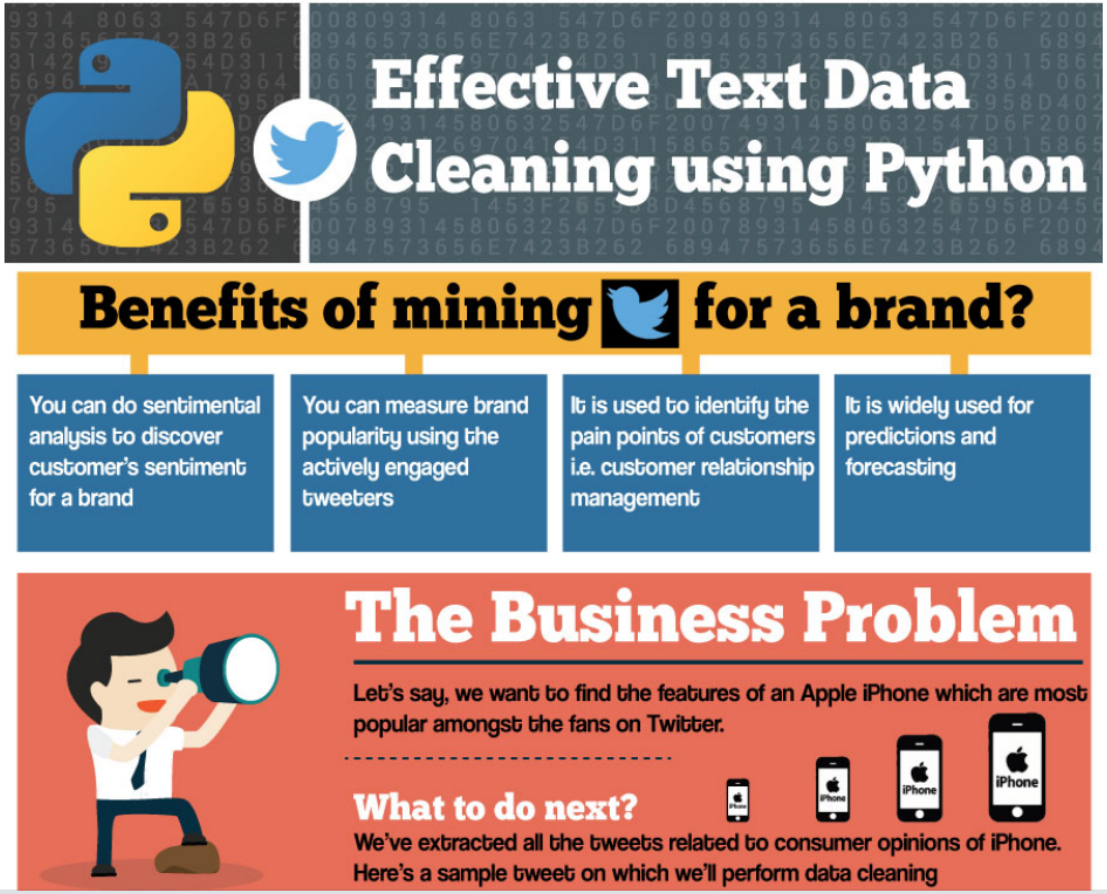


– 40 Questions to test a Data Scientist on Clustering Techniques (Skill test Solution)
– Introductory guide on Linear Programming for (aspiring) data scientists
– 40 Questions to test a data scientist on Machine Learning [Solution: SkillPower – Machine Learning, DataFest 2017]
– 6 Easy Steps to Learn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

CAREER RESOURCES

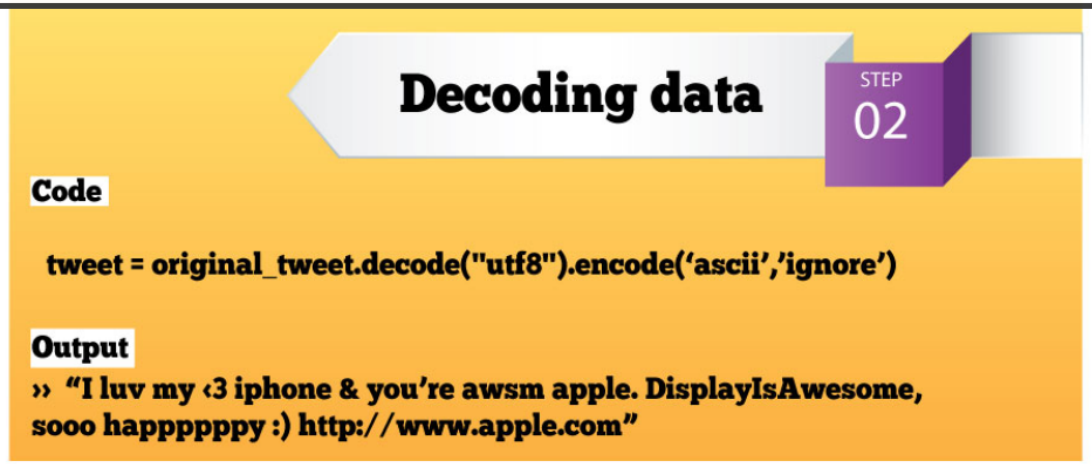

16 Key Questions You Should Answer Before Transitioning into Data Science
NOVEMBER 23, 2020

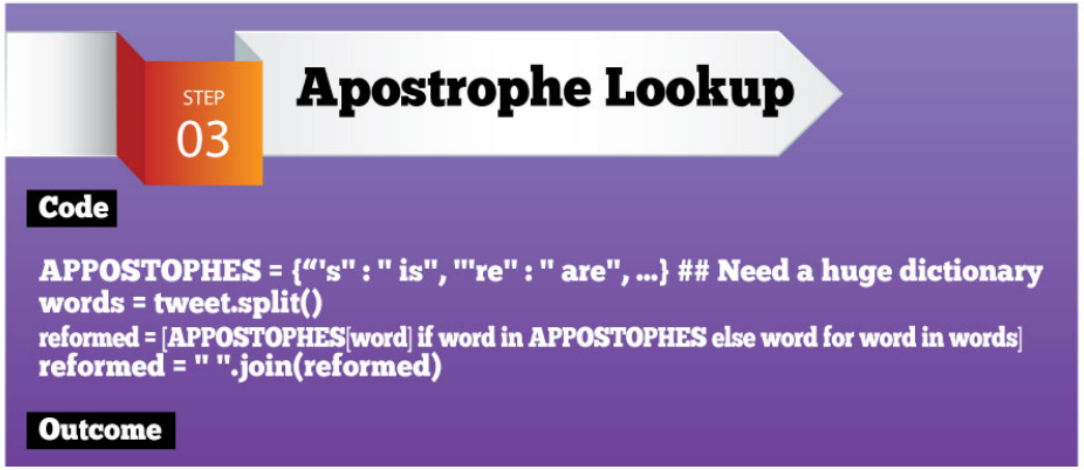

Here's What You Need to Know to Become a Data Scientist!
JANUARY 22, 2021


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

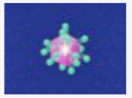

Should I become a data scientist (or a business analyst)?
NOVEMBER 24, 2020





analyst)?
NOVEMBER 24, 2020

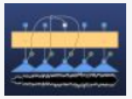
RECENT POSTS



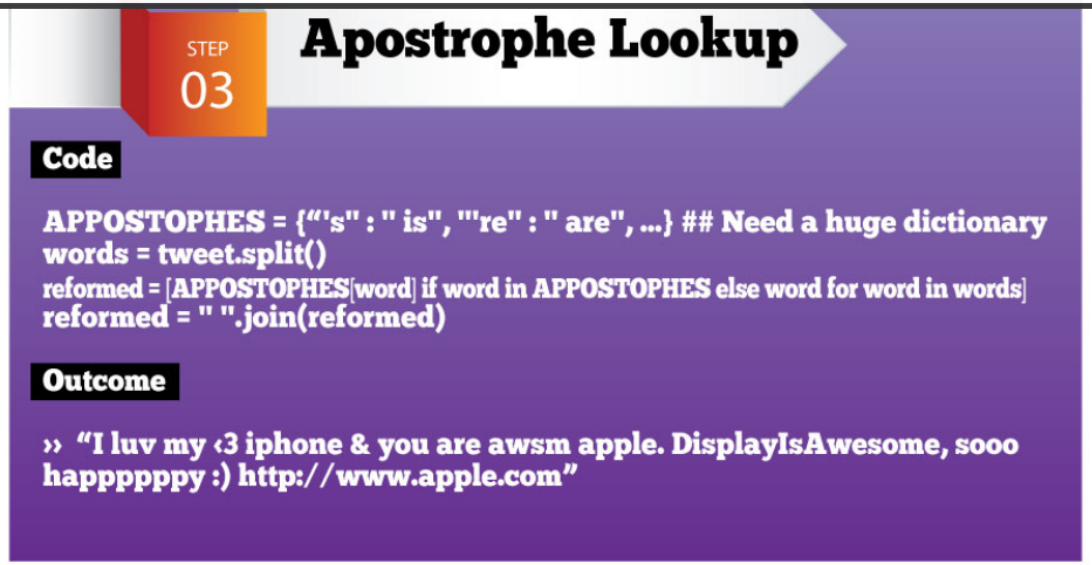
The Ultimate Swiss Army Knife of 'apply' Family in R
FEBRUARY 11, 2021

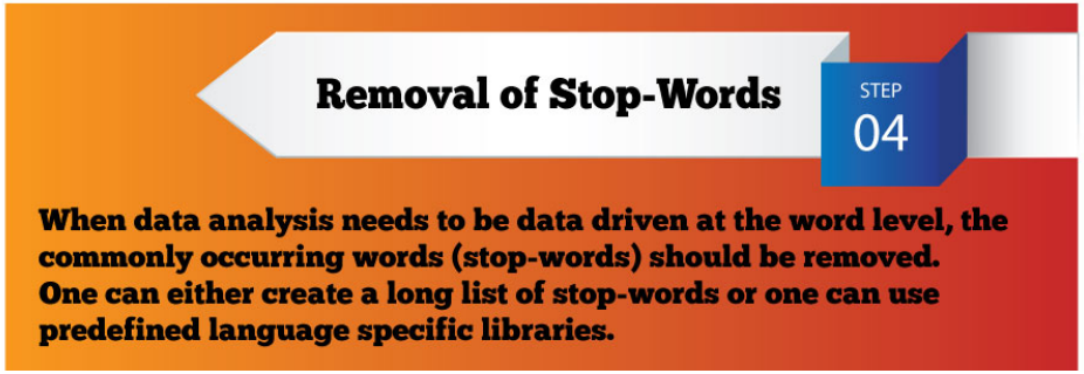
An Introduction to Normalization Theory
FEBRUARY 11, 2021

Getting Started with Analytics: The Actionable and Measurable Way
FEBRUARY 11, 2021

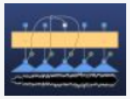


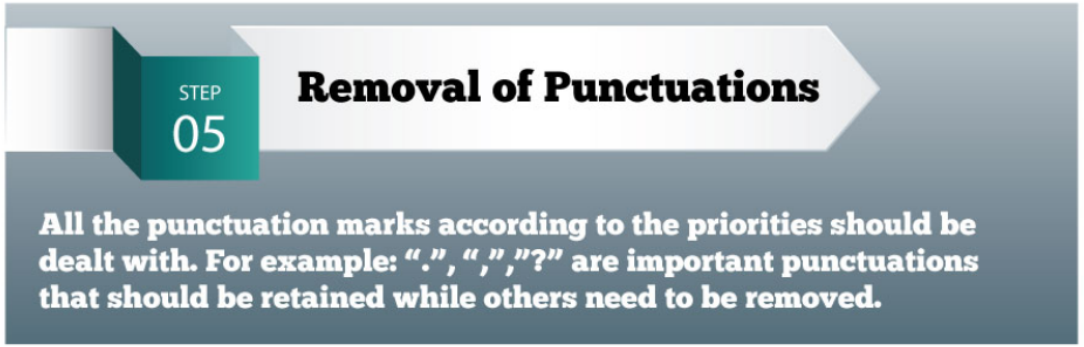
Introduction to Hugging Face's Transformers v4.3.0 and its First Automatic Speech Recognition Model – Wav2Vec2
FEBRUARY 10, 2021

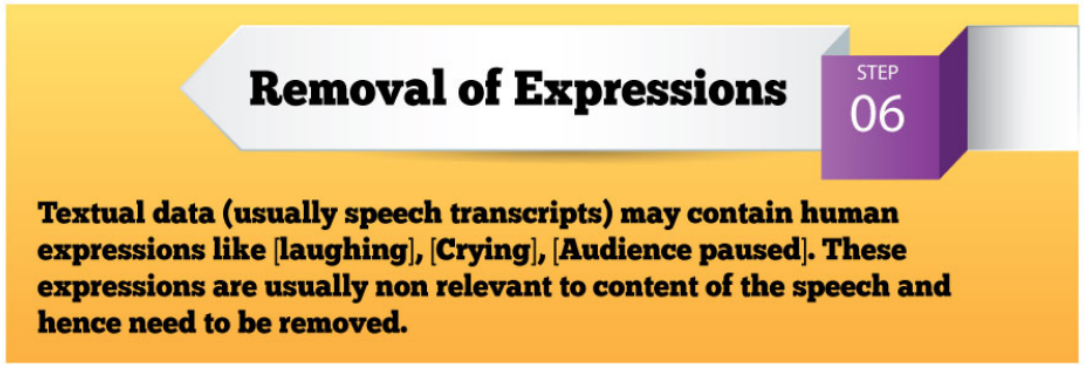





Analytics: The Actionable and Measurable Way
FEBRUARY 11, 2021


Introduction to Hugging Face's Transformers v4.3.0 and its First Automatic Speech Recognition Model – Wav2Vec2
FEBRUARY 10, 2021


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.

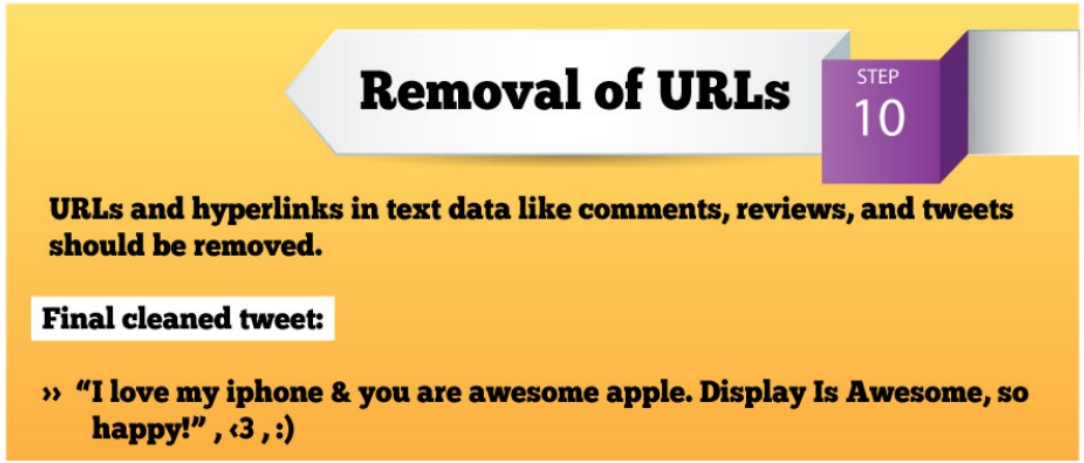

Removal of Expressions
STEP
06
Textual data (usually speech transcripts) may contain human expressions like [laughing], [Crying], [Audience paused]. These expressions are usually non relevant to content of the speech and hence need to be removed.


Split Attached Words


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 happppppy :) http://www.apple.com"


Slangs lookup
STEP
08
Code
tweet = _slang_loopup(tweet)
Outcome
›› "I love my ‹3 iphone & you are awesome apple. Display Is Awesome, sooo happppppy :) http://www.apple.com"


STEP
09
Standardizing word
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"


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 , :)

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