



Benefits of mining Twitter 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



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

- 40 Questions to test a Data Scientist on Clustering Techniques (Skill test Solution)
- Introductory guide on Linear Programming for (aspiring) data scientists
- 40 Questions to test a data scientist on Machine Learning [Solution: SkillPower – Machine Learning, DataFest 2017]
- 6 Easy Steps to Learn Naive Bayes Algorithm with codes in Python and R
- 25 Questions to test a Data Scientist on Support Vector Machines

Tableau Data Storytelling

GET THE WHITEPAPER

analyticsvidhya.com/blog/2015/06/quick-guide-text-data-cleaning-python/?utm_source=linkedin.com&utm_medium=social

Analytics Vidhya

BLOG COURSES HACKATHONS JOBS AI & ML BLACKBELT ASCEND PRO WRITE FOR US CONTACT

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

Steps for Data Cleaning

STEP 01

Escaping HTML characters

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 hahpppppy http://www.apple.com"

CAREER RESOURCES

 16 Key Questions You Should Answer Before Transitioning into Data Science
NOVEMBER 23, 2020

 Here's What You Need to Know to Become a Data Scientist!
JANUARY 22, 2021

 These 7 Signs Show you have Data Scientist Potential!
DECEMBER 3, 2020

 How To Have a Career in Data Science (Business Analytics)?
NOVEMBER 26, 2020

 Should I become a data scientist (or a business analyst)?
NOVEMBER 24, 2020



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 haooooo :) http://www.apple.com"
```

STEP
03

Apostrophe Lookup

Code

```
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

RECENT POSTS



The Ultimate Swiss Army
Knife of 'apply' Family in R

FEBRUARY 11, 2021



An Introduction to
Normalization Theory

FEBRUARY 11, 2021



Getting Started with
Analytics: The Actionable
and Measurable Way

FEBRUARY 11, 2021



Introduction to Hugging
Face's Transformers v4.3.0
and its First Automatic
Speech Recognition Model –
Wav2Vec2

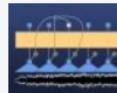
FEBRUARY 10, 2021





Analytics: The Actionable and Measurable Way

FEBRUARY 11, 2021



Introduction to Hugging Face's Transformers v4.3.0 and its First Automatic Speech Recognition Model – Wav2Vec2

FEBRUARY 10, 2021



Your Journey to Become an AI & ML Professional Starts Here!



- 75+ Mentorship Sessions
- 39+ Real-World Projects
- Interview Preparation

Certified AI & ML BlackBelt Plus Program

Know More

Removal of Stop-Words

STEP
04

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.





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.



Your Journey to Become an
AI & ML Professional Starts Here!

- 75+ Mentorship Sessions
- 39+ Real-World Projects
- Interview Preparation



Certified AI & ML BlackBelt Plus Program

Know More

STEP
06

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.



Ascend Pro

Mastering Data Science For Industry

Become a Job-Ready
Data Science Professional...

- Acquire Industry-Relevant Skills
- Work-on Industry Projects
- Get Placed in Top Companies



Download Brochure

STEP

Split Attached Words

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

Importing_Data_Py....pdf

Python_SciPy_Che....pdf

Pandas_Cheat_Sheet.pdf

Numpy_Python_C....pdf

Show all



STEP 07 Split Attached Words

Code

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

Outcome

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

Your Journey to Become an **AI & ML Professional** Starts Here!

- 75+ Mentorship Sessions
- 39+ Real-World Projects
- Interview Preparation

Certified AI & ML BlackBelt *Plus* Program

[Know More](#)

STEP 08 Slangs lookup

Code

```
tweet = _slang_lookup(tweet)
```

Outcome

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

Ascend Pro

Mastering Data Science For Industry

Become a Job-Ready Data Science Professional...

- Acquire Industry-Relevant Skills
- Work-on Industry Projects
- Get Placed in Top Companies

[Download Brochure](#)



Standardizing word

STEP
09

Code

```
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"



Your Journey to Become an
AI & ML Professional Starts Here!

- 75+ Mentorship Sessions
- 39+ Real-World Projects
- Interview Preparation



Certified AI & ML BlackBelt Plus Program

Know More

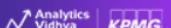
Removal of URLs

STEP
10

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 ,:)



Ascend Pro

Mastering Data Science For Industry

Become a Job-Ready
Data Science Professional...

- Acquire Industry-Relevant Skills
- Work-on Industry Projects
- Get Placed in Top Companies



Download Brochure



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.

Spelling correction

In natural language, misspelled errors are encountered. One can use algorithms like the Levenshtein Distances, Dictionary Lookup etc. other modules and packages to fix these errors.

Your Next Steps...

Now that the data (tweet) is cleaned, you are ready to practice and learn the following techniques (in no order) of Text Mining-

1. Framework to build a niche dictionary for text mining

<http://bit.ly/TextMining>



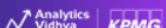
Your Journey to Become an
AI & ML Professional Starts Here!

- 75+ Mentorship Sessions
- 39+ Real-World Projects
- Interview Preparation



Certified AI & ML BlackBelt Plus Program

Know More



Ascend Pro

Mastering Data Science For Industry

Become a Job-Ready
Data Science Professional...

- Acquire Industry-Relevant Skills
- Work-on Industry Projects
- Get Placed in Top Companies



Download Brochure