

x
-
x
-
x
-

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

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

 LOGIN / REGISTER ([HTTPS://ID.ANALYTICSVIDHYA.COM/ACCOUNTS/LOGIN/?](https://id.analyticsvidhya.com/accounts/login/?)

NEXT=[HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2018/02/THE-DIFFERENT-METHODS-DEAL-TEXT-DATA-PREDICTIVE-](https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/)



PYTHON/) LOGIN / REGISTER ([HTTPS://ID.ANALYTICSVIDHYA.COM/ACCOUNTS/LOGIN/?](https://id.analyticsvidhya.com/accounts/login/?)

NEXT=[HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2018/02/THE-DIFFERENT-METHODS-DEAL-TEXT-DATA-PREDICTIVE-](https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/)

PYTHON/)




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



Applied Machine Learning

16 - 17TH FEB, 2019 | DELHI NCR



Early Bird Offer

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+AMLW101+AMLW101_Jan2019/about?utm_source=AVtopBanner)

[MACHINE LEARNING \(HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/MACHINE-LEARNING/\)](https://www.analyticsvidhya.com/blog/category/machine-learning/)

[NLP \(HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/NLP/\)](https://www.analyticsvidhya.com/blog/category/nlp/)

[PYTHON \(HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/PYTHON-2/\)](https://www.analyticsvidhya.com/blog/category/python-2/)

Ultimate guide to deal with Text Data (using Python) – for Data Scientists & Engineers


```
TextBlob(train['tweet'][0]).ngrams(2)
> [WordList(['user', 'when']),
  WordList(['when', 'a']),
  WordList(['a', 'father']),
  WordList(['father', 'is']),
  WordList(['is', 'dysfunctional']),
  WordList(['dysfunctional', 'and']),
  WordList(['and', 'is']),
  WordList(['is', 'so']),
  WordList(['so', 'selfish']),
  WordList(['selfish', 'he']),
  WordList(['he', 'drags']),
  WordList(['drags', 'his']),
  WordList(['his', 'kids']),
  WordList(['kids', 'into']),
  WordList(['into', 'his']),
  WordList(['his', 'dysfunction']),
  WordList(['dysfunction', 'run'])]
```

3.2 Term frequency

Term frequency is simply the ratio of the count of a word present in a sentence, to the length of the sentence.

Therefore, we can generalize term frequency as:

TF = (Number of times term T appears in the particular row) / (number of terms in that row)

To understand more about Term Frequency, have a look at [this article](https://www.analyticsvidhya.com/blog/2015/04/information-retrieval-system-explained/).

(<https://www.analyticsvidhya.com/blog/2015/04/information-retrieval-system-explained/>).

Below, I have tried to show you the term frequency table of a tweet.

End Notes

I hope that now you have a basic understanding of how to deal with text data in predictive modeling. These methods will help in extracting more information which in return will help you in building better models.

I would recommend practising these methods by applying them in machine learning/deep learning competitions. You can also start with the Twitter sentiment problem we covered in this article (the dataset is available on the datahack (<https://datahack.analyticsvidhya.com/contest/all/>) platform of AV).


Did you find this article helpful? Please share your opinions/thoughts in the comments section below.


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





https://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/2018/02/the-different-methods-deal-text-data-predictive-python/?share=linkedin&nb=1&nb=1>)

 (<https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/?share=facebook&nb=1&nb=1>)

 (<https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/?share=google-plus-1&nb=1&nb=1>)

 (<https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/?share=twitter&nb=1&nb=1>)

 (<https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/?share=pocket&nb=1&nb=1>)

 (<https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/?share=reddit&nb=1&nb=1>)

Like this:



(<https://www.analyticsvidhya.com/blog/author/shubham-jain/>)

Shubham Jain

(<https://www.analyticsvidhya.com/blog/author/shubham-jain/>)

I am currently pursuing my B.Tech in Ceramic Engineering from IIT (B.H.U) Varanasi. I am an aspiring data scientist and a ML enthusiast. I am really passionate about changing the world by using artificial intelligence.

✉ (<mailto:shubham.jain.cer14@itbhu.ac.in>)

in (<https://www.linkedin.com/in/shubham-jain-25a104108/>)

RELATED ARTICLES

KUNAL JAIN (<https://www.analyticsvidhya.com/blog/2016/10/xgboost-machine-learning-iphone-apple-coreml-study-of-factors-contributing-to-air-pollution/>)



(<https://www.analyticsvidhya.com/blog/2016/10/xgboost-machine-learning-iphone-apple-coreml-study-of-factors-contributing-to-air-pollution/>)

Winners Approach & Codes from Knocktober on iPhone (Intro to Apple's CoreML) Complete Study of Factors Contributing to Air Pollution
How to build your first Machine Learning model on iPhone (Intro to Apple's CoreML) Complete Study of Factors Contributing to Air Pollution
(<https://www.analyticsvidhya.com/blog/2016/10/xgboost-machine-learning-iphone-apple-coreml-study-of-factors-contributing-to-air-pollution/>)



(<https://www.analyticsvidhya.com/blog/2017/01/sentiment-analysis-of-twitter-posts-on-chennai-floods-using-python/>) (<https://www.analyticsvidhya.com/blog/2016/11/interesting-data-science-games-to-break-the-monday-blues/>) (<https://www.analyticsvidhya.com/blog/2016/11/building-a-mask-r-cnn-model-to-detect-damage-in-cars-using-python/>)

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 Discussion portal (<https://disc.python.analyticsvidhya.com/>) to get your queries resolved.

[Sentiment Analysis of Twitter Posts on Chennai Floods using Python](https://www.analyticsvidhya.com/blog/2017/01/sentiment-analysis-of-twitter-posts-on-chennai-floods-using-python/) ([interesting-data-science-games-to-break-the-monday-blues/](https://www.analyticsvidhya.com/blog/2016/11/interesting-data-science-games-to-break-the-monday-blues/)) ([building-a-mask-r-cnn-model-to-detect-damage-in-cars-using-python/](https://www.analyticsvidhya.com/blog/2016/11/building-a-mask-r-cnn-model-to-detect-damage-in-cars-using-python/))

26 COMMENTS



YONGDUEK SEO

[Reply](#)

February 27, 2018 at 12:48 pm (<https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/#comment-151603>)

str(x).split() instead produces better result without empty words.



SOURAV MAHARANA

[Reply](#)

February 27, 2018 at 2:55 pm (<https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/#comment-151606>)

Regarding your last section. You used glove model to find similarity between words or find a similar word to the target word. If i want to find a similar document to my target document, then can I achieve this by word embedding? how?

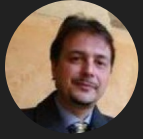


SHUBHAM JAIN

[Reply](#)

May 14, 2018 at 11:09 am (<https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/#comment-153236>)

For finding similarity between documents, you can try with help of building document vector using doc2vec.



GIANNI

[Reply](#)

February 27, 2018 at 4:08 pm (<https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/#comment-151610>)

Great job Shubham !

Every Time I peek in AV I got mesmerized 😊 thank you all folks !



SHUBHAM JAIN

[Reply](#)

March 1, 2018 at 12:19 am (<https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/#comment-151656>)

Glad you liked the article. 😊



JEFF

[Reply](#)

February 28, 2018 at 2:14 am (<https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/#comment-151627>)

Excellent write-up. Keep up the good work. Thank you so much.



PRADYUT DASGUPTA

[Reply](#)

February 28, 2018 at 8:34 am (<https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/#comment-151631>)

Hi , I am not able to find the data set. Kindly help.!



SHUBHAM JAIN

[Reply](#)

March 1, 2018 at 12:18 am (<https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/#comment-151655>)

You can find the dataset from here.

<https://datahack.analyticsvidhya.com/contest/practice-problem-twitter-sentiment-analysis/>
(<https://datahack.analyticsvidhya.com/contest/practice-problem-twitter-sentiment-analysis/>)



MAHESH

[Reply](#)

July 25, 2018 at 2:38 pm (<https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/#comment-154246>)

I'm not able to find the dataset in the above link. "Data" link present in that page doesn't perform any action at all so, I guess it's removed from that link. Can you pls check once and provide the link witch which I can directly download the dataset?



PULKIT SHARMA

[Reply](#)

July 25, 2018 at 8:03 pm (<https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/#comment-154260>)

Hi Mahesh,

You can find the dataset from here.

<https://datahack.analyticsvidhya.com/contest/practice-problem-twitter-sentiment-analysis/>
(<https://datahack.analyticsvidhya.com/contest/practice-problem-twitter-sentiment-analysis/>)



VIVEK

[Reply](#)

February 28, 2018 at 11:35 am (<https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/#comment-151637>)

Excellent article, easy to understand.



SUSANT

[Reply](#)

March 6, 2018 at 8:02 pm (<https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/#comment-151727>)

Ultimate guide ,Shubham..very well written..



PRAKASH

[Reply](#)

March 9, 2018 at 6:22 am (<https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/#comment-151787>)

Very useful article.

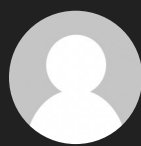


MARCEL

[Reply](#)

March 12, 2018 at 1:39 pm (<https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/#comment-151845>)

My compliments for this nice article.



SATISH

[Reply](#)

March 16, 2018 at 6:32 pm (<https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/#comment-151943>)

Can you please elaborate on N-grams.. what the use of n-grams and what happens if we choose high n values



SHUBHAM JAIN

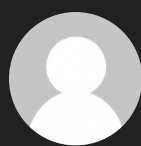
[Reply](#)

May 14, 2018 at 11:20 am (<https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/#comment-153237>)

N-grams are generally preferred to learn some sequential order in our model. We prefer small values of N because otherwise our model will become very slow and will also require higher computational power. So, instead of using higher values of N, we generally prefer using sequential modeling techniques like RNN, LSTM.

Hope this helps.

Shubham



VANI

[Reply](#)

March 19, 2018 at 2:11 pm (<https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/#comment-151996>)

can u suggest some topic related to textdata for research



MEL

[Reply](#)

April 2, 2018 at 3:45 pm (<https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/#comment-152314>)

Good day – Thank you for the example. It provides good guidelines to newbies like me.

I was able to follow your example right up til 3.3 Inverse Document Frequency, but sample code does not seem to work, Additionally, the output provided seems to come from another dataset or rather a copy /paste from a previous article ?

Finally, the numerical sections following are not labeled correctly. Jumping from 3.3 to 3.34 then 4.5, 4.6

Could you kindly update?



SHUBHAM JAIN

[Reply](#)

April 3, 2018 at 8:11 pm (<https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/#comment-152334>)

The code seems to be fine with me. And the output is also correct. Try to follow the preprocessing steps properly and then run it again.

As far as the numbering of sections is concerned, they were just mistakenly put by me. Not a big issue though since it is clear from the table of content. Still, I have updated it.

Regards,
Shubham



RUBEN

[Reply](#)

April 22, 2018 at 9:59 pm (<https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/#comment-152768>)

Hi Shubham, great tutorial!

Only thing is that I'm getting stuck at the same point (3.3 ITF):

[CODE]

NameError Traceback (most recent call last)

in ()

1 for i, word in enumerate(tf1['words']):

--> 2 tf1.loc[i, 'idf'] = np.log(train.shape[0]/(len(train[train['tweet'].str.contains(word)])))

3

4 tf1

NameError: name 'np' is not defined [/CODE]

I've cleared the notebook output multiple times, but it keeps giving me the same error.

I'll appreciate any help, thanks!



GANESH

[Reply](#)

June 19, 2018 at 12:55 pm (<https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/#comment-153863>)

use below one and proceed

```
import numpy as np
```



DENNIS

[Reply](#)

April 18, 2018 at 8:04 pm (<https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/#comment-152659>)

how do you now use the trained model



HAKAN

[Reply](#)

April 25, 2018 at 2:38 am (<https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/#comment-152821>)

Hi Shubham,

Thank you for the article. It is really helpful for text analysis. One thing I cannot quite understand is how can I use features I extracted from text such as number of numerics, number of uppercase with TFIDF vector. I couldn't find an intuitive explanation or example of this. Could you be able to make an example of it ?

Thanks again.



DATTA SAI

[Reply](#)

May 18, 2018 at 12:46 pm (<https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/#comment-153367>)

what is the pd there in :

```
freq = pd.Series(' '.join(train['tweet']).split()).value_counts()[-10:]
```

thanks in advance.



AISHWARYA SINGH

[Reply](#)

May 18, 2018 at 1:51 pm (<https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/#comment-153370>)

Hi Datta,

pd here represents pandas. The library pandas is imported as pd.







SRIRAM

[Reply](#)

July 4, 2018 at 2:19 am (<https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/#comment-154050>)

Hi Shubham, great article, thanks.

TOP ANALYTICS VIDHYA USERS

Rank	Name		Points
1		SRK (https://datahack.analyticsvidhya.com/user/profile/SRK)	9688
2		Rohan Rao (https://datahack.analyticsvidhya.com/user/profile/Rohan_Rao)	9200
3		aayushmnit (https://datahack.analyticsvidhya.com/user/profile/aayushmnit)	7779
4		mark12 (https://datahack.analyticsvidhya.com/user/profile/mark12)	7212



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

POPULAR POSTS

- 24 Ultimate Data Science Projects To Boost Your Knowledge and Skills (& can be accessed freely) (<https://www.analyticsvidhya.com/blog/2018/05/24-ultimate-data-science-projects-to-boost-your-knowledge-and-skills/>)
- A Complete Tutorial to Learn Data Science with Python from Scratch

Introduction to Monte Carlo Tree Search: The Game-Changing Algorithm behind DeepMind's AlphaGo (<https://www.analyticsvidhya.com/blog/2019/01/monte-carlo-tree-search-introduction-algorithm-deepmind-alphago/>)

JANUARY 24, 2019

Top 10 Presentations from rstudio::conf 2019 – The Best R Conference of the Year! (<https://www.analyticsvidhya.com/blog/2019/01/top-highlights-rstudioconf-2019-best-r-conference/>)

JANUARY 22, 2019

Must-Read Tutorial to Learn Sequence Modeling (deeplearning.ai Course #5) (<https://www.analyticsvidhya.com/blog/2019/01/sequence-models-deeplearning/>)

JANUARY 21, 2019



([http://www.edvancer.in/certified-data-scientist-with-](http://www.edvancer.in/certified-data-scientist-with-python-course?utm_source=AV&utm_medium=AVads&utm_campaign=AVadsnonfc&utm_content=pythonavad)

[python-course?](http://www.edvancer.in/certified-data-scientist-with-python-course?utm_source=AV&utm_medium=AVads&utm_campaign=AVadsnonfc&utm_content=pythonavad)

[utm_source=AV&utm_medium=AVads&utm_campaign=AVadsnonfc&utm_content=pythonavad](http://www.edvancer.in/certified-data-scientist-with-python-course?utm_source=AV&utm_medium=AVads&utm_campaign=AVadsnonfc&utm_content=pythonavad))

Learn Python, Statisti



(<https://trainings.analyticsvidhya.com/courses/course->

v1:AnalyticsVidhya+DS101+2018T2/about?utm_source=Blog&utm_medium=Sticky_banner1)

Almost 1.5 Billion Seconds of Videos Are Being Uploaded To Youtube Every Day !

(<https://trainings.analyticsvidhya.com/courses/course->

v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?
utm_source=Blog&utm_medium=Sticky_banner2&utm_campaign=cv10percent)

ANALYTICS
VIDHYA

DATA
SCIENTISTS

COMPANIES

JOIN OUR COMMUNITY :

About Us

(http://www.analyticsvidhya.com/about-me/)

Our Team

(https://www.analyticsvidhya.com/about-me/team/)

Career

(https://www.analyticsvidhya.com/career-analytcs-vidhya/)

Contact Us

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

Write for us

(https://www.analyticsvidhya.com/about-me/write/)

Blog

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

Hackathon

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

Discussions

(https://discuss.analyticsvidhya.com/join)

Apply Jobs

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

Leaderboard

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

Post Jobs

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

Trainings

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

Filling

Hackathons

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

Advertising

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

Reach Us

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



(https://www.facebook.com/AnalyticsVidhya)



(https://twitter.com/AnalyticsVidhya)



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



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

3057

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

Followers

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



(https://twitter.com/AnalyticsVidhya)



(https://twitter.com/AnalyticsVidhya)



19433



(https://twitter.com/AnalyticsVidhya)



(https://twitter.com/AnalyticsVidhya)



(https://twitter.com/AnalyticsVidhya)



(https://twitter.com/AnalyticsVidhya)

Subscribe to emailer

