

f (<https://www.facebook.com/AnalyticsVidhya>)

t (<https://twitter.com/analytcsvidhya>)

g+ (<https://plus.google.com/+Analyticsvidhya/posts>)

in (<https://www.linkedin.com/groups/Analytics-Vidhya-Learn-everything-about-5057165>)



Analytics Vidhya
Learn everything about analytics

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



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

Home (<https://www.analyticsvidhya.com/>) > Big data (<https://www.analyticsvidhya.com/blog/category/big-data/>) > Steps for effective

Steps for effective text data cleaning (with case study using Python)

BIG DATA (<https://www.analyticsvidhya.com/blog/category/big-data/>) BUSINESS ANALYTICS

(<https://www.analyticsvidhya.com/blog/category/business-analytics/>) PYTHON

(<https://www.analyticsvidhya.com/blog/category/python-2/>)

<https://www.facebook.com/sharer.php?u=https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-for-effective-text-data-cleaning-with-case-study-using-python>) t (<https://twitter.com/home?for=effective-text-data-cleaning-with-case-study-using-python+https://www.analyticsvidhya.com/blog/2014/11/text-eps-python/>) g+ (<https://plus.google.com/share?url=https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/>) p .com/pin/create/button/?url=https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-&description=Steps%20for%20effective%20text%20data%20cleaning%20(with%20case%20study%20using%20Python))



Get Certified by TECH MAHINDRA

REGISTER FOR FREE INFO-SESSION

(http://events.upxacademy.com/infosession?utm_source=MLWeek-

The days when one would get data in tabulated spreadsheets are truly behind us. A moment of silence for the data residing in the spreadsheet pockets. Today, more than 80% of the data is unstructured – it is either present in data silos or scattered around the digital archives. Data is being produced as we speak – from every conversation we make in the social media to every content generated from news sources. In order to produce any meaningful actionable insight from data, it is important to know how to work with it in its unstructured form. As a Data Scientist at one of the fastest growing Decision Sciences firm, my bread and butter comes from deriving meaningful insights from unstructured text information.



(https://www.analyticsvidhya.com/blog/wp-content/uploads/2014/11/Mining_Twitter_Data.jpg)

One of the first steps in working with text data is to pre-process it. It is an essential step before the data is ready for analysis. Majority of available text data is highly unstructured and noisy in nature – to achieve better insights or to build better algorithms, it is necessary to play with clean data. For example, social media data is highly unstructured – it is an informal communication – typos, bad grammar, usage of slang, presence of unwanted content like URLs, Stopwords, Expressions etc. are the usual suspects.

In this blog, therefore I discuss about these possible noise elements and how you could clean them step by step. I am providing ways to clean data using Python.

As a typical business problem, assume you are interested in finding: which are the features of an iPhone which are more popular among the fans. You have extracted consumer opinions related to iPhone and here is a tweet you extracted:

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

Steps for data cleaning:

Here is what you do:

1. **Escaping HTML characters:** Data obtained from web usually contains a lot of html entities like < > & which gets embedded in the original data. It is thus necessary to get rid of these entities. One approach is to directly remove them by the use of specific regular expressions. Another approach is to use appropriate packages and modules (for example htmlparser of Python), which can convert these entities to standard html tags. For example: < is converted to "<" and & is converted to "&".

Snippet:

```
import HTMLParser
```

```
html_parser = HTMLParser.HTMLParser()
```

```
tweet = html_parser.unescape(original_tweet)
```

Output:

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

2. **Decoding data:** This is the process of transforming information from complex symbols to simple and easier to understand characters. Text data may be subject to different forms of decoding like "Latin", "UTF8" etc. Therefore, for better analysis, it is necessary to keep the complete data in standard encoding format. UTF-8 encoding is widely accepted and is recommended to use.

Snippet:

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

Output:

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

3. **Apostrophe Lookup:** To avoid any word sense disambiguation in text, it is recommended to maintain proper structure in it and to abide by the rules of context free grammar. When apostrophes are used, chances of disambiguation increases.

For example "it's is a contraction for it is or it has".

All the apostrophes should be converted into standard lexicons. One can use a lookup table of all possible keys to get rid of disambiguates.

Snippet:

```
APPOSTOPHES = {"'s" : " is", "'re" : " are", ...} ## Need a huge dictionary
```

```
words = tweet.split()
```

```
reformed = [APPOSTOPHES[word] if word in APPOSTOPHES else word for word in words]
```

```
reformed = " ".join(reformed)
```

Outcome:

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

4. **Removal of Stop-words:** When data analysis needs to be data driven at the word level, the commonly occurring words (stop-words) should be removed. One can either create a long list of stop-words or one can use predefined language specific libraries.
5. **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.
6. **Removal of Expressions:** Textual data (usually speech transcripts) may contain human expressions like [laughing], [Crying], [Audience paused]. These expressions are usually non relevant to content of the speech and hence need to be removed. Simple regular expression can be useful in this case.
7. **Split Attached Words:** We humans in the social forums generate text data, which is completely informal in nature. Most of the tweets are accompanied with multiple attached words like RainyDay, PlayingInTheCold etc. These entities can be split into their normal forms using simple rules and regex.

Snippet:

```
cleaned = " ".join(re.findall('[A-Z][^A-Z]*', original_tweet))
```

Outcome:

```
>> "I luv my <3 iphone & you are awsm apple. Display Is Awesome, sooo happpppppy 😊  
http://www.apple.com"
```

8. **Slangs lookup:** Again, social media comprises of a majority of slang words. These words should be transformed into standard words to make free text. The words like luv will be converted to love, Helo to Hello. The similar approach of apostrophe look up can be used to convert slangs to standard words. A number of sources are available on the web, which provides lists of all possible slangs, this would be your holy grail and you could use them as lookup dictionaries for conversion purposes.

Snippet:

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

Outcome:

```
>> "I love my <3 iphone & you are awesome apple. Display Is Awesome, sooo happpppppy 😊  
http://www.apple.com"
```

9. **Standardizing words:** Sometimes words are not in proper formats. For example: "I looooveeee you" should be "I love you". Simple rules and regular expressions can help solve these cases.

Snippet:

```
tweet = ''.join(''.join(s)[:2] for _, s in itertools.groupby(tweet))
```

Outcome:

```
>> "I love my <3 iphone & you are awesome apple. Display Is Awesome, so happy 😊  
http://www.apple.com"
```

10. **Removal of URLs:** URLs and hyperlinks in text data like comments, reviews, and tweets should be removed.

Final cleaned tweet:

```
>> "I love my iphone & you are awesome apple. Display Is Awesome, so happy!" , <3 , 😊
```

Advanced data cleaning:

1. **Grammar checking:** Grammar checking is majorly learning based, huge amount of proper text data is learned and models are created for the purpose of grammar correction. There are many online tools that are available for grammar correction purposes.
2. **Spelling correction:** In natural language, misspelled errors are encountered. Companies like Google and Microsoft have achieved a decent accuracy level in automated spell correction. One can use algorithms like the Levenshtein Distances, Dictionary Lookup etc. or other modules and packages to fix these errors.

End Notes:

Hope you found this article helpful. These were some tips and tricks, I have learnt while working with a lot of text data. If you follow the above steps to clean the data, you can drastically improve the accuracy of your results and draw better insights. Do share your views/doubts in the comments section and I would be happy to participate.

Go Hack 😊

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>).

Share this:

 (<https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/?share=linkedin&nb=1>) 179

 (<https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/?share=facebook&nb=1>) 66

 (<https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/?share=google-plus-1&nb=1>)

 (<https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/?share=twitter&nb=1>)

 (<https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/?share=pocket&nb=1>)

 (<https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/?share=reddit&nb=1>)

RELATED

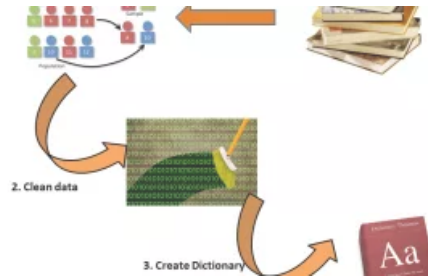


(<https://www.analyticsvidhya.com/blog/2016/06/exclusive-python-tutorials-talks-pycon-2016-portland-oregon/>)

Exclusive Python Tutorials & Talks from PyCon 2016 Portland, Oregon (<https://www.analyticsvidhya.com/blog/2016/06/exclusive-python-tutorials-talks-pycon-2016-portland-oregon/>)

June 7, 2016

In "Big data"



(<https://www.analyticsvidhya.com/blog/2014/08/step-step-guide-extract-information-free-text-unstructured-data/>)

Step by step guide to extract insights from free text (unstructured data) (<https://www.analyticsvidhya.com/blog/2014/08/step-step-guide-extract-information-free-text-unstructured-data/>)

August 19, 2014

In "Big data"

Data Science Consultant & Text Mining Gurgaon/Pune (3 to 7 years of Experience) gurgaonpune-3-to-7-years-of-experience/) (<https://www.analyticsvidhya.com/blog/2016/12/data-science-consultant-text-mining-gurgaonpune-3-to-7-years-of-experience/>)

Experience : 3 - 7 years

Requirements : strong understanding of Natural Language Processing Task Info : Pioneers in sales force and marketing analytics December 27, 2016 In "Jobs"

TAGS: HTMLPARSER ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/HTMLPARSER/](https://www.analyticsvidhya.com/blog/tag/htmlparser/)), PYTHON ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/PYTHON/](https://www.analyticsvidhya.com/blog/tag/python/)), SOCIAL MEDIA ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/SOCIAL-MEDIA/](https://www.analyticsvidhya.com/blog/tag/social-media/)), TEXT CLEANING ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/TEXT-CLEANING/](https://www.analyticsvidhya.com/blog/tag/text-cleaning/)), TEXT MINING ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/TEXT-MINING/](https://www.analyticsvidhya.com/blog/tag/text-mining/)), TWITTER DATA ANALYSIS ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/TWITTER-DATA-ANALYSIS/](https://www.analyticsvidhya.com/blog/tag/twitter-data-analysis/))

Next Article

Analytics Engineer – Inside view – India (2-4 years of experience)
(<https://www.analyticsvidhya.com/blog/2014/11/analytics-engineer-view-india-2-4-years-experience/>)

Previous Article

Data Scientist – Numerify – Bangalore (3-5 Years of experience)

(<https://www.analyticsvidhya.com/blog/2014/11/data-scientist-numerify-bangalore-3-5-years-experience/>)



(<https://www.analyticsvidhya.com/blog/author/shivam5992/>)

Author

Shivam Bansal (<https://www.analyticsvidhya.com/blog/author/shivam5992/>)

Shivam Bansal is a data scientist with exhaustive experience in Natural Language Processing and Machine Learning in several domains. He is passionate about learning and always looks forward to solve challenging analytical problems.

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.

13 COMMENTS



Manikandan T V says:

REPLY (<https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/?replytocom=42651#respond>)
NOVEMBER 17, 2014 AT 10:47 AM (<https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/#comment-42651>)

Can All this be done in R?



Shivam Bansal says:

REPLY (<https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/?replytocom=42732#respond>)
NOVEMBER 17, 2014 AT 5:28 PM (<https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/#comment-42732>)

Yes !!

Packages like "tm" (text mining) provides support for many of these functions. Rest of them can be explicitly written in R.



himani says:

REPLY (<https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/?replytocom=109078#respond>)
APRIL 6, 2016 AT 9:19 PM (<https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/#comment-109078>)

hey, i am getting these errors can you help me

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

```
NameError: name '_slang_loopup' is not defined
```



sreesha says:

REPLY (<https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/?replytocom=69665#respond>)
JANUARY 18, 2015 AT 4:39 PM (<https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/#comment-69665>)

Can you explain words standardization in more detail ,what are the improper formats ?



Saif Ali says:

REPLY (<https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/?replytocom=104200#respond>)



JANUARY 20, 2016 AT 2:35 PM ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/11/TEXT-DATA-CLEANING-STEPS-PYTHON/#COMMENT-104200](https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/#comment-104200))

Hey Shivam,

I want to tokenize 1000 tweets stored in a text file using a single loop in PYTHON , please can you help me with that?

please reply asap.

Thanks



Anurag Arora says: REPLY ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/11/TEXT-DATA-CLEANING-STEPS-PYTHON/?REPLYTOCOM=110469#RESPOND](https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/?replytocom=110469#respond))
MAY 5, 2016 AT 8:13 AM ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/11/TEXT-DATA-CLEANING-STEPS-PYTHON/#COMMENT-110469](https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/#comment-110469))

Can you please explain the code in Standardizing words(Point 9) ? Also, what other approaches can I take for the standardization of words ?



Sofea says: REPLY ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/11/TEXT-DATA-CLEANING-STEPS-PYTHON/?REPLYTOCOM=114479#RESPOND](https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/?replytocom=114479#respond))
AUGUST 5, 2016 AT 11:08 AM ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/11/TEXT-DATA-CLEANING-STEPS-PYTHON/#COMMENT-114479](https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/#comment-114479))

how to remove repetitive dots in tweets? for example "i love...."



Aneesh says: REPLY ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/11/TEXT-DATA-CLEANING-STEPS-PYTHON/?REPLYTOCOM=118198#RESPOND](https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/?replytocom=118198#respond))
NOVEMBER 10, 2016 AT 4:35 AM ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/11/TEXT-DATA-CLEANING-STEPS-PYTHON/#COMMENT-118198](https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/#comment-118198))

is there any downloadable dictionary for APPOSTOPHES

"APPOSTOPHES = {"s" : " is", "re" : " are", ...} ## Need a huge dictionary"

"



Shivam Bansal (<http://shivambansal.com>) says: REPLY ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/11/TEXT-DATA-CLEANING-STEPS-PYTHON/?REPLYTOCOM=118200#RESPOND](https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/?replytocom=118200#respond))
NOVEMBER 10, 2016 AT 5:46 AM ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/11/TEXT-DATA-CLEANING-STEPS-PYTHON/#COMMENT-118200](https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/#comment-118200))

Hi Aneesh,

Share your email, I will forward you the list.



Aneesh says:

REPLY (<https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/?replytocom=118202#respond>)
NOVEMBER 10, 2016 AT 6:10 AM (<https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/#comment-118202>)

a4aneeshc@gmail.com (<mailto:a4aneeshc@gmail.com>)



Partho says:

REPLY (<https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/?replytocom=118616#respond>)
NOVEMBER 21, 2016 AT 10:22 AM (<https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/#comment-118616>)

Hi Shivam,

Can you please provide dictionary for APPOSTOPHES, _slang_loopup and set of rules for word standardization ? It will be really helpful. My mail id : partho.iitm@gmail.com (<mailto:partho.iitm@gmail.com>) . Thanks.



Kevin Mahendra (<http://kevinsitumorang.myportfolio.com>) says:

REPLY (<https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/?replytocom=119824#respond>)
DECEMBER 16, 2016 AT 3:05 AM (<https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/#comment-119824>)

can i have what Partho wants to? Thanks! I really need that for my school assignment
@kaemsitumorang@gmail.com



Hunter Red says:

REPLY (<https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/?replytocom=123486#respond>)
MARCH 1, 2017 AT 9:12 AM (<https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/#comment-123486>)

Would you tell me "for_"in " tweet = ".join(".join(s)[:2] for _, s in itertools.groupby(tweet)) " what's meaning?thank you.

LEAVE A REPLY

Your email address will not be published.

Comment






Name (required)

Email (required)

Website

SUBMIT COMMENT

TOP ANALYTICS VIDHYA USERS

Rank	Name		Points
1		vopani (https://datahack.analyticsvidhya.com/user/profile/Rohan Rao)	7876
2		SRK (https://datahack.analyticsvidhya.com/user/profile/SRK)	6547
3		Aayushmnit (https://datahack.analyticsvidhya.com/user/profile/aayushmnit)	6101
4		binga (https://datahack.analyticsvidhya.com/user/profile/binga)	5044
5		Nalin Pasricha (https://datahack.analyticsvidhya.com/user/profile/Nalin)	4417

More Rankings (<http://datahack.analyticsvidhya.com/users>)



POST GRADUATE PROGRAM IN BUSINESS ANALYTICS

Admissions open

APPLY NOW

*In Chennai, Gurgaon, Bengaluru, Pune,
Hyderabad & Mumbai*

([http://www.greatlearning.in/great-lakes-pgpba?](http://www.greatlearning.in/great-lakes-pgpba?utm_source=avm&utm_medium=avmbanner&utm_campaign=pgpba)

[utm_source=avm&utm_medium=avmbanner&utm_campaign=pgpba](http://www.greatlearning.in/great-lakes-pgpba?utm_source=avm&utm_medium=avmbanner&utm_campaign=pgpba))

A promotional banner for UpGrad and IIT-B. The left side is black with white and yellow text. The right side shows a man with glasses smiling. The text includes "UpGrad | IIT-B", "GET HIRED BY THE BEST WITH IIT-B CAREER SUPPORT", a list of companies (Thomson Reuters, Opera, PwC), and an "APPLY NOW" button.

UpGrad | IIT-B

**GET HIRED
BY THE BEST WITH
IIT-B CAREER SUPPORT**

- THOMSON REUTERS
- OPERA
- **pwc** & more

APPLY NOW

([https://upgrad.com/data-analytics?](https://upgrad.com/data-analytics?utm_source=AV&utm_medium=Display&utm_campaign=DA_AV_Banner&utm_term=DA_AV_Banner&utm_content=DA_AV_Banner)

[utm_source=AV&utm_medium=Display&utm_campaign=DA_AV_Banner&utm_term=DA_AV_Banner&utm_content=DA_AV_Banner](https://upgrad.com/data-analytics?utm_source=AV&utm_medium=Display&utm_campaign=DA_AV_Banner&utm_term=DA_AV_Banner&utm_content=DA_AV_Banner))