## Steps to ingest data:

- 1. Downloaded WinSCP Software from NYU Data Services Lab
  - https://guides.nyu.edu/software
  - https://winscp.net/eng/download.php
- 2. Install and Open installed WinSCP Application
- 3. Use IP address of dumbo to login to WinSCP
- 4. Drag and drop data from local to dumbo.
- 5. Check in the dumbo terminal for the required folder using -ls command.



- 6. Folder "project\_testing" is present which has realdonaldtrump.json, stopwords,obama.txt
- 7. Use the following commands to make folder project\_testing in hdfs and ingest this hdfs dfs -mkdir project\_testing

hdfs dfs -put project\_testing/realdonaldtrump.json project\_testing/realdonaldtrump.json hdfs dfs -put project\_testing/obama.txt project\_testing/obama.txt hdfs dfs -put project\_testing/stopwords project\_testing/stopword

## Snippet of realdonaldtrump.json Dataset:

"url":{"urls":[{"display\_url":"Instagram.com/realDonaldTrum description":{"urls":[]}, ww.Instagram.com/realDonaldTrump", "indices":[0,23], "url":"https://t.co/OKx80x7xC5"}]}}, "favourites\_count":7, "follow\_request\_sent":false, "followers\_count":61858883, "followers\_count":61858888, "followers\_count":61858888, "followers\_count\_ wing":true, "friends\_count":47, "geo\_enabled":true, "has\_extended\_profile":false, "id":25073877, "id\_str":"25073877", "is\_translation\_enabled":true, "is\_translator":false, "le ng":null, "listed\_count":105141, "location":"Washington, DC", "name":"Donald J. Trump", "notifications":false, "profile background color":"6D5C18", "profile background image url": "http://abs.twimg.com/images/themes/themel/bg.png", "profile background image url https": "https://abs.twimg.com/images/themes/themel/bg.png", "profile background til true, profile\_banner\_url": "https://pbs.twimg.com/profile\_banners/25073877/1560920145", profile\_image\_url": "http://pbs.twimg.com/profile\_images/874276197357596672/kU at00m normal.jpg", "profile image url https":"https://pbs.twlmg.com/profile images/874276197357596672/kUuht00m normal.jpg", "profile link color":"1895E0", "profile sidebar border\_color":"BDDCAD", "profile\_sidebar\_fill\_color":"CSCECO", "profile\_text\_color":"333333", "profile\_use\_background\_image":true, "protected":false, "screen\_name":"realDor aldTrump", "statuses count":42807, "time zone":null, "translator type":"regular", "url":"https://t.co/OMxB0x7xC5", "utc offset":null, "verified":true}}
("contributors":null, "coordinates":null, "created\_at":"Thu Jul 11 15:52:19 +0000 2019", "entities":("hashtags":[], "symbols":[], "urls":[("display\_url":"twitter.com/i/web/ tatus/l...", "expanded url": "https://twitter.com/i/web/status/1149345683138367491", "indices":[117,140], "url": "https://t.co/zAYNhLI8Fi"}], "user\_mentions":[]}, "favorite\_cou Tri22785, "favorited":false, "geo":null, "id":1149345681333367491, "id\_str":"1149345683133367491", "in\_reply\_to\_screen\_name":"realDonaldTrump", "in\_reply\_to\_screen\_name":"realDonaldTrump", "in\_reply\_to\_screen\_name":"realDonaldTrump", "in\_reply\_to\_screen\_name":"realDonaldTrump", "in\_reply\_to\_screen\_name":"realDonaldTrump", "in\_reply\_to\_screen\_name":"realDonaldTrump", "in\_reply\_to\_screen\_name":"realDonaldTrump", "in\_reply\_to\_screen\_name":"realDonaldTrump", "in\_reply\_to\_user\_id":25073877, "in\_reply\_to\_user\_id str":"25073877", "is\_quote\_scatus":false, "lang":"en"
"place":null, "retrieved\_uto":1562866735.0, "retweet\_count":5236, "retweeted":false, "source":"<a href=\"http://twitter.com/download/iphone\" rel=\"nofollow\">Twitter\_for\_: none</a>", "text": "...or a very nervous and skinny version of Pocahontas (1/1024th), as your President, rather than what you have now, ... https://t.co/zAYNALI8Fi", "trunc ed":true, "user":("contributors enabled":false, "created at":"Wed Mar 18 13:46:38 +0000 2009", "default profile":false, "default profile image":false, "description":"45th resident of the United States of Americaus", "entities": "description": ("urls":[]), "url": ("urls":[["display\_url": "Instagram.com/realDonaldTrump", "expanded\_url": "http:// ww.Instagram.com/realDonaldTrump", "indices":[0,23], "url":"https://t.co/OMx80x7xC5"}]}}, "favourites\_count":7, "follow\_request\_sent":false, "followers\_count":61858883, "followers\_count":61858883, "followers\_count":7, "follow\_request\_sent":false, "followers\_count":61858883, "followers\_count":7, "follow\_request\_sent":false, "followers\_count":61858883, "followers\_count":7, "follow\_request\_sent":false, "followers\_count":61858883, "followers\_count":7, "follow\_request\_sent":false, "followers\_count":7, "followers\_count.":7, "fol wing":true, "friends\_count":47, "geo\_enabled":true, "has\_extended\_profile":false, "id":25073877, "id\_str":"25073877", "is\_translation\_enabled":true, "is\_translator":false, "la ng":null, "listed count":105141, "location": "Washington, DC", "name": "Donald J. Trump", "notifications":false, "profile background color": "6D5C18", "profile background image url":"http://abs.twimg.com/images/themes/themel/bg.png","profile\_background\_image\_url\_https":"https://abs.twimg.com/images/themes/themel/bg.png","profile\_background\_til true, profile\_banner\_url": "https://pbs.twimg.com/profile\_banners/25073877/1560920145", "profile\_image\_url": "http://pbs.twimg.com/profile\_images/874276197357596672/kU:" t00m\_normal.jpg", "profile\_image\_url\_https": "https://pbs.twimg.com/profile\_images/874276197357596672/kUuht00m\_normal.jpg", "profile\_link\_color": "lB95E0", "profile\_sidebar border\_color":"BDDCAD", "profile\_sidebar\_fill\_color":"C5CECO", "profile\_text\_color":"333333", "profile\_use\_background\_image":true, "protected":false, "screen\_name":"realDoraldTrump", "statuses\_count":42807, "time\_zone":null, "translator\_type":"regular", "url":"https://t.co/OMxBOx7xC5", "utc\_offset":null, "verified":true}} [hsw268@login-2-1 ~]\$

## Snippet of all the required files in the hdfs for twitter analysis:

```
508728609 2020-05-07 12:23 project_testing/.jar
                     3 hsw268 users
rw-r--r-+
                      3 hsw268 users 379650048 2020-05-07 12:23 project_testing/BdadProject-assembly-0.1.jar._COPYING_
3 hsw268 users 8467508 2020-04-10 15:10 project_testing/WHtweets.csv
3 hsw268 users 6328881 2020-04-07 16:51 project_testing/newrealdonaldtrump.csv
rw-r--r-+
                                                         963732 2020-05-05 01:30 project_testing/obama.txt
0 2020-04-17 20:25 project_testing/obama5rdd2
3931815 2020-05-03 17:12 project_testing/obamaData
                          hsw268 users
rwxr-xr-x+
                          hsw268 users
                      3 hsw268 users
                         hsw268 users 115942426 2020-04-07 16:51 project_testing/realdonaldtrump.csv hsw268 users 15901693 2020-04-07 16:51 project_testing/realdonaldtrump.json hsw268 users 300554 2020-04-16 03:48 project_testing/sentiment140-mini.csv hsw268 users 238803811 2020-04-16 03:48 project_testing/sentiment140.csv
                      3 hsw268 users
rw-r--r-+
                      3 hsw268 users
rw-r--r-+
rw-r--r--+
                      3 hsw268 users
                          hsw268 users
                                                               2483 2020-04-09 00:26 project_testing/stopwords
```