×

(http://play.google.com/store/apps/details?id=com.analyticsvidhya.android)

End of Decade Sale: Flat 20% OFF on courses | Use Code: EODS20 - Enroll Today (https://courses.analyticsvidhya.com/collections/?utm\_source=flashstrip&utm\_medium=blog)

**f** (https://www.facebook.com/AnalyticsVidhya) **y** (https://twitter.com/analyticsvidhya)

G+ (https://plus.google.com/+Analyticsvidhya/posts)

in (https://in.linkedin.com/company/analytics-vidhya)

**1** LOGIN / REGISTER (HTTPS://ID.ANALYTICSVIDHYA.COM/ACCOUNTS/LOGIN/?NEXT=HTTPS:

//WWW.ANALYTICSVIDHYA.COM/BLOG/2015/06/QUICK-GUIDE-TEXT-DATA-CLEANING-PYTHON

/?UTM\_SOURCE=LINKEDIN.COM&UTM\_MEDIUM=SOCIAL)

HOME (HTTPS://WWW.ANALYTICSVIDHYA.COM)

BLOG ARCHIVE (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG-ARCHIVE/)

**DISCUSS (HTTPS://DISCUSS.ANALYTICSVIDHYA.COM)** 

CORPORATE (HTTPS://WWW.ANALYTICSVIDHYA.COM/CORPORATE/)



BLOG (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/?UTM\_SOURCE=HOME\_BLOG\_NAVBAR) ~ Q

COURSES (HTTPS://COURSES.ANALYTICSVIDHYA.COM) ~

HACKATHONS (HTTPS://DATAHACK.ANALYTICSVIDHYA.COM/CONTEST/ALL)

DSAT (HTTPS://DSAT.ANALYTICSVIDHYA.COM/?UTM\_SOURCE=HOME\_BLOG\_NAVBAR)

BOOTCAMP (HTTPS://WWW.ANALYTICSVIDHYA.COM/DATA-SCIENCE-IMMERSIVE-BOOTCAMP

/?UTM\_SOURCE=HOME\_BLOG\_NAVBAR)

CONTACT (HTTPS://WWW.ANALYTICSVIDHYA.COM/CONTACT/)

Home (https://www.analyticsvidhya.com/) > Beginner
(https://www.analyticsvidhya.com/blog/category/beginner/) > Quick Guide: Steps To
Perform Text Data Cleaning in Python (https://www.analyticsvidhya.com/blog/2015
/06/quick-guide-text-data-cleaning-python/)

BEGINNER (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/BEGINNER/)

BUSINESS ANALYTICS (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/BUSINESS-ANALYTICS/)

INFOGRAPHIC (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/INFOGRAPHIC/)

INFOGRAPHICS (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/INFOGRAPHICS/)

NLP (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/NLP/)

PYTHON\_(HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/PYTHON-2/)

TECHNIQUE (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/TECHNIQUE/)

TEXT (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/TEXT/)

UNSTRUCTURED DATA (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY

/UNSTRUCTURED-DATA/)

# **Quick Guide: Steps To Perform Text Data Cleaning in Python**

ANALYTICS VIDHYA CONTENT TEAM (HTTPS://WWW.ANALYTICSVIDHYA.CO...



(<a href="https://www.analyticsvidhya.com/data-science-immersive-bootcamp/">https://www.analyticsvidhya.com/data-science-immersive-bootcamp//>
/?utm\_source=bannerbelowblogtitle&utm\_medium=blogdisplay</a>)

#### 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 /

# RECOMMENDED READS



A Complete
Python Tutorial
to Learn Data
Science from
Scratch

(https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-learn-data-science-python-scratch-2/?utm\_source=quick-guide-text-data-cleaning-python)



Commonly
used Machine
Learning
Algorithms
(with Python
and R Codes)

(https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/vutm\_source=quick-guide-

text-data-cleaning-python)

Bootcamp 30/12/19, 8:12 am

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 you mine this data, you need to perform a lot of cleaning. These tweets, once extracted can come with unwanted html characters, bad grammar and poor spellings – making the mining very difficult.

Below is the infographic, which displays the steps of cleaning this data related to tweets before mining them. While the example in use is of Twitter, you can of course apply these methods to any text mining problem. We've used Python to execute these cleaning steps.



7 Regression Techniques you should know!

(https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/?utm\_source=quick-guide-text-data-cleaning-python)

# RECOMMENDED RESOURCES



Free Course for you
Comprehensive
Learning path to
become a data
scientist

(https://courses.analyticsvidhya.c /courses/a-comprehensivelearning-path-to-becomea-data-scientist-in-2019 /?utm\_source=Recommendation\_ utm\_medium=blog& utm\_campaign=Novlist)



Practice & Learn

Loan Prediction

(https://datahack.analyticsvidhya./contest/practice-problem-loan-prediction-iii/?utm\_source=Recommendatioutm\_medium=blog&utm\_campaign=Novlist)



### Benefits of mining for a brand?

You can do sentimental analysis to discover customer's sentiment for a brand You can measure brand popularity using the actively engaged tweeters It is used to identify the pain points of customers i.e. customer relationship management It is widely used for predictions and forecasting



## → Steps for Data Cleaning ← ←



#### Code

import HTMLParser
html\_parser = HTMLParser.HTMLParser()
tweet = html parser.unescape(original tweet)



#### Output

» "I luv my <3 iphone & you're awsm apple. Display Is Awesome, sooo happppppy http://www.apple.com"

### **Decoding data**

STEP 02

#### Code

tweet = original\_tweet.decode("utf8").encode('ascii','ignore')

#### Output

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

### **91.**

Try for free
Applied Machine
Learning - Beginner
to Professional

(https://courses.analyticsvidhya.c /courses/applied-machinelearning-beginner-toprofessional /?utm\_source=Recommendation\_ utm\_medium=blog& utm\_campaign=Novlist)

Bootcamp

(https://www.analyticsvidhya.com/wp-content/uploads/2015/06/New-Info.jpg)

Download the PDF Version of this infographic and refer the python codes to perform Text Mining and follow your 'Next Steps...' -> <u>Download Here</u> (<a href="http://discuss.analyticsvidhya.com/t/download-infographic-step-for-text-data-cleaning-using-python/1952">http://discuss.analyticsvidhya.com/t/download-infographic-step-for-text-data-cleaning-using-python/1952</a>)

To view the complete article on effective steps to perform data cleaning using python -> visit here (https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/)

If you like what you just read & want to continue your analytics learning, subscribe to our emails (http://feedburner.google.com/fb/a/mailverify?uri=analyticsvidhya), follow us on twitter (http://twitter.com/analyticsvidhya) or like our facebook page (http://facebook.com/analyticsvidhya).

You can also read this article on Analytics Vidhya's Android APP



(//play.google.com/store

/apps/details?id=com.analyticsvidhya.android& utm\_source=blog\_article&utm\_campaign=blog&pcampaignid=MKT-Other-global-all-co-prtnr-py-PartBadge-Mar2515-1)

#### Share this:

- (https://www.analyticsvidhya.com/blog/2015/06/quick-guide-text-data-cleaning-python/?share=linkedin&nb=1)
- [1] (https://www.analyticsvidhya.com/blog/2015/06/quick-guide-text-data-cleaning-python/?share=facebook&nb=1)
- (https://www.analyticsvidhya.com/blog/2015/06/quick-guide-text-data-cleaning-python/?share=twitter&nb=1)
- (https://www.analyticsvidhya.com/blog/2015/06/quick-guide-text-data-cleaning-python/?share=pocket&nb=1)
- (https://www.analyticsvidhya.com/blog/2015/06/quick-guide-text-data-cleaning-python/?share=reddit&nb=1)

Like this:

Loading...

TAGS: ADVANCED DATA CLEANING (HTTPS://WWW.ANALYTICSVIDHYA.COM /BLOG/TAG/ADVANCED-DATA-CLEANING/), DATA CLEANING (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/DATA-CLEANING/), DATA SCIENCE (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/DATA-SCIENCE/), INFOGRAPHIC (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG /INFOGRAPHIC/), INFOGRAPHICS (HTTPS://WWW.ANALYTICSVIDHYA.COM /BLOG/TAG/INFOGRAPHICS/), PYTHON (HTTPS://WWW.ANALYTICSVIDHYA.COM /BLOG/TAG/PYTHON/), TEXT MINING (HTTPS://WWW.ANALYTICSVIDHYA.COM /BLOG/TAG/TEXT-MINING/)

PREVIOUS ARTICLE

Beware interviewer for analytics job is observing you closely!

> (https://www.analyticsvidhya.com /blog/2015/06/analyticsinterview-behaviour-toavoid/)

**NEXT ARTICLE** 

**Senior Analyst -**Marketelligent -Mumbai - (3 - 5 years of **Experience**)

(https://www.analyticsvidhya.com /blog/2015/06/senioranalyst-marketelligentmumbai-3-5-yearsexperience/)



Analytics Vidhya Content Team

(<u>Https://www.analyticsvidhya.com/blog/author</u>(<u>https://www.analyticsvidhya.com</u>

/avcontentteam/)

/blog/author

/avcontentteam/)

Analytics Vidhya Content team

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

#### **6 COMMENTS**



#### **INDRANEEL PISE**

<u>Reply</u>

<u>June 30, 2015 at 4:37 am (https://www.analyticsvidhya.com/blog/2015</u>/06/quick-guide-text-data-cleaning-python/#comment-89497)

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



**ROHIT** 

Reply

<u>June 30, 2015 at 12:22 pm (https://www.analyticsvidhya.com/blog/2015 /06/quick-guide-text-data-cleaning-python/#comment-89514)</u>

remember to deal with character encodings.



#### **ROHIT SHETTY**

Reply

<u>June 30, 2015 at 12:25 pm (https://www.analyticsvidhya.com/blog/2015 /06/quick-guide-text-data-cleaning-python/#comment-89515)</u>

Thou r awwssomee



#### **TERESA ROTHAAR**

Reply

<u>June 30, 2015 at 2:49 pm (https://www.analyticsvidhya.com/blog/2015 /06/quick-guide-text-data-cleaning-python/#comment-89519)</u>

This is great. I'm doing a project on social media data mining for my capstone class for my MS in MIS, and this is very helpful. Thanks so much!



**WHALY** 

<u>Reply</u>

July 20, 2015 at 1:35 am (https://www.analyticsvidhya.com/blog/2015

Bootcamp 30/12/19, 8:12 am

/06/quick-guide-text-data-cleaning-python/#comment-90768)



#### **MZKARIM**

<u>Reply</u>

October 30, 2015 at 6:09 am (https://www.analyticsvidhya.com/blog/2015/06/quick-guide-text-data-cleaning-python/#comment-98471)

Great resource. Thanks.



(https://www.analyticsvidhya.com/)

### Download App



(https://play.google.com/store

/apps/details?id=com.analyticsvidhya.android)



(https://apps.apple.com /us/app/analyticsvidhya/id1470025572)

### **Analytics Vidhya**

#### **Data Science**

About Us (https://www.analyticsvidhya.com Blog (https://www.analyticsvidhya.com

/about-me/) /blog/)

Our Team Hackathon

(https://www.analyticsvidhya.com/about- (https://datahack.analyticsvidhya.com/)

me/team/) Discussions

Careers (https://www.analyticsvidhya.com (https://discuss.analyticsvidhya.com/)

/about-me/career-analytics-vidhya/) Apply Jobs

Contact us (https://www.analyticsvidhya.com/jobs/)

(https://www.analyticsvidhya.com

/contact/)

**Companies** 

Post Jobs

(https://www.analyticsvidhya.com

/corporate/)

**Trainings** 

(https://courses.analyticsvidhya.com/)

Hiring Hackathons

(https://datahack.analyticsvidhya.com/)

Advertising

(https://www.analyticsvidhya.com

/contact/)

#### Visit us

in

(https://www.linkedin.com

/comp@mtyps://www.youtube.com

(https:///annarlyntfesebo@httpsn//twitter.com

/Analytictlyidl/NAAH6@Dtelytld4higthy2343iObA)

© Copyright 2013-2020 Analytics Vidhya

Privacy Policy Terms of Use Refund Policy

(https://www.addtoany.com/add\_to/facebook?linkurl=https%3A%2F

%2Fwww.analyticsvidhya.com%2Fblog%2F2015%2F06%2Fquick-guide-text-data-cleaning-python

%2F%3Futm\_source%3Dlinkedin.com%26utm\_medium%3Dsocial&linkname=Python%20Infographic%20%7C

%20Steps%20To%20Clean%20Tweets%20Before%20Mining%20Them%20In%20Python)

(https://www.addtoany.com/add\_to/twitter?linkurl=https%3A%2F

%2Fwww.analyticsvidhya.com%2Fblog%2F2015%2F06%2Fquick-guide-text-data-cleaning-python

%2F%3Futm\_source%3Dlinkedin.com%26utm\_medium%3Dsocial&linkname=Python%20Infographic%20%7C

%20Steps%20To%20Clean%20Tweets%20Before%20Mining%20Them%20In%20Python)

(https://www.addtoany.com/add\_to/linkedin?linkurl=https%3A%2F

%2Fwww.analyticsvidhya.com%2Fblog%2F2015%2F06%2Fquick-guide-text-data-cleaning-python Bootcamp

30/12/19, 8:12 am

```
%2F%3Futm_source%3Dlinkedin.com%26utm_medium%3Dsocial&linkname=Python%20Infographic%20%7C
%20Steps%20To%20Clean%20Tweets%20Before%20Mining%20Them%20In%20Python)
(https://www.addtoany.com/add_to/flipboard?linkurl=https%3A%2F
%2Fwww.analyticsvidhya.com%2Fblog%2F2015%2F06%2Fguick-guide-text-data-cleaning-python
%2F%3Futm_source%3Dlinkedin.com%26utm_medium%3Dsocial&linkname=Python%20Infographic%20%7C
%20Steps%20To%20Clean%20Tweets%20Before%20Mining%20Them%20In%20Python)
(https://www.addtoany.com/add_to/whatsapp?linkurl=https%3A%2F
%2Fwww.analyticsvidhya.com%2Fblog%2F2015%2F06%2Fguick-guide-text-data-cleaning-python
%2F%3Futm_source%3Dlinkedin.com%26utm_medium%3Dsocial&linkname=Python%20Infographic%20%7C
%20Steps%20To%20Clean%20Tweets%20Before%20Mining%20Them%20In%20Python)
(https://www.addtoany.com/add_to/facebook?linkurl=https%3A%2F
%2Fwww.analyticsvidhya.com%2Fblog%2F2015%2F06%2Fquick-guide-text-data-cleaning-python
%2F%3Futm_source%3Dlinkedin.com%26utm_medium%3Dsocial&linkname=Python%20Infographic%20%7C
%20Steps%20To%20Clean%20Tweets%20Before%20Mining%20Them%20In%20Python)
(https://www.addtoany.com/add_to/twitter?linkurl=https%3A%2F
%2Fwww.analyticsvidhya.com%2Fblog%2F2015%2F06%2Fguick-guide-text-data-cleaning-python
%2F%3Futm_source%3Dlinkedin.com%26utm_medium%3Dsocial&linkname=Python%20Infographic%20%7C
%20Steps%20To%20Clean%20Tweets%20Before%20Mining%20Them%20In%20Python)
(https://www.addtoany.com/add_to/linkedin?linkurl=https%3A%2F
%2Fwww.analyticsvidhya.com%2Fblog%2F2015%2F06%2Fguick-guide-text-data-cleaning-python
%2F%3Futm_source%3Dlinkedin.com%26utm_medium%3Dsocial&linkname=Python%20Infographic%20%7C
%20Steps%20To%20Clean%20Tweets%20Before%20Mining%20Them%20In%20Python)
(https://www.addtoany.com/add_to/flipboard?linkurl=https%3A%2F
%2Fwww.analyticsvidhya.com%2Fblog%2F2015%2F06%2Fquick-guide-text-data-cleaning-python
%2F%3Futm_source%3Dlinkedin.com%26utm_medium%3Dsocial&linkname=Python%20Infographic%20%7C
%20Steps%20To%20Clean%20Tweets%20Before%20Mining%20Them%20In%20Python)
(https://www.addtoany.com/add_to/whatsapp?linkurl=https%3A%2F
%2Fwww.analyticsvidhya.com%2Fblog%2F2015%2F06%2Fquick-guide-text-data-cleaning-python
%2F%3Futm_source%3Dlinkedin.com%26utm_medium%3Dsocial&linkname=Python%20Infographic%20%7C
%20Steps%20To%20Clean%20Tweets%20Before%20Mining%20Them%20In%20Python)
(https://www.addtoany.com/add_to/facebook?linkurl=https%3A%2F
%2Fwww.analyticsvidhya.com%2Fblog%2F2015%2F06%2Fquick-guide-text-data-cleaning-python
%2F%3Futm_source%3Dlinkedin.com%26utm_medium%3Dsocial&linkname=Python%20Infographic%20%7C
%20Steps%20To%20Clean%20Tweets%20Before%20Mining%20Them%20In%20Python)
(https://www.addtoany.com/add_to/twitter?linkurl=https%3A%2F
%2Fwww.analyticsvidhya.com%2Fblog%2F2015%2F06%2Fquick-guide-text-data-cleaning-python
%2F%3Futm_source%3Dlinkedin.com%26utm_medium%3Dsocial&linkname=Python%20Infographic%20%7C
%20Steps%20To%20Clean%20Tweets%20Before%20Mining%20Them%20In%20Python)
(https://www.addtoany.com/add_to/linkedin?linkurl=https%3A%2F
%2Fwww.analyticsvidhya.com%2Fblog%2F2015%2F06%2Fquick-guide-text-data-cleaning-python
%2F%3Futm_source%3Dlinkedin.com%26utm_medium%3Dsocial&linkname=Python%20Infographic%20%7C
%20Steps%20To%20Clean%20Tweets%20Before%20Mining%20Them%20In%20Python)
(https://www.addtoany.com/add_to/flipboard?linkurl=https%3A%2F
                                                                               Bootcamp
10 of 11
                                                                                30/12/19, 8:12 am
```

%2Fwww.analyticsvidhya.com%2Fblog%2F2015%2F06%2Fquick-guide-text-data-cleaning-python %2F%3Futm\_source%3Dlinkedin.com%26utm\_medium%3Dsocial&linkname=Python%20Infographic%20%7C %20Steps%20To%20Clean%20Tweets%20Before%20Mining%20Them%20In%20Python) (https://www.addtoany.com/add\_to/whatsapp?linkurl=https%3A%2F %2Fwww.analyticsvidhya.com%2Fblog%2F2015%2F06%2Fquick-guide-text-data-cleaning-python %2F%3Futm\_source%3Dlinkedin.com%26utm\_medium%3Dsocial&linkname=Python%20Infographic%20%7C %20Steps%20To%20Clean%20Tweets%20Before%20Mining%20Them%20In%20Python) (https://www.addtoany.com/add\_to/facebook?linkurl=https%3A%2F %2Fwww.analyticsvidhya.com%2Fblog%2F2015%2F06%2Fguick-guide-text-data-cleaning-python %2F%3Futm\_source%3Dlinkedin.com%26utm\_medium%3Dsocial&linkname=Python%20Infographic%20%7C %20Steps%20To%20Clean%20Tweets%20Before%20Mining%20Them%20In%20Python) (https://www.addtoany.com/add\_to/twitter?linkurl=https%3A%2F %2Fwww.analyticsvidhya.com%2Fblog%2F2015%2F06%2Fquick-guide-text-data-cleaning-python %2F%3Futm\_source%3Dlinkedin.com%26utm\_medium%3Dsocial&linkname=Python%20Infographic%20%7C %20Steps%20To%20Clean%20Tweets%20Before%20Mining%20Them%20In%20Python) (https://www.addtoany.com/add\_to/linkedin?linkurl=https%3A%2F %2Fwww.analyticsvidhya.com%2Fblog%2F2015%2F06%2Fquick-guide-text-data-cleaning-python %2F%3Futm\_source%3Dlinkedin.com%26utm\_medium%3Dsocial&linkname=Python%20Infographic%20%7C %20Steps%20To%20Clean%20Tweets%20Before%20Mining%20Them%20In%20Python) (https://www.addtoany.com/add\_to/flipboard?linkurl=https%3A%2F %2Fwww.analyticsvidhya.com%2Fblog%2F2015%2F06%2Fquick-guide-text-data-cleaning-python %2F%3Futm\_source%3Dlinkedin.com%26utm\_medium%3Dsocial&linkname=Python%20Infographic%20%7C

%20Steps%20To%20Clean%20Tweets%20Before%20Mining%20Them%20In%20Python)

%20Steps%20To%20Clean%20Tweets%20Before%20Mining%20Them%20In%20Python)

%2Fwww.analyticsvidhya.com%2Fblog%2F2015%2F06%2Fguick-guide-text-data-cleaning-python

%2F%3Futm\_source%3Dlinkedin.com%26utm\_medium%3Dsocial&linkname=Python%20Infographic%20%7C

(https://www.addtoany.com/add\_to/whatsapp?linkurl=https%3A%2F