.0 in d:\anaconda\anaconda file\lib\site-packages (from tweepy) (1.3.0)
Requirement already satisfied: chardet<4,>=3.0.2 in d:\anaconda\anaconda file\lib\site-packages (from requests[socks]>=2.11.1->tweepy) (3.0.4)
Requirement already satisfied: certifi>=2017.4.17 in d:\anaconda\anaconda file\lib\site-packages (from requests[socks]>=2.11.1->tweepy) (2020.6.20)
Requirement already satisfied: idna<3,>=2.5 in d:\anaconda\anaconda file\lib\site-packages (from requests[socks]>=2.11.1->tweepy) (2.9)
Requirement already satisfied: urllib3!=1.25.0,!<1.25.1,<1.26,>=1.21.1 in d:\anaconda\anaconda file\lib\site-packages (from requests[socks]>=2.11.1->tweepy) (1.25.9)
Requirement already satisfied: PySocks!=1.5.7,>=1.5.6; extra == "socks" in d:\anaconda\anaconda file\lib\site-packages (from requests[socks]>=2.11.1->tweepy) (1.7.1)
Requirement already satisfied: oauthlib>=3.0.0 in d:\anaconda\anaconda file\lib\site-packages (from requests-oauthlib>=0.7.0->tweepy) (3.1.0)

(base) C:\Users\Owner>
```

```
oldest = alltweets[-1].id - 1

# keep grabbing tweets until there are no tweets left to grab
while len(new_tweets) > 0:
    # print "getting tweets before %s" % (oldest)

    # all subsequent requests use the max_id param to prevent duplicates
    new_tweets = api.user_timeline(screen_name=screen_name, count=200, max_id=oldest)
```

# DATA COLLECTION – APPLY TWITTER DEVELOPER ACCOUNT

Download Link: <https://developer.twitter.com/en/apply-for-access>

The screenshot shows a web browser window displaying the Twitter Developer 'Apply for access' page. The browser's address bar shows the URL 'developer.twitter.com/en/apply-for-access'. The page features a purple header with the Twitter Developer logo and navigation links. The main content area includes a sub-header 'Get started with Twitter APIs and tools', a large 'Apply for access' title, and a paragraph stating that all new developers must apply for a developer account. Below this, there are two buttons: 'Apply for a developer account' and 'Restricted use cases >'. A horizontal line separates this section from a grid of four API categories: Standard APIs, Premium APIs, Enterprise APIs, and Ads APIs. Each category has a brief description and a corresponding 'Apply' button.

HP 11.6" Chromebook Intel Celeron | Getting started — Twitter Developer | project/social media tutorial/Tutorial | Social\_Media\_Data\_Analysis - Jupyter | Apply for access – Twitter Developer

developer.twitter.com/en/apply-for-access

Developer Use cases Products Docs More Labs Apply Apps

Get started with Twitter APIs and tools

## Apply for access

All new developers must apply for a developer account to access Twitter APIs.

[Apply for a developer account](#) [Restricted use cases >](#)

---

Standard APIs	Premium APIs	Enterprise APIs	Ads APIs
Our free, standard APIs are great for getting started, testing an integration, or validating a concept.	Our premium APIs offer scalable access to Twitter data for those looking to grow, experiment, and innovate.	Our enterprise APIs offer the highest level of access and reliability to those who depend on Twitter data.	The Ads API gives partners a programmatic way to integrate with the Twitter Ads platform.
<a href="#">Apply for a developer account &gt;</a>	<a href="#">Apply for premium access &gt;</a>	<a href="#">Apply for enterprise access &gt;</a>	<a href="#">Apply for Ads APIs access &gt;</a>

12:41 AM 7/22/2020

# DATA COLLECTION – USER ACCOUNTS



## 1 Data Collection (user accounts) ¶

Necessary Packages

In [10]:

```
#!/usr/bin/env python  
# encoding: utf-8
```

```
import tweepy # https://github.com/tweepy/tweepy  
import csv, pdb  
import time
```

```
# Twitter API credentials  
twitter_app_auth = {  
    'consumer_key':  
    'consumer_secret'  
    'access_token':  
    'access_token_secret':  
}
```

Path - List\_User

```
List_User = r'D:\project\social media tutorial\example data>List_User.csv'
```

Path - Save File

```
def get_all_tweets(screen_name):
```

```
# Twitter only allows access to a users most recent 3240 tweets with this method  
Collection_Path = r'D:\project\social media tutorial\example data\collection data\%s_June_10_tweets.csv' % screen_name
```

# DATA COLLECTION – USER ACCOUNTS

AutoSave Off | List\_User - Excel | zlyjsl123@gmail.com

File Home Insert Page Layout Formulas Data Review View Help Search

Clipboard: Paste, Cut, Copy, Format Painter

Font: Calibri, 11, Bold, Italic, Underline, Text Color, Background Color

Alignment: Wrap Text, Merge & Center

Number: General, Currency, Percentage, Decimals

Styles: Conditional Formatting, Format as Table, Cell Styles

Cells: Insert, Delete, Format

Editing: AutoSum, Fill, Clear, Sort & Filter, Find & Select, Ideas

Share

fx: [Empty]

10

**List\_User File Example**

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	FL511_Estatal																				
2	321Transit																				
3	965traffic																				
4	BikeWalkCFL																				
5	fl_511_i4																				
6	FL511_95Express																				
7	fl511_central																				
8	fl511_i10																				
9	fl511_i75																				

AutoSave Off | Alexander Spring\_December\_06\_11\_tweets - Excel | zlyjsl123@gmail.com

File Home Insert Page Layout Formulas Data Review View Help Search

Clipboard: Paste, Cut, Copy, Format Painter

Font: Calibri, 11, Bold, Italic, Underline, Text Color, Background Color

Alignment: Wrap Text, Merge & Center

Number: Custom, Currency, Percentage, Decimals

Styles: Conditional Formatting, Format as Table, Cell Styles

Cells: Insert, Delete, Format

Editing: AutoSum, Fill, Clear, Sort & Filter, Find & Select, Ideas

Share Comments

fx: 6/10/2020 6:49:32 PM

1

**User Accounts Data File Example**

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	##### b"@Brexit JacksellArt	1	0	1.27E+18				BrexitBass	1.27E+18	1.27E+18	3.09E+09	FALSE	FALSE	Twitter W User_api=	TRUE	FALSE	#####	FALSE			0	{'hashtags' b"	7.
2	##### b"RT @Hic spring_art	0	10	1.27E+18								FALSE	FALSE	Twitter fo User_api=	FALSE	FALSE	#####	TRUE			333333	{'hashtags' b"\xe2\x8c	2.
3	##### b"Tailorbyr BestShopp	0	0	1.27E+18																		{'hashtags' b"	1.
4	##### b'.@AmCa AmCaritas	0	0	1.27E+18																		{'hashtags' b"Care is th	1.
5	##### b"RT @naf R4137688	0	9	1.27E+18																		{'hashtags' b"\xf0\x9f	1.
6	##### b"RT @wal Spring_Soc	0	540	1.27E+18																		{'hashtags' b"\xf0\x9f	1.
7	##### b"@Alexan AmiesPhili	0	0	1.27E+18				AmiesPhili	1.27E													{'hashtags' b"Naturalis	8.
8	##### b"RT @dio lau_lfrchi	0	418	1.27E+18								FALSE	FALSE	Twitter fo User_api=	FALSE	FALSE	#####	FALSE			333333	{'hashtags' b"Plus vif q	2.
9	##### b"RT @Doi CharlotteE	0	1	1.27E+18								FALSE	FALSE	Twitter fo User_api=	FALSE	FALSE	#####	TRUE			333333	{'hashtags' b"Director	3.

# DATA COLLECTION – KEYWORDS

## 2 Data Collection (keywords)

```
[4]: import tweepy
import csv
from tweepy import Stream
from tweepy import OAuthHandler
from tweepy.streaming import StreamListener

consumer_key = 
consumer_secret = 
access_token = '
access_token_secret = 

auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth)

# Open/Create a file to append data
#csvFile = open('%s_tweets.csv' % q, 'w', encoding='utf-8')
#Use csv Writer
#csvWriter = csv.writer(csvFile)

List_User = r'D:\project\social media tutorial\example data\0. List_KW.csv' #define the path

for line in open(List_User, 'r', encoding='utf-8'):
    keyword = line.strip()

    save_path = r'D:\project\social media tutorial\example data\key word\%s_December_06_11_tweets.csv' % keyword #define the path
```

# DATA COLLECTION – KEYWORDS

0.List\_KW File Example

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	DeLeon Springs																						
2	lynx bus																						
3	Salt Springs																						
4	Alexander Spring																						
5	Blue Spring																						
6	florida bus																						
7	florida crime																						
8	florida sidewalk																						
9	Florida Spring																						
10	florida walking																						
11	I4 Construction																						
12	I4 Crash																						
13	I4 Ultimate																						
14	Iuice Bike																						

Keywords Data File Example

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	##### b"@Brexit	JacksellArt	1	0	1.27E+18	BrexitB	1.27E+18	1.27E+18	3.08E+08	FALSE	FALSE	TRUE	FALSE	#####	#####	#####	#####	#####	#####	#####	0	{'hashtags'	b"
2	##### b"RT @Hic	spring_art	0	10	1.27E+18																333333	{'hashtags'	b"\xe2\x8c
3	##### b"Tailorbyr	BestShopp	0	0	1.27E+18																333333	{'hashtags'	b"
4	##### b'.@AmCa	AmCaritas	0	0	1.27E+18							FALSE	FALSE	Twitter W	User_	api=	TRUE	FALSE	#####	FALSE	333333	{'hashtags'	b"Care is tf
5	##### b"RT @naf	R4137688	0	9	1.27E+18							FALSE	FALSE	Twitter fo	User_	api=	FALSE	FALSE	#####	TRUE	333333	{'hashtags'	b"\xf0\x9f
6	##### b"RT @wal	Spring_Sou	0	540	1.27E+18							FALSE	FALSE	Twitter fo	User_	api=	FALSE	FALSE	#####	TRUE	333333	{'hashtags'	b"\xf0\x9f
7	##### b"@Alexan	AmiesPhili	0	0	1.27E+18	AmiesPhili	1.27E+18	1.27E+18	8.23E+08	FALSE	FALSE	Twitter fo	User_	api=	TRUE	FALSE	#####	TRUE	#####	TRUE	333333	{'hashtags'	b"Naturalis
8	##### b"RT @dio	lau_Ifrcbi	0	418	1.27E+18							FALSE	FALSE	Twitter fo	User_	api=	FALSE	FALSE	#####	FALSE	333333	{'hashtags'	b"Plus vif n

# **Demo 2: Sentiment Analysis and Visualization**

# SENTIMENT ANALYSIS

## 3 Sentiment Analysis - keywords

```
24]: import os
import pandas as pd
from textblob import TextBlob
```

```
path = r'D:\project\social media tutorial\example data\key word' #define the path
```

```
files = os.listdir(path) #define the files in the path
```

```
def modifystr(s):
    #s = s.str.replace('[^\w\s]', '')
    s = s.replace('/', '')
    s = s.replace(', ', '')
    s = s.replace('@', '')
    s = s.replace('///', '')
    s = s.replace('#', '')
    s = s.replace('%', '')
    s = s.replace(' ', '')
    s = s.replace('\ ', '')
    s = s.replace('!', '')
```



Path – Input Folder

# SENTIMENT ANALYSIS

```
for i in range(0,len(files)):
```

```
df = files[i]
```

```
df_name = path + "\\ " + df
```

```
df_name_save_path = r'D:\\project\\social media tutorial\\example data\\key word senti\\' +df
```

```
print(df_name_save_path)
```

```
df_final = pd.read_csv(df_name,
```

```
names = ['1','2','3','4','5','6','7','8','9','10','11','12','13',  
         '14','15','16','17','18','19','20','21','22','23','24'] )
```

```
df_final.dropna(axis = 0, how = 'all', inplace = True)
```

```
df_final['2']=list(map(modifystr,df_final['2']))
```

```
df_final['sentiment']=list(map(sentiment,df_final['2']))
```

```
df_final.index = range(0,len(df_final))
```

```
df_final_sentiment = pd.DataFrame(df_final[['1','2','6','7','sentiment']])
```

Path – Output Files

# SENTIMENT ANALYSIS

Alexander Spring\_December\_06\_11\_tweets - Excel

zlyjs123@gmail.com

File Home Insert Page Layout Formulas Data Review View Help Search

Clipboard Font Alignment Number Styles Cells Editing Ideas

6/10/2020 6:49:32 PM

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	
1	#####	b"@Brexit JacksellArt		1	0	1.27E+18		BrexitB	1.27E+18	1.27E+18	2.09E+08	FALSE	FALSE	Twitter W	User_api=	TRUE	FALSE	#####	FALSE			0	{'hashtags' b'	7
2	#####	b'RT @Hic spring_art		0	10	1.27E+18																333333	{'hashtags' b'\xe2\x8c	2
3	#####	b'Tailorbyr BestShopp		0	0	1.27E+18																333333	{'hashtags' b'	1
4	#####	b'.@AmCa AmCaritas		0	0	1.27E+18						FALSE	FALSE	Twitter W	User_api=	TRUE	FALSE	#####	FALSE			333333	{'hashtags' b'Care is th	1
5	#####	b'RT @naf R4137688		0	9	1.27E+18						FALSE	FALSE	Twitter fo	User_api=	FALSE	FALSE	#####	TRUE			333333	{'hashtags' b'\xf0\x9f	1
6	#####	b'RT @wal Spring_Sou		0	540	1.27E+18						FALSE	FALSE	Twitter fo	User_api=	FALSE	FALSE	#####	TRUE			333333	{'hashtags' b'\xf0\x9f	1
7	#####	b'RT @Alexan AmiesPhili		0	0	1.27E+18		AmiesPhili	1.27E+18	1.27E+18	8.23E+08	FALSE	FALSE	Twitter fo	User_api=	TRUE	FALSE	#####	TRUE			333333	{'hashtags' b'Naturalis	8
8	#####	b'RT @dio lau ifrchi		0	418	1.27E+18						FALSE	FALSE	Twitter fo	User_api=	FALSE	FALSE	#####	FALSE			333333	{'hashtags' b'Plus vif n	2

Input File Example

Alexander Spring\_December\_06\_11\_tweets - Excel

zlyjs123@gmail.com

File Home Insert Page Layout Formulas Data Review View Help Search

Clipboard Font Alignment Number Styles Cells Editing Ideas

M3

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1		1	2	6	7	sentiment	polarity	subjectively															
2	6/10/2020 18:49	BrexitBass	1.27E+18			Sentiment	-0.29688	0.5															
3	6/10/2020 18:31	RT Hich866	1.27E+18			Sentiment	0	0															
4	6/10/2020 18:02	Tailoryrd E	1.27E+18			Sentiment	-0.06667	0.133333															
5	6/10/2020 17:51	AmCaritas	1.27E+18			Sentiment	0	0															
6	6/10/2020 16:37	RT nafasei	1.27E+18			Sentiment	0	0															
7	6/10/2020 15:33	RT walaasi	1.27E+18			Sentiment	0	0															
8	6/10/2020 15:29	AlexanderI	1.27E+18			Sentiment	0	0															
9	6/10/2020 14:29	RT diorang	1.27E+18			Sentiment	0	0															
10	6/10/2020 14:24	RT Downe	1.27E+18			Sentiment	0.275	0.266667															
11	6/10/2020 14:20	RT Wildest	1.27E+18			Sentiment	0	0															
12	6/10/2020 14:16	RT Wildest	1.27E+18			Sentiment	0	0															
13	6/10/2020 14:04	RT Wildest	1.27E+18			Sentiment	0	0															
14	6/10/2020 14:00	FIRST TIMI	1.27E+18			Sentiment	0.2125	0.527083															

Output File Example

# SENTIMENT ANALYSIS VISUALIZATION

## 5 Sentiment Analysis Visualization

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
```

```
path = r'F:\sentiment result\final_data\whole result\Sunshine Skyway.csv'
df = pd.read_csv(path, header = 0, names = ['id', 'time', 'text', 'account', 'geotagged', 'sentiment', 'polarity', 'subjectivity'])
output_path = r'F:\sentiment result\final_data\whole result\Sunshine Skyway.png'
```

```
df.time = pd.to_datetime(df.time)
```

```
#select data based on the time (half year)
```

```
df_1 = df[(df.time.dt.year == 2017)&(df.time.dt.month>1)&(df.time.dt.month<8)]
df_2 = df[(df.time.dt.year == 2017)&(df.time.dt.month>7)&(df.time.dt.month<13)]
df_3 = df[(df.time.dt.year == 2018)&(df.time.dt.month>0)&(df.time.dt.month<9)]
```

```
fig, axes = plt.subplots(3, 1, sharex=True, sharey=True)
```

```
fig.set_size_inches(5,10)
```

```
axes[0].hist(df_1.polarity, density = 1, bins=20, color='r',)
axes[0].set_title('February 2017 - July 2017')
axes[0].set_ylabel('Density')
```

```
axes[1].hist(df_2.polarity, density = 1, bins=20, color='r')
axes[1].set_title('August 2017 - December 2017')
axes[1].set_ylabel('Density')
```

Path – Input Files

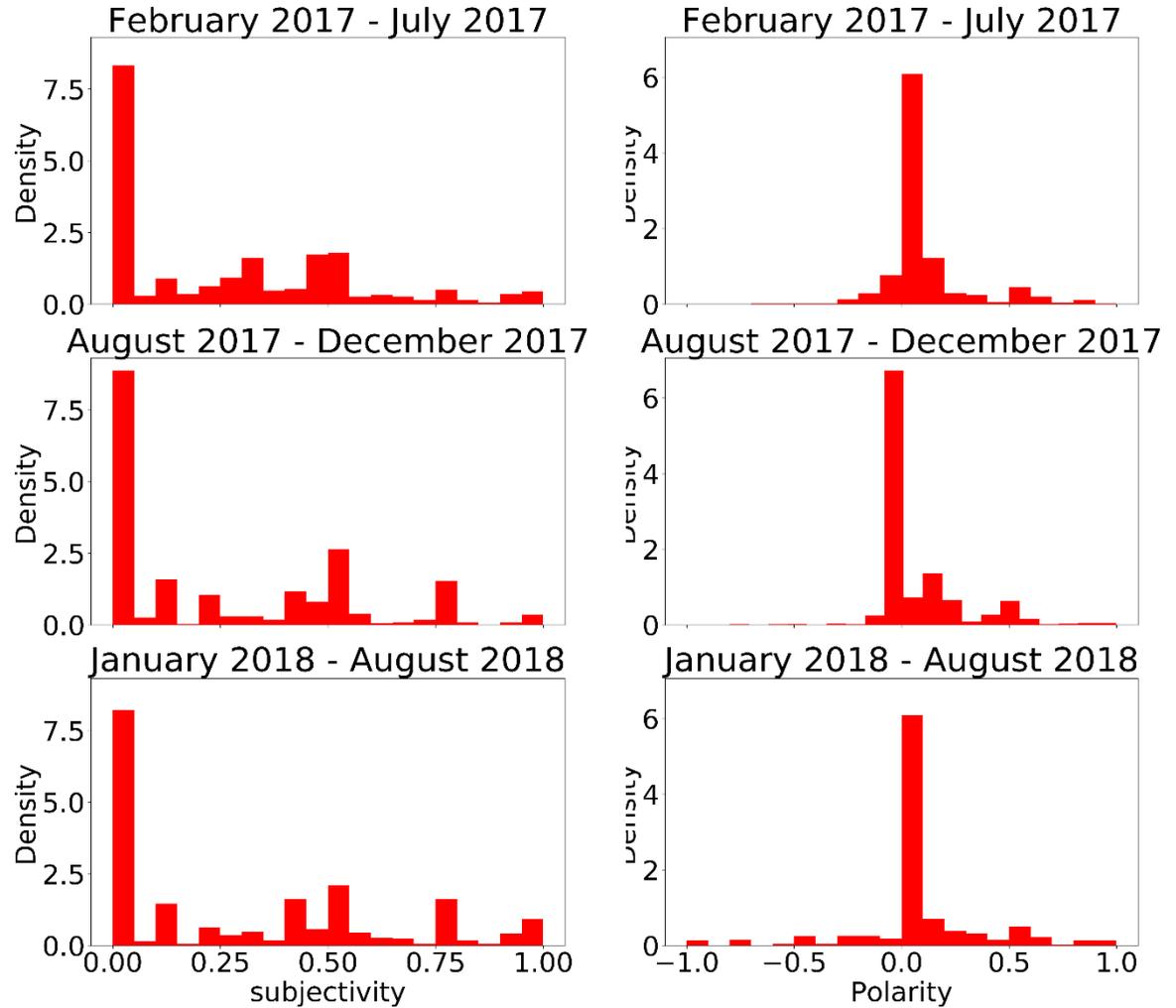
Path – Output Files

Set up the time periods

Time periods name in the figure

# SENTIMENT ANALYSIS VISUALIZATION

Example – Sentiment Analysis Visualization



# Demo 3: Topic Analysis

# TOPIC ANALYSIS – DATA PROCESSING

## 7 Topic Model Data Processing

In [68]:

```
import pandas as pd
import os
from textblob import TextBlob

path = 'F:\\sentiment result\\final_data\\whole result'

path_save = 'D:\\project\\social media tutorial\\example data\\topic model'

files = os.listdir(path)

for j in range(0, len(files)):

    df = files[j]
    path1 = path + "\\ " + df

    df_txt = df.replace('.csv', '')
    path_final = path_save + "\\ " + df_txt + '.txt'

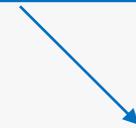
    names = ['index', 'time', 'text', 'id', 'geolocation', 'sentiment', 'polarity', 'subjectivity']

    df = pd.read_csv(path1)
```

Path – Input Folder



Path – Output Files



# TOPIC ANALYSIS – DATA PROCESSING

AutoSave Off | florida bus - Excel

File Home Insert Page Layout Formulas Data Review View Help Search

Clipboard Font Alignment Number

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		1	2	6	7	sentiment	polarity	subjectively					
2		0	##### bRT ChrisK	1.03E+18		Sentiment	0	0.0625					
3		1	##### bRT ChrisK	1.03E+18		Sentiment	0	0.0625					
4		2	##### bRT ChrisK	1.03E+18		Sentiment	0	0.0625					
5		3	##### bRT joann	1.03E+18		Sentiment	0.519048	0.590476					
6		4	##### bRT hamn	1.03E+18		Sentiment	0	0					
7		5	##### brealDona	1.03E+18		Sentiment	0.25	0.333333					
8		6	##### bRT Gwen	1.03E+18		Sentiment	0.2	0.45					
9		7	##### bSubtle hir	1.03E+18		Sentiment	0	0					
10		8	##### bRT ChrisK	1.03E+18		Sentiment	0	0.0625					
11		9	##### bRT Gwen	1.03E+18		Sentiment	0.5	0.6					
12		10	##### bRT ChrisK	1.03E+18		Sentiment	0	0.0625					
13		11	##### bEverywh	1.03E+18		Sentiment	0	0.0625					
14		12	##### bRT joann	1.03E+18		Sentiment	0.519048	0.590476					
15		13	#										
16		14	#										
17		15	#										
18		16	#										
19		17	##### bRT joann	1.03E+18		Sentiment	0.519048	0.590476					
20		18	##### bRT joann	1.03E+18		Sentiment	0.519048	0.590476					
21		19	##### bAwesome	1.03E+18		Sentiment	0.278571	0.385714					
22		20	##### bRT hamn	1.03E+18		Sentiment	0.5	0.6					
23		21	##### bRT hamn	1.03E+18		Sentiment	0	0					
24		22	##### bRT Gwen	1.03E+18		Sentiment	0.2	0.45					
25		23	##### bRT Gwen	1.03E+18		Sentiment	0.2	0.45					
26		24	##### bRT hamn	1.03E+18		Sentiment	0	0					
27		25	##### bRT Gwen	1.03E+18		Sentiment	0.2	0.45					

florida bus

Input File Example

florida bus - Notepad

File Edit Format View Help

```

1.0309664470679512e+18 "RT ChrisKingFL Everywhere we go in
1.0309663678377124e+18 "RT ChrisKingFL Everywhere we go in
1.030964885092528e+18 "RT ChrisKingFL Everywhere we go in
1.0309567547865496e+18 "RT joannefea Awesome bus tour educ
1.0309562966049956e+18 "RT hammel11 I havenxe2x80x99t said
1.0309554328677088e+18 "realDonaldTrumpnNRA FLGovScott Thi
1.030954321599447e+18 "RT GwenGraham South Florida friends
1.0309539232496681e+18 "Subtle hint from the Panama CityFl
1.0309516505743809e+18 "RT ChrisKingFL Everywhere we go in
1.0309511678561894e+18 "RT GwenGraham Our South Florida Ea
1.0309509267633888e+18 "RT ChrisKingFL Everywhere we go in
1.0309498999935384e+18 "Everywhere we go in Southwest Flor
1.0309494540426076e+18 "RT joannefea Awesome bus tour educ
1.0309452314060841e+18 "RT Cissy4Judge Cissy is very proud
    
```

```

1.0309363598239908e+18 "RT joannefea Awesome bus tour educ
1.0309362734255226e+18 "Awesome bus tour educating the vot
1.0309319836381184e+18 "RT GwenGraham Our South Florida Ea
1.0309318630980198e+18 "RT hammel11 I havenxe2x80x99t said
1.0309313480962662e+18 "RT GwenGraham South Florida friend
1.0309252249870458e+18 "RT GwenGraham South Florida friend
1.030918707630039e+18 "RT hammel11 I havenxe2x80x99t said
1.0309062709482291e+18 "RT GwenGraham South Florida friend
1.0309047633936957e+18 "RT GwenGraham South Florida friend
1.0309013396505025e+18 "RT hammel11 I havenxe2x80x99t said
1.0309007449484737e+18 "I havenxe2x80x99t said a damn thin
1.0309000108194816e+18 "WDWToday Ray from Davenport Florid
    
```

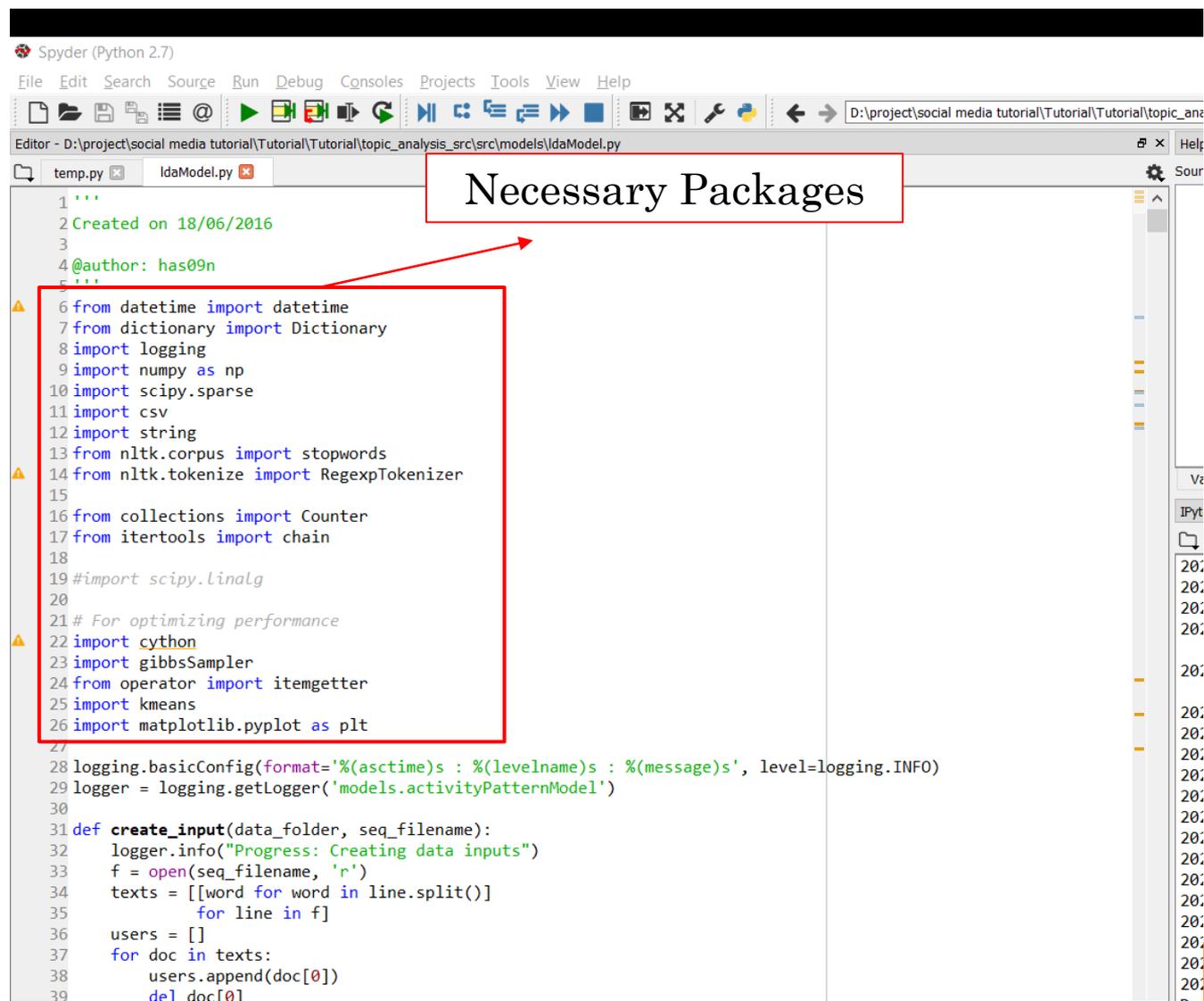
Output File Example

# TOPIC ANALYSIS – RUN TOPIC MODEL

- Download Anaconda (python 2.7) 32-bit Graphical Installer
- Open Spyder from the anaconda navigator (python 2.7) version
- Open the ldaModel.py
- Install all the necessary python packages
- Change the input path and file name
- Run the model

# TOPIC ANALYSIS – RUN TOPIC MODEL

## Example – Topic Analysis Model in Spyder



```
Spyder (Python 2.7)
File Edit Search Source Run Debug Consoles Projects Tools View Help
D:\project\social media tutorial\Tutorial\Tutorial\topic_an...
Editor - D:\project\social media tutorial\Tutorial\Tutorial\topic_analysis_src\src\models\ldaModel.py
temp.py x IdaModel.py x
1 '''
2 Created on 18/06/2016
3
4 @author: has09n
5 '''
6 from datetime import datetime
7 from dictionary import Dictionary
8 import logging
9 import numpy as np
10 import scipy.sparse
11 import csv
12 import string
13 from nltk.corpus import stopwords
14 from nltk.tokenize import RegexpTokenizer
15
16 from collections import Counter
17 from itertools import chain
18
19 #import scipy.linalg
20
21 # For optimizing performance
22 import cython
23 import gibbsSampler
24 from operator import itemgetter
25 import kmeans
26 import matplotlib.pyplot as plt
27
28 logging.basicConfig(format='%(asctime)s : %(levelname)s : %(message)s', level=logging.INFO)
29 logger = logging.getLogger('models.activityPatternModel')
30
31 def create_input(data_folder, seq_filename):
32     logger.info("Progress: Creating data inputs")
33     f = open(seq_filename, 'r')
34     texts = [[word for word in line.split()]
35             for line in f]
36     users = []
37     for doc in texts:
38         users.append(doc[0])
39         del doc[0]
```

Necessary Packages

# TOPIC ANALYSIS – RUN TOPIC MODEL

```
Spyder (Python 2.7)
File Edit Search Source Run Debug Consoles Projects Tools View Help
D:\project\social media tutorial\Tutorial\Tutorial\
Editor - D:\project\social media tutorial\Tutorial\Tutorial\topic_analysis_src\src\models\ldaModel.py
temp.py x ldaModel.py x
1398 bad_words_list = ['RT','http','https']
1399 texts = [[word.translate(table, string.punctuation) for word in text if not any(b in word for b in bad_words_list)
1400 for text in texts]
1401 texts = [' '.join(text) for text in texts]
1402
1403 logger.info("Progress: Writing the sanitized input file")
1404 seqFileName = data_folder + 'sequence_sanitized' + '.dat'
1405 seqFile = open(seqFileName, 'w')
1406 for text in texts :
1407     seqFile.write("%s\n" % text)
1408
1409 if __name__ == "__main__":
1410     #data_folder = 'C:/research/ActivityPatternModel/twitter/output/Sandy/' changed to E:/Going abroad/UCF/Dr.
1411     data_folder = 'D:/project/social media tutorial/example data/topic model/florida bus/'
1412
1413     #raw_input_file= data_folder + 'sandy_100_times.dat'
1414     raw_input_file= data_folder + 'florida bus.txt'
1415     input_file= data_folder + 'sequence_sanitized.dat'
1416     matrix_file = data_folder + 'activity.mm'
1417     mention_matrix_file = data_folder + 'mention.mm'
1418     dic_file = data_folder + 'dictionary.dat'
1419     user_file = data_folder + 'user.dat'
1420
1421     #Run it once to create the input files
1422     sanitize_input(data_folder, raw_input_file)
1423
1424     WS, DS , US, W0, UL = create_input(data_folder, input_file)
1425
1426     #analyzeDictionary is needed only for missing activities
1427     #analyzeDictionary(dic_file)
1428
1429     #K = number of patterns
1430     runLDAModel(data_folder, matrix_file, dic_file, user_file, K=10, perplex=0) #k is the number of pattern I want
1431
1432     #runLDAModel(data_folder, matrix_file, dic_file, user_file, K=10, perplex=1)
1433
1434     #runUserPatternLDAModel(data_folder, matrix_file, dic_file, user_file, K=50, perplex=0)
1435
1436     #runCommunityUserPatternLDAModel(data_folder, matrix_file, mention_matrix_file, dic_file, user_file, K=10, perplex=0)
1437
1438     #analyzePatterns(data_folder, dic_file, user_file, num_patterns)
```

Example – Topic Analysis Model in Spyder

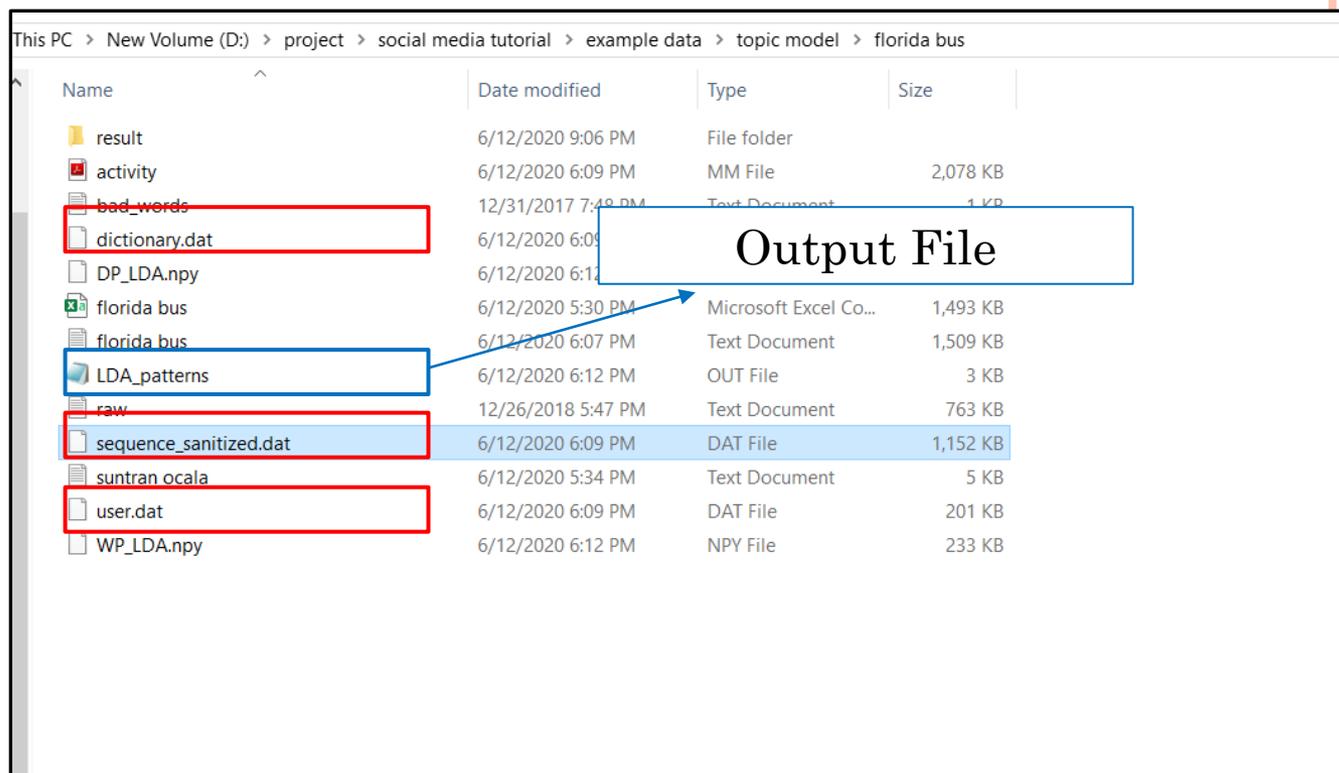
Input Folder

File Name

Number of Topics

# TOPIC ANALYSIS – RUN TOPIC MODEL

Tips: In the input file folder, there should be two necessary files – *user.dat*, *dictionary.dat* and *sequence\_sanitized.dat*, which can be seen as right figure.



# TOPIC ANALYSIS – DATA PROCESSING FOR VISUALIZATION

Input File

## 8 Data Processing for Visualization

```
[92]: df = pd.read_csv(r'D:\project\social media tutorial\example data\topic model\florida bus\LDA_patterns.out', sep = ' ', names = [
print(df)
num_class = 10
topic_num = 10

topic_list = []
words_list = []
probability_list = []
final_name = ['Topic', 'Words']
df_final = pd.DataFrame(columns=

for i in range(0, topic_num):
    topic_num = i + 1
    for j in range(2, num_class+2):
        id_num = j+i*12
        topic_name = 'Topic #' +str(topic_num)
        topic_list.append(topic_name)
        words_list.append(df.type[id_num])
        probability_list.append(df.prob[id_num])

df_final['Topic'] = topic_list
df_final['Words'] = words_list
df_final['Probability'] = probability_list

df final
```

Number of Topics

# TOPIC ANALYSIS – DATA PROCESSING FOR VISUALIZATION

```
LDA_patterns - Notepad
File Edit Format View Help

Pattern1 Prob. 0.109347

Activity Prob.
school 0.091001
There 0.073810
Florida 0.057103
bus 0.052753
week 0.050682
new 0.049716
year 0.048473
see 0.044330
ICE 0.040948
last 0.040809

Patte

Activity Prob.
tour 0.044066
bus 0.038879
MayorLevine 0.018654
The 0.017526
Tampa 0.013240
days 0.012939
LiveFromLivingRooms 0.012112
4 0.011436
Floridas 0.011135
1 0.010534

Pattern3 Prob. 0.096877
```

Input File

	A	B	C	D	E
1	Topic	Words	Probability		
2	Topic #1	amp	0.053861		
3	Topic #1	11	0.040138		
4	Topic #1	near	0.039083		
5	Topic #1	like	0.035388		
6	Topic #1	He	0.033541		
7	Topic #1	Hes	0.031166		
8	Topic #1	moving	0.030638		
9					
10					
11	Topic #1	allowed	0.030111		
12	Topic #2	bus	0.167361		
13	Topic #2	Florida	0.06021		
14	Topic #2	driver	0.042303		
15	Topic #2	one	0.01735		
16	Topic #2	WATCH	0.017056		
17	Topic #2	street	0.016175		
18	Topic #2	florida	0.015295		
19	Topic #2	hitting	0.015295		
20	Topic #2	pedestrian	0.015001		
21	Topic #2	fell	0.014708		
22	Topic #3	bus	0.163686		
23	Topic #3	Florida	0.142392		
24	Topic #3	way	0.019499		
25	Topic #3	ON	0.013719		
26	Topic #3	Im	0.012502		
27	Topic #3	video	0.011894		

Output File

# TOPIC ANALYSIS –VISUALIZATION

## 8 Topic Model Visualization ¶

```
In [93]: import csv, pdb
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from datetime import datetime
import pickle
from datetime import datetime
from matplotlib import style
import matplotlib.ticker as mticker
import matplotlib.dates as mdates
import matplotlib.cm as cm
import math

SMALL_SIZE = 12
MEDIUM_SIZE = 32
BIGGER_SIZE = 40

plt.rc('font', size=SMALL_SIZE)
plt.rc('axes', titlesize=BIGGER_SIZE)
plt.rc('axes', labelsizem= MEDIUM_SIZE)
plt.rc('xtick', labelsizem= MEDIUM_SIZE)
plt.rc('ytick', labelsizem= MEDIUM_SIZE)
plt.rc('legend', fontsize=SMALL_SIZE)
plt.rc('figure', titlesize=BIGGER_SIZE)
```

```
path_input = r"F:\topic model\tm\florida bus\result\florida bus.csv"
```

```
path_output = r"D:\project\social media tutorial\example data\topic model\florida bus\Topic_non_RT_user_heatmap_1.png"
```

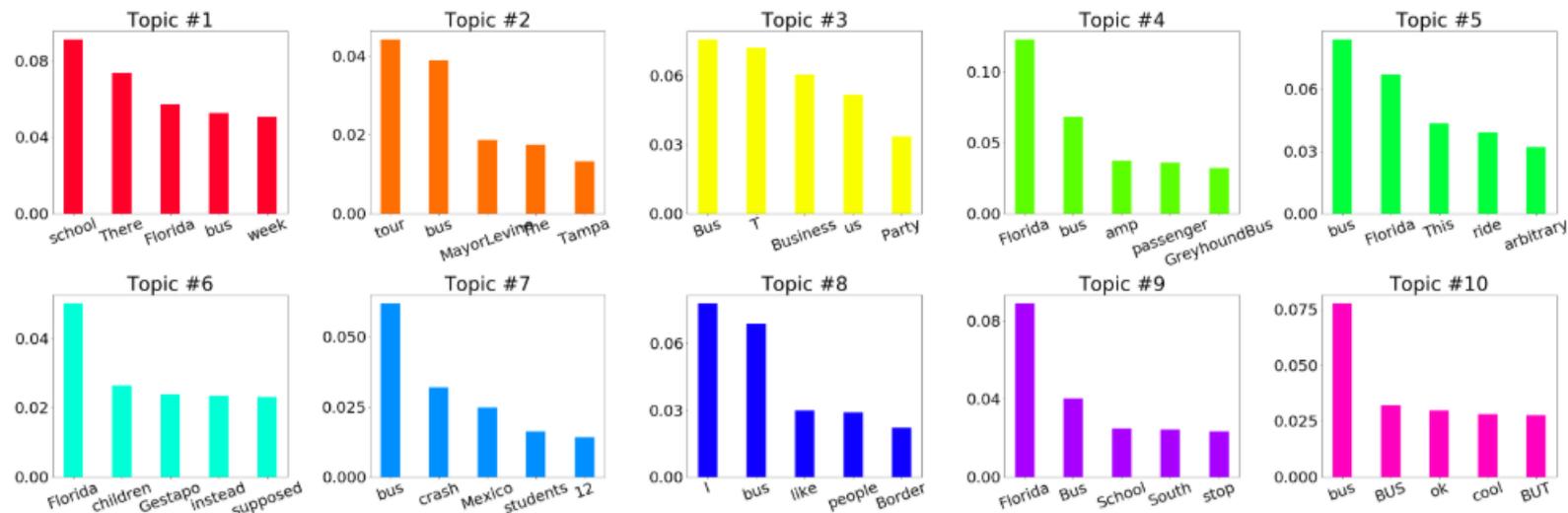
Input File

Output Figure

# TOPIC ANALYSIS –VISUALIZATION

Example of output figure

```
plt.tight_layout()  
plt.show()
```



# QUESTIONS