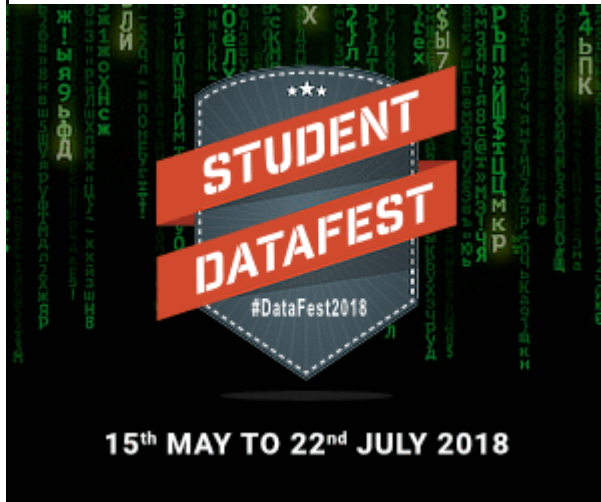


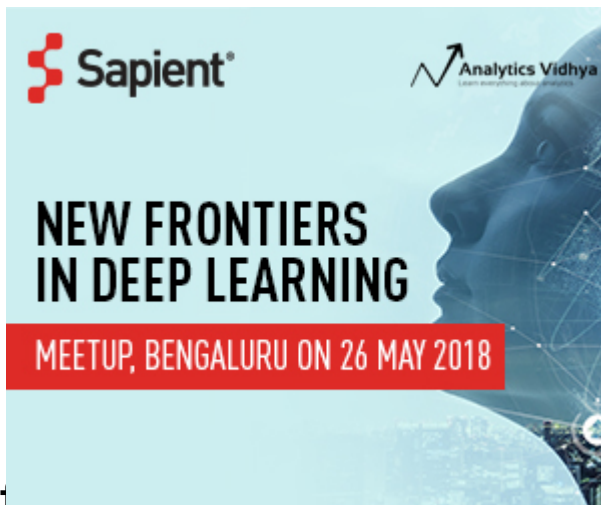
f (<https://www.facebook.com/AnalyticsVidhya>)t (<https://twitter.com/analytcsvidhya>)g+ (<https://plus.google.com/+Analyticsvidhya/posts>)

-Vidhya-Learn-everything-about-5057165)

(https://analyticsvidhya.com/student-datafest-2018/?utm_source=AVblogbottom)

Analytics Vidhya
Learn everything about analytics
[//www.analyticsvidhya.com](https://www.analyticsvidhya.com))

INNOPLEXUS™
HACKATHON: TAKE THE ML-CHALLENGE
27 MAY 2018~

Analytics (<https://www.analyticsvidhya.com/blog/category/business-analytic...>)

Perform Text Data Cleaning in

M/BLOG/CATEGORY/BUSINESS-ANALYTICS/) INFOGRAPHICS
FOGRAPHICS/) PYTHON
THON-2/)

E lyticsvidhya.com/blog/2015/06/quick-guide-text-data-cleaning-

(https://datahack.analyticsvidhya.com/contest/new-frontiers-in-deep-learning/?utm_source=AVBlogbottom) t (<https://twitter.com/home?n/&t=Quick%20Guide:%20Steps%20To%20Perform%20Text%20Data%20Cleaning%20in%20Python>)

:=Quick%20Guide:%20Steps%20To%20Perform%20Text%20Data%20Cleaning%20in%20Python+https://www.analyticsvidhya.com/blog/2015/06/quick-text-data-cleaning-python/) g+ (<https://plus.google.com/share?url=https://www.analyticsvidhya.com/blog/2015/06/quick-guide-text-data-cleaning-n/>)

n/) P (<http://pinterest.com/pin/create/button/?url=https://www.analyticsvidhya.com/blog/2015/06/quick-guide-text-data-cleaning-n/&media=https://www.analyticsvidhya.com/wp-content/uploads/2015/06/screenshot.png&description=Quick%20Guide:%20Steps%20To%20Perform%20Text%20Data%20Cleaning%20in%20Python>)

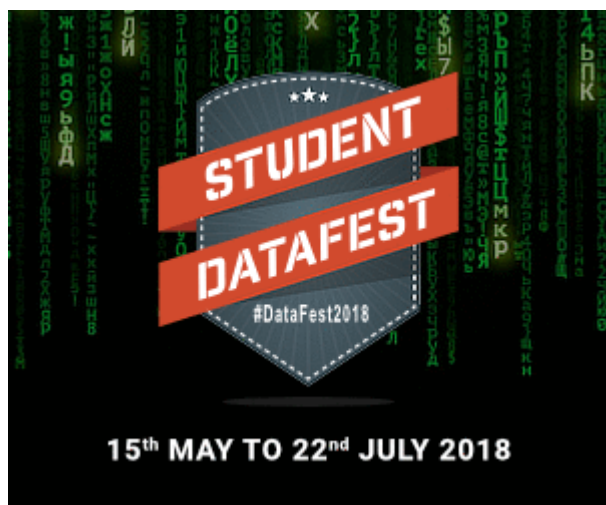
Introduction

Twitter has become an inevitable channel for brand management. It has compelled brands to become more responsive to their customers. On the other hand, the damage it would cause can't be undone. The 140 character tweets has now become a powerful tool for customers / users to directly

convey messages to brands.

For companies, these tweets carry a lot of information like sentiment, engagement, reviews and features of its products and what not. However, mining these tweets isn't easy. Why? Because, before mining a lot of cleaning. These tweets, once extracted can come with bad grammar and poor spellings – making the mining very difficult.

As the steps of cleaning this data related to tweets before mining is of Twitter, you can of course apply these methods to any other source to execute these cleaning steps.



(https://analyticsvidhya.com/student-datafest-2018/?utm_source=AV/blogbottom)



(https://datahack.analyticsvidhya.com/contest/new-frontiers-in-deep-learning/?utm_source=AV_Blogbottom)



Effective Text Data Cleaning using Python

Minining for a brand?

are brand
ing the
ed

It is used to identify the pain points of customers i.e. customer relationship management

It is widely used for predictions and forecasting

The Business Problem

Let's say, we want to find the features of an Apple iPhone which are most popular amongst the fans on Twitter.

What to do next?

We've extracted all the tweets related to consumer opinions of iPhone. Here's a sample tweet on which we'll perform data cleaning



TWEET

"I luv my <3 iphone & you're awsm apple. DisplaysAwesome, sooo happpppppy :) http://www.apple.com"

Guide for Data Cleaning

**STUDENT
DATAFEST**

#DataFest2018

15th MAY TO 22nd JULY 2018

(https://analyticsvidhya.com/student-datafest-2018/?utm_source=AVblogbottom)

ing HTML characters



```
HTMLParser.HTMLParser()  
e(original_tweet)
```

e awsm apple. Display Is Awesome, sooo
le.com"

ecoding data

STEP
02

(https://datahack.analyticsvidhya.com/contest/new-frontiers-in-deep-learning/?utm_source=AV_Blogbottom)

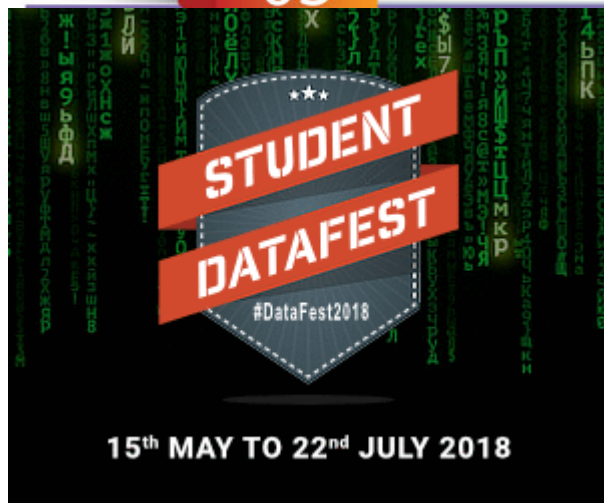
```
tweet = original_tweet.decode("utf8").encode('ascii','ignore')
```

Output

```
>> "I luv my <3 iphone & you're awsm apple. DisplayIsAwesome,  
sooo happpppppy :) http://www.apple.com"
```


STEP
03

Apostrophe Lookup



(https://analyticsvidhya.com/student-datafest-2018/?utm_source=AVblogbottom)

, "re" : " are", ...} ## Need a huge dictionary

word in APPOSTOPHES else word for word in words]

are awsm apple. DisplayIsAwesome, sooo

happy : http://www.apple.com"



NEW FRONTIERS
IN DEEP LEARNING

MEETUP, BENGALURU ON 26 MAY 2018

(https://datahack.analyticsvidhya.com/contest/new-frontiers-in-deep-learning/?utm_source=AV_Blogbottom)

Removal of Stop-Words

STEP
04

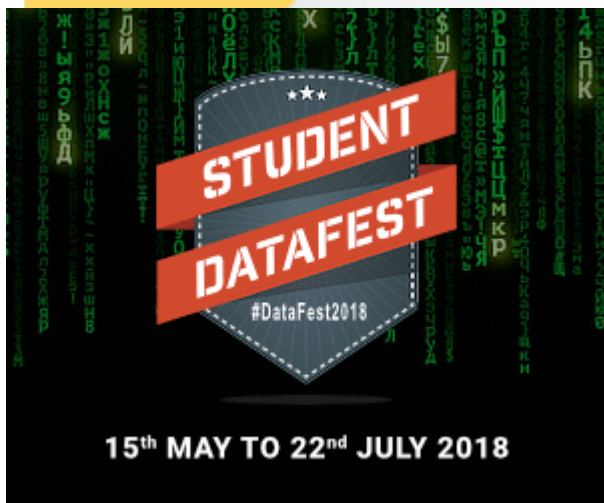
to be data driven at the word level, the
s (stop-words) should be removed.
g list of stop-words or one can use
predefined language specific libraries.

STEP
05

Removal of Punctuations

All the punctuation marks according to the priorities should be dealt with. For example: ".", ",", "?" are important punctuations that should be retained while others need to be removed.

that should be retained while others need to be removed.



(https://analyticsvidhya.com/student-datafest-2018/?utm_source=AVblogbottom)



(https://datahack.analyticsvidhya.com/contest/new-frontiers-in-deep-learning/?utm_source=AV_Blogbottom)

1 of Expressions

STEP
06

in transcripts) may contain human [Crying], [Audience paused]. These relevant to content of the speech and

Split Attached Words

`[A-Z][^A-Z]*', original_tweet))`

are awsm apple. Display Is Awesome, sooo pple.com"

Slangs lookup

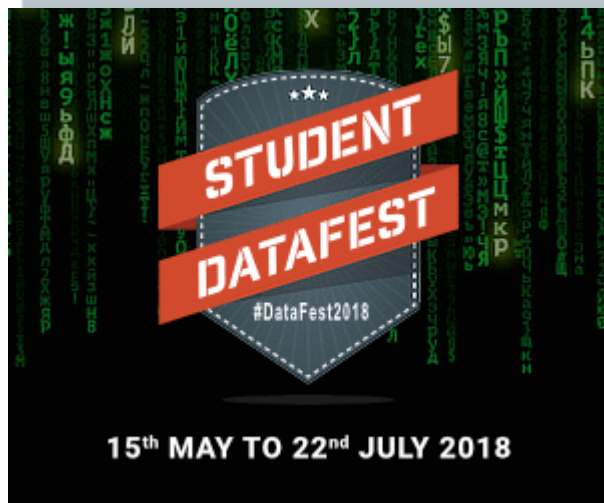
STEP
08

Code

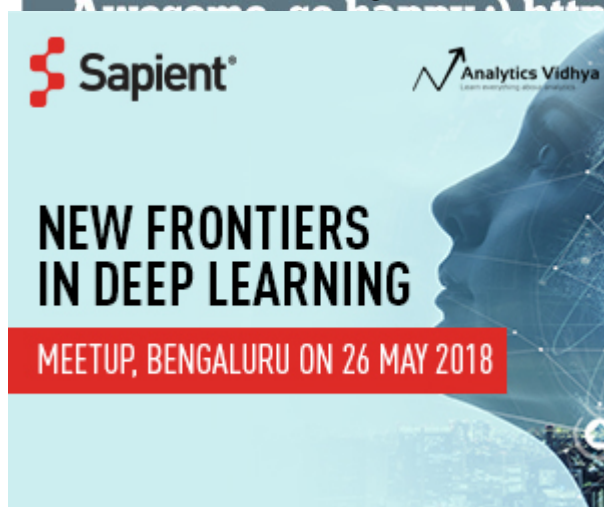
```
tweet = _slang_loopup(tweet)
```

Outcome

» **“I love my <3 iphone & you are awesome apple. Display Is Awesome, sooo happpppppy :) http://www.apple.com”**



(https://analyticsvidhya.com/student-datafest-2018/?utm_source=AVblogbottom)



(http://datahack.analyticsvidhya.com/contest/new-frontiers-in-deep-learning/?utm_source=AV_Blogbottom)

Final cleaned tweet:

» **“I love my iphone & you are awesome apple. Display Is Awesome, so happy!” , <3 , :)**

Standardizing word

for _, s in itertools.groupby(tweet))

Removal of URLs

STEP
10

at data like comments, reviews, and tweets

Advanced Data Cleaning

Grammar checking

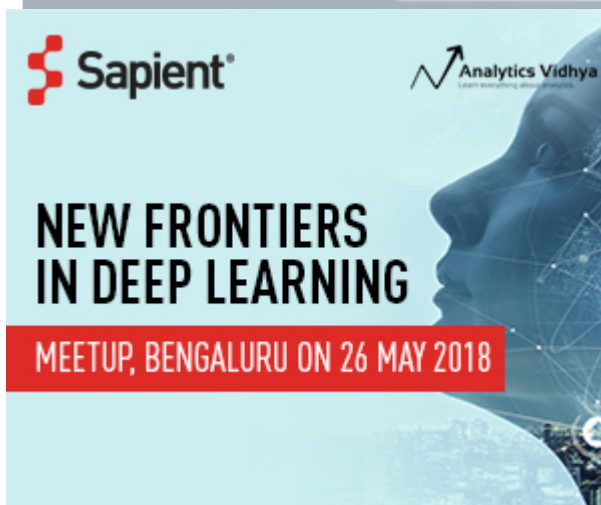
Grammar checking is majorly learning based, huge amount of proper text data is learned and models are created. Many online tools are available for grammar correction purposes.



15th MAY TO 22nd JULY 2018

(https://analyticsvidhya.com/student-datafest-2018/?utm_source=AVblogbottom)

Your Next Steps...



(https://datahack.analyticsvidhya.com/contest/new-frontiers-in-deep-learning/?utm_source=AV_Blogbottom)

4. Text Mining Hack using Google API

<http://bit.ly/1LDPF6c>

cleaned, you are ready to practice and learn the () of Text Mining-

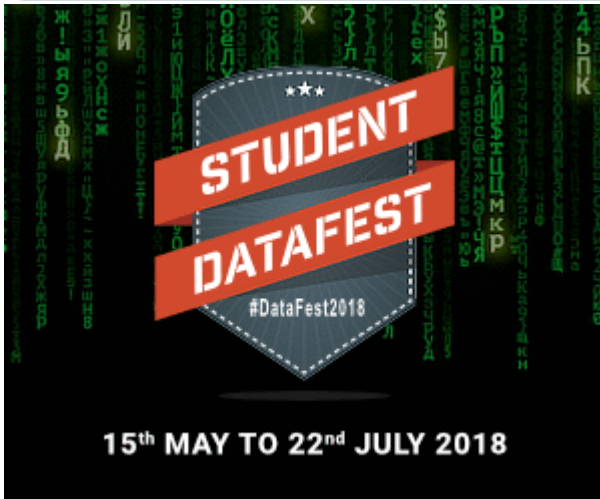
tionary for text mining

insights from free text

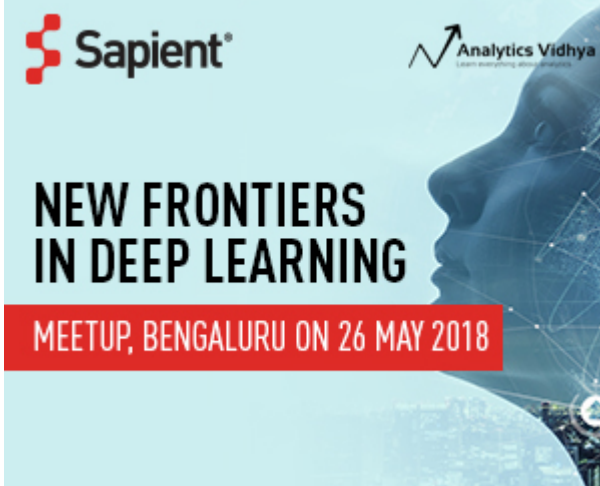


For more resources on analytics/data science, visit

analyticsvidhya.com



(https://analyticsvidhya.com/student-datafest-2018/?utm_source=AVblogbottom)



(https://datahack.analyticsvidhya.com/contest/new-frontiers-in-deep-learning/?utm_source=AV_Blogbottom)

[content/uploads/2015/06/New-Info.jpg\)](#)

and refer the python codes to perform Text Mining and follow your [s.analyticsvidhya.com/t/download-infographic-step-for-text-data-](https://analyticsvidhya.com/t/download-infographic-step-for-text-data-)

steps to perform data cleaning using python -> visit here [\(2014/11/text-data-cleaning-steps-python/\)](https://analyticsvidhya.com/2014/11/text-data-cleaning-steps-python/)

Want to continue your analytics learning, subscribe to our emails (<http://eepurl.com/800g>), follow us on twitter (<http://twitter.com/analyticsvidhya>) or like our facebook page (<http://facebook.com/analyticsvidhya>).

Analytics Vidhya's Andorid APP



[e=blog_article&utm_campaign=blog&pcampaignid=MKT-Other-1\)](#)

[k-guide-text-data-cleaning-python/?share=linkedin&nb=1\)](#)

[k-guide-text-data-cleaning-python/?share=facebook&nb=1\)](#)

318

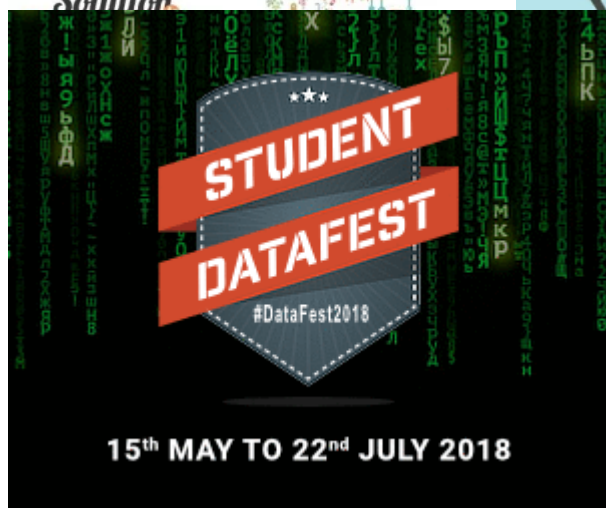
[k-guide-text-data-cleaning-python/?share=google-plus-1&nb=1\)](#)

[k-guide-text-data-cleaning-python/?share=twitter&nb=1\)](#)

[k-guide-text-data-cleaning-python/?share=pocket&nb=1\)](#)

[k-guide-text-data-cleaning-python/?share=reddit&nb=1\)](#)

RELATED



(https://analyticsvidhya.com/student-datafest-2018/?utm_source=AVblogbottom)
December 23, 2015



(<https://datahack.analyticsvidhya.com/contest/new-frontiers-in-deep-learning>)
utm_source=AVblogbottom



(<https://www.analyticsvidhya.com/blog/2016/06/exclusive-python-tutorials-talks-pycon-2016-portland-oregon/>)
November 16, 2014
In "Big data"



(<https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/>)
Steps for effective text data cleaning (with case study using Python)
(<https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/>)
November 16, 2014
In "Big data"

(<https://www.analyticsvidhya.com/blog/2016/06/exclusive-python-tutorials-talks-pycon-2016-portland-oregon/>), DATA CLEANING
(<https://www.analyticsvidhya.com/blog/tag/data-science/>), INFOGRAPHIC
INFOGRAPHICS (<https://www.analyticsvidhya.com/blog/tag/infographics/>), PYTHON
MINING (<https://www.analyticsvidhya.com/blog/tag/text-mining/>)

Next Article

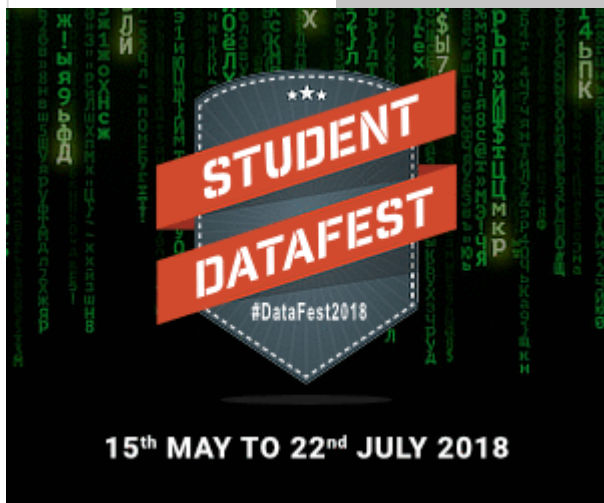
Senior Analyst – MarketIntelligent – Mumbai – (3 – 5 years of Experience)

(<https://www.analyticsvidhya.com/blog/2015/06/senior-analyst-marketintelligent-mumbai-3-5-years-experience/>)

Previous Article

Beware – interviewer for analytics job is observing you closely!

(<https://www.analyticsvidhya.com/blog/2015/06/analytics-interview-behaviour-to-avoid/>)



(https://analyticsvidhya.com/student-datafest-2018/?utm_source=AVblogbottom)



(https://datahack.analyticsvidhya.com/contest/new-frontiers-in-deep-learning/?utm_source=AV-Blogbottom)

(https://datahack.analyticsvidhya.com/contest/new-frontiers-in-deep-learning/?utm_source=AV-Blogbottom)

This article is quite old now and you might not get a prompt response from the author. We would request you to post this comment on Analytics Vidhya **Discussion portal** (<https://discuss.analyticsvidhya.com/>) to get your queries resolved.

6 COMMENTS



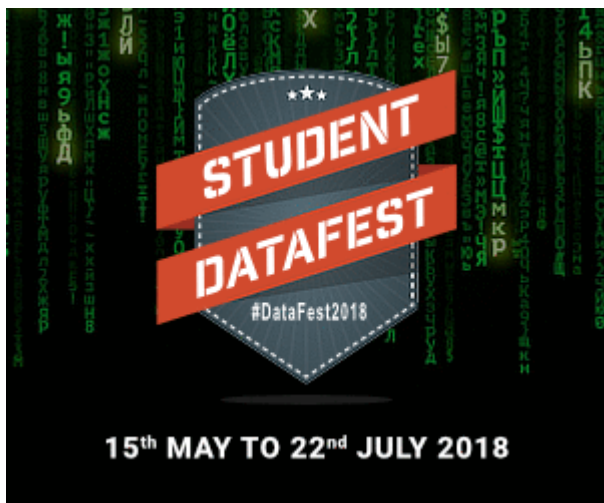
Indraneel Pise says:

REPLY



JUNE 30, 2015 AT 4:37 AM ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2015/06/QUICK-GUIDE-TEXT-DATA-CLEANING-PYTHON/#COMMENT-89497](https://www.analyticsvidhya.com/blog/2015/06/quick-guide-text-data-cleaning-python/#comment-89497))

Stemming is also an important step in text mining. You could include that too.



(https://analyticsvidhya.com/student-datafest-2018/?utm_source=AVblogbottom)

REPLY

ALYTICSVIDHYA.COM/BLOG/2015/06/QUICK-GUIDE-TEXT-DATA-CLEANING-
encodings.

REPLY

ALYTICSVIDHYA.COM/BLOG/2015/06/QUICK-GUIDE-TEXT-DATA-CLEANING-



(https://datahack.analyticsvidhya.com/contest/new-frontiers-in-deep-learning/?utm_source=AV_Blogbottom)
This is awesome, can you check the urls in the last step? They seem not working 😊

REPLY

ALYTICSVIDHYA.COM/BLOG/2015/06/QUICK-GUIDE-TEXT-DATA-CLEANING-

in social media data mining for my capstone class for my MS in
ks so much!

REPLY

ALYTICSVIDHYA.COM/BLOG/2015/06/QUICK-GUIDE-TEXT-DATA-CLEANING-



mzkarim says:

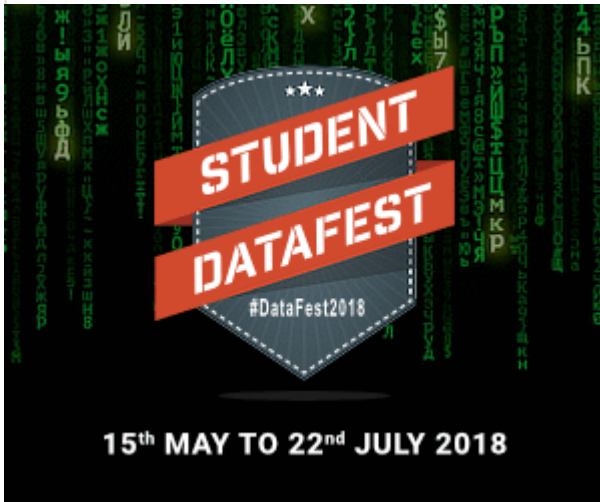
OCTOBER 30, 2015 AT 6:09 AM ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2015/06/QUICK-GUIDE-TEXT-DATA-CLEANING-PYTHON/#COMMENT-98471](https://www.analyticsvidhya.com/blog/2015/06/quick-guide-text-data-cleaning-python/#comment-98471))

Great resource. Thanks.

REPLY

LEAVE A REPLY





Your email address will not be published.



(https://analyticsvidhya.com/student-datafest-2018/?utm_source=AVblogbottom)
Email (required)



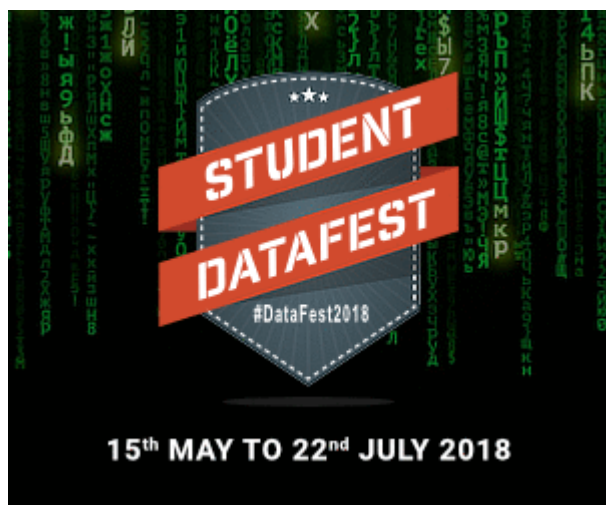
SUBMIT COMMENT

Rank	Name	Points
1	 https://datahack.analyticsvidhya.com/user/profile/Rohan_Rao	8714
2	 SRK (https://datahack.analyticsvidhya.com/user/profile/SRK)	8287
3	 aayushmnit (https://datahack.analyticsvidhya.com/user/profile/aayushmnit)	7439
4	 mark12 (https://datahack.analyticsvidhya.com/user/profile/mark12)	6269

5

sonny (<https://datahack.analyticsvidhya.com/user/profile/sonny>)

5937

[More Rankings \(http://datahack.analyticsvidhya.com/users\)](http://datahack.analyticsvidhya.com/users)


(https://analyticsvidhya.com/student-datafest-2018/?utm_source=AV_blogbottom)

A Complete Tutorial to Learn Data Science with Python from Scratch

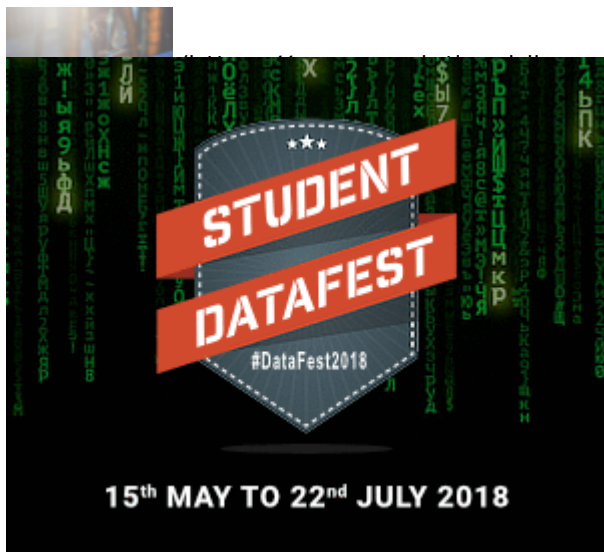


(<https://datahack.analyticsvidhya.com/contest/new-frontiers-in-deep-learning/>)

— 6 Easy Steps to Learn Naive Bayes Algorithm (with codes in Python and R)
(<https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/>)
utm_source=AV_Blogbottom)

- A Complete Tutorial on Tree Based Modeling from Scratch (in R & Python)
(<https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/>)
- A Complete Tutorial on Time Series Modeling in R
(<https://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling/>)

RECENT POSTS



(https://analyticsvidhya.com/student-datafest-2018/?utm_source=AVblogbottom)

[om/blog/2018/05/analytics-vidhya-trainings-and-datahack-](https://analyticsvidhya.com/blog/2018/05/analytics-vidhya-trainings-and-datahack-)

[dhya – Launch of Trainings and DataHack platform update!
/blog/2018/05/analytics-vidhya-trainings-and-datahack-platform-update/\)](https://analyticsvidhya.com/blog/2018/05/analytics-vidhya-trainings-and-datahack-platform-update/)

[om/blog/2018/05/deep-learning-faq/\)](https://analyticsvidhya.com/blog/2018/05/deep-learning-faq/)

12 Frequently Asked Questions on Deep Learning (with their

[analyticsvidhya.com/blog/2018/05/deep-learning-faq/\)](https://analyticsvidhya.com/blog/2018/05/deep-learning-faq/)



(<https://datahack.analyticsvidhya.com/contests/new-frontiers-in-deep-learning>)
(<https://www.analyticsvidhya.com/blog/2018/05/19-data-science-tools-for-people-dont-understand-coding/>)
MARSHAY JAIN, MAY 16, 2018
utm_source=AV_Blogbottom)

[om/blog/2018/05/announcing-datahack-summit-2018-](https://analyticsvidhya.com/blog/2018/05/announcing-datahack-summit-2018-)

[18 – Bengaluru, 22 – 24 November 2018
/blog/2018/05/announcing-datahack-summit-2018-bengaluru/\)](https://analyticsvidhya.com/blog/2018/05/announcing-datahack-summit-2018-bengaluru/)

[om/blog/2018/05/19-data-science-tools-for-people-dont-](https://analyticsvidhya.com/blog/2018/05/19-data-science-tools-for-people-dont-)



(<http://www.edvancer.in/certified-data-scientist-with-python->

[ds&utm_campaign=AVadsnonfc&utm_content=pythonavd](https://www.edvancer.in/certified-data-scientist-with-python-))

(https://analyticsvidhya.com/student-datafest-2018/?utm_source=AVblogbottom)



(https://datahack.analyticsvidhya.com/contest/new-frontiers-in-deep-learning/?utm_source=AV_Blogbottom)



FOLLOWERS


(<http://www.facebook.com/Analyticsvidhya>)



Email

SUBSCRIBE

(<http://feedburner.google.com/fb/a/mailverify?uri=analyticsvidhya>)



ANALYTICS
VIDHYA

About Us
(<http://www.analyticsvidhya.com/about>)

Hackathon
(<https://datahack.analyticsvidhya.com/>)

Discussions
(<https://discuss.analyticsvidhya.com/>)

Apply Jobs
(<https://www.analyticsvidhya.com/career-apply/>)

Leaderboard
(<https://datahack.analyticsvidhya.com/users/>)

Contact Us
(<https://www.analyticsvidhya.com/contact/>)

[analyticsvidhya.com/about](https://www.analyticsvidhya.com/about)

DATA SCIENTISTSCOMPANIES

Blog
(<https://www.analyticsvidhya.com/blog/>)

Post Jobs
(<https://www.analyticsvidhya.com/corporate/>)

Trainings
(<https://datahack.analyticsvidhya.com/>)

Hiring
(<https://discuss.analyticsvidhya.com/>)

Aids withing
(<https://www.analyticsvidhya.com/jobs/>)

Real time
(<https://www.analyticsvidhya.com/users/>)

<https://www.analyticsvidhya.com/contact/>

JOIN OUR COMMUNITY :

f

(<https://www.facebook.com/analyticsvidhya>)

16096

t

(<https://twitter.com/analyticsvidhya>)

Followers

in

(<https://www.facebook.com/analyticsvidhya>)

Followers

plus

(<https://plus.google.com/+AnalyticsVidhya>)

2830

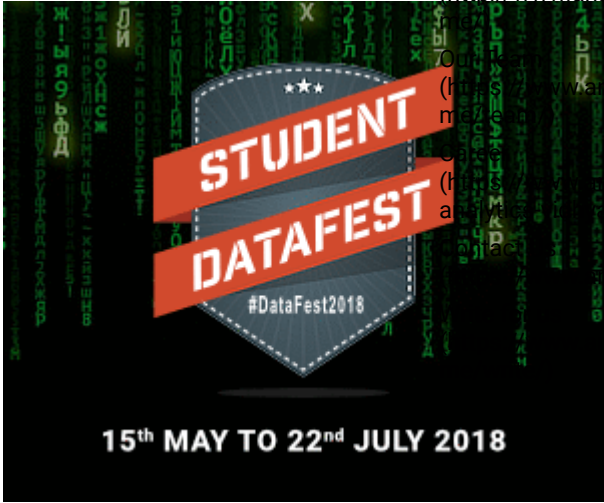
plus

(<https://plus.google.com/+AnalyticsVidhya>)


7513

Subscribe to emailer

>



(https://analyticsvidhya.com/student-datafest-2018/?utm_source=AVblogbottom)



© Copyright 2013-2018 Analytics Vidhya

Privacy Policy (<https://www.analyticsvidhya.com/privacy-policy/>)

Terms of Use (<https://www.analyticsvidhya.com/terms/>)

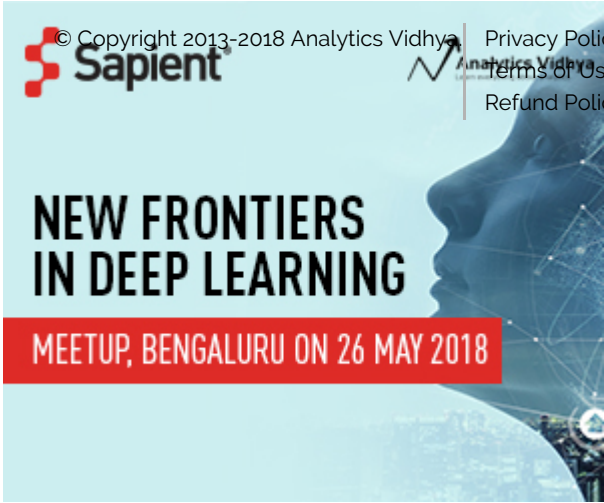
Refund Policy (<https://www.analyticsvidhya.com/refund-policy/>)

NEW FRONTIERS
IN DEEP LEARNING

MEETUP, BENGALURU ON 26 MAY 2018

Analytics Vidhya

Don't have an account? Sign up (l



(https://datahack.analyticsvidhya.com/contest/new-frontiers-in-deep-learning/?utm_source=AV_Blogbottom)

<https://www.analyticsvidhya.com/blog/2015/06/quick-guide-text-data-cleaning-python/>

16/16