Text Analysis

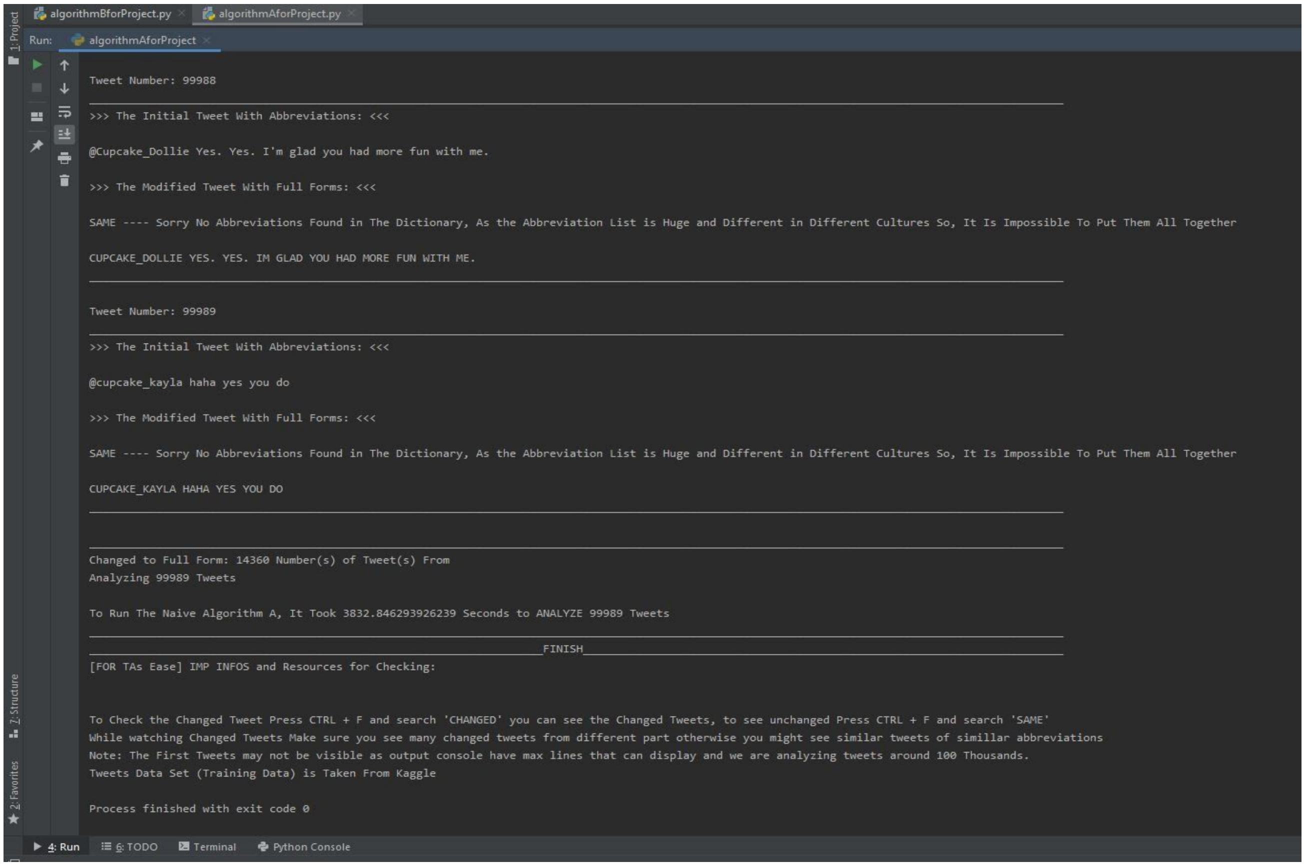
Reading Twitter Data Set and Replacing the Abbreviation (Slang words) with Full Form.

ReadMe + Instruction Guide

File Descriptions:

1. algorithmAforProject: is the python file for the naive algorithm. It took around 1 hours to run the program, as it is the most inefficient solution (mentioned on project topic pdf) and analyzing 100 thousand tweets against 235 abbreviations list.

Summary of Output in Screenshot:



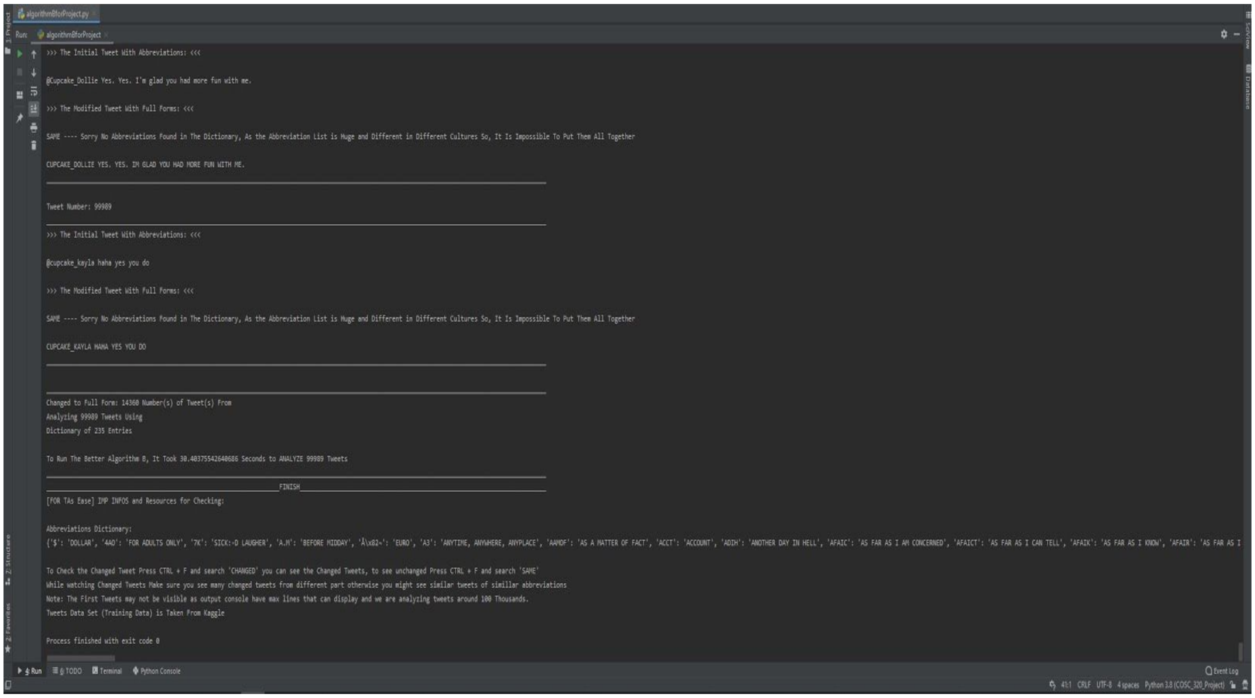
```
Tweet Number: 99988
________________________________________________________________________________________________
>>> The Initial Tweet With Abbreviations: <<<

@Cupcake_Dollie Yes. Yes. I'm glad you had more fun with me.

>>> The Modified Tweet With Full Forms: <<<

SAME ---- Sorry No Abbreviations Found in The Dictionary, As the Abbreviation List is Huge and Different in Different Cultures So, It Is Impossible To Put Them All Together

CUPCAKE_DOLLIE YES. YES. IM GLAD YOU HAD MORE FUN WITH ME.
________________________________________________________________________________________________

Tweet Number: 99989
________________________________________________________________________________________________
>>> The Initial Tweet With Abbreviations: <<<

@cupcake_kayla haha yes you do

>>> The Modified Tweet With Full Forms: <<<

SAME ---- Sorry No Abbreviations Found in The Dictionary, As the Abbreviation List is Huge and Different in Different Cultures So, It Is Impossible To Put Them All Together

CUPCAKE_KAYLA HAHA YES YOU DO
________________________________________________________________________________________________

________________________________________________________________________________________________
Changed to Full Form: 14360 Number(s) of Tweet(s) From
Analyzing 99989 Tweets

To Run The Naive Algorithm A, It Took 3832.846293926239 Seconds to ANALYZE 99989 Tweets
________________________________________________________________________________________________
_____________________________________________FINISH_____________________________________________
[FOR TAs Ease] IMP INFOS and Resources for Checking:


To Check the Changed Tweet Press CTRL + F and search 'CHANGED' you can see the Changed Tweets, to see unchanged Press CTRL + F and search 'SAME'
While watching Changed Tweets Make sure you see many changed tweets from different part otherwise you might see similar tweets of simillar abbreviations
Note: The First Tweets may not be visible as output console have max lines that can display and we are analyzing tweets around 100 Thousands.
Tweets Data Set (Training Data) is Taken From Kaggle

Process finished with exit code 0
```

* Screenshot is provided in the repository also

(Note: Running algorithmAforProject, the naive one took around 1 hour, so I would suggest run the Program when you will be taking meals or shower or watching movies)

2. <u>algorithmBforProject:</u> is the python file for the efficient algorithm. It took around 30 seconds to run the program, as it is the most efficient solution and it is analyzing 100 thousand tweets against 235 abbreviations hash: dictionary.

   Summary of Output in Screenshot:


algorithmBforProject.py
Run: algorithmBforProject
Run
6: TODO
Terminal
Python Console
Event Log
4:1 CRLF UTF-8 4 spaces Python 3.8 (COSC_320_Project)

```
>>> The Initial Tweet With Abbreviations: <<<

@Cupcake_Dollie Yes. Yes. I'm glad you had more fun with me.

>>> The Modified Tweet With Full Forms: <<<

SAME ---- Sorry No Abbreviations Found in The Dictionary, As the Abbreviation List is Huge and Different in Different Cultures So, It Is Impossible To Put Them All Together

CUPCAKE_DOLLIE YES. YES. IM GLAD YOU HAD MORE FUN WITH ME.
_______________________________________________________________________________

Tweet Number: 99989
_______________________________________________________________________________
>>> The Initial Tweet With Abbreviations: <<<

@cupcake_kayla haha yes you do

>>> The Modified Tweet With Full Forms: <<<

SAME ---- Sorry No Abbreviations Found in The Dictionary, As the Abbreviation List is Huge and Different in Different Cultures So, It Is Impossible To Put Them All Together

CUPCAKE_KAYLA HAHA YES YOU DO
_______________________________________________________________________________

_______________________________________________________________________________
Changed to Full Form: 14360 Number(s) of Tweet(s) From
Analyzing 99989 Tweets Using
Dictionary of 235 Entries

To Run The Better Algorithm B, It Took 30.40375542640686 Seconds to ANALYZE 99989 Tweets
_______________________________________________________________________________
_____________________________________FINISH____________________________________
[FOR TAs Ease] IMP INFOS and Resources for Checking:

Abbreviations Dictionary:
{'$': 'DOLLAR', '4AO': 'FOR ADULTS ONLY', '7K': 'SICK:-D LAUGHER', 'A.M': 'BEFORE MIDDAY', 'Â\x82¬': 'EURO', 'A3': 'ANYTIME, ANYWHERE, ANYPLACE', 'AAMOF': 'AS A MATTER OF FACT', 'ACCT': 'ACCOUNT', 'ADIH': 'ANOTHER DAY IN HELL', 'AFAIC': 'AS FAR AS I AM CONCERNED', 'AFAICT': 'AS FAR AS I CAN TELL', 'AFAIK': 'AS FAR AS I KNOW', 'AFAIR': 'AS FAR AS I

To Check the Changed Tweet Press CTRL + F and search 'CHANGED' you can see the Changed Tweets, to see unchanged Press CTRL + F and search 'SAME'
While watching Changed Tweets Make sure you see many changed tweets from different part otherwise you might see similar tweets of simillar abbreviations
Note: The First Tweets may not be visible as output console have max lines that can display and we are analyzing tweets around 100 Thousands.
Tweets Data Set (Training Data) is Taken From Kaggle

Process finished with exit code 0
```

* Screenshot is provided in the repository also

3. <u>train.csv:</u> is the dataset for the tweets collected from kaggle. It contains around 100,000 (100 thousands) Tweets, among 100,000 tweets 14,360 Tweet's abbreviation is updated to full form, as the abbreviations and slang list is huge and it depends on many different cultures it is impossible to collect the whole list of abbreviations/slang, managed to collect 235 nos of abbreviation/slang

4. <u>finalSlang.csv:</u> combined collection of all the dataset of the abbreviations and slang list used in the python program.

5. <u>Book1.csv:</u> initial collection of the abbreviations and slang list, not used in the program, then it is updated to more collections and changed to finalSlang.csv which is used in the program.

6. <u>Venv + .idea:</u> python build files along with the libraries.

Efficiency of The Programs:

algorithmAforProject Takes approx. 1 hours to run completely where algorithmBforProject takes 30 seconds to run completely. Both are analyzing same no. of tweets using the same no. of abbreviations. So the efficient solution algorithmBforProject is 120 times faster than the naive solution algorithmAforProject.

Instructions to Run The Program:

1. Make sure you have python installed with proper IDE.

2. Clone the repository.

3. Open the repository/python file (algorithmAforProject, algorithmBforProject) with IDE and run it

4. It will take around 1 hour to run algorithmAforProject as I mentioned above, so suggestion would be to run algorithmBforProject before.

5. See the last lines of the console after running the program to see some summary of the programs.