

Benefits of mining for a brand?

analysis to discover customer's sentiment for a brand

You can do sentimental

popularity using the actively engaged tweeters

You can measure brand

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

TWEET

"I luv my 8lt;3 iphone 6amp; you're awsm apple. DisplayIsAwesome, sooo happppppy:) http://www.apple.com"



Escaping HTML characters

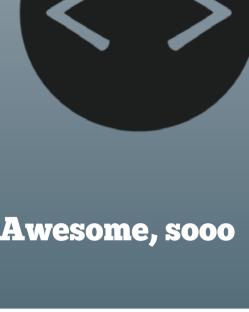
01 Code import HTMLParser

STEP

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

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

Output

Code

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

» "I luv my <3 iphone & you're awsm apple. DisplayIsAwesome,

STEP

03 Code

Apostrophe Lookup

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

04 When data analysis needs to be data driven at the word level, the

commonly occurring words (stop-words) should be removed.

predefined language specific libraries.

Removal of Stop-Words



STEP

Removal of Punctuations STEP 05 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

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

Split Attached Words

STEP

08

STEP

06

Outcome >> "I luv my <3 iphone & you are awsm apple. Display Is Awesome, sooo

Code

hence need to be removed.

STEP

07

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

Slangs lookup

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

Outcome

STEP

09

Code

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

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

>> "I love my <3 iphone & you are awesome apple. Display Is

Awesome, so happy:) http://www.apple.com

Standardizing word

tweet = _slang_loopup(tweet)

Code

Outcome

URLs and hyperlinks in text data like comments, reviews, and tweets should be removed.

Removal of URLs

Advanced Data Cleaning

STEP

10

» "I love my iphone & you are awesome apple. Display Is Awesome, so happy!", <3,:)

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

errors.

Grammar checking

Final cleaned tweet:

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

Your Next Steps...

http://bit.ly/1JjslYe

following techniques (in no order) of Text Mining-1. Framework to build a niche dictionary for text mining http://bit.ly/1eetMw6

Now that the data (tweet) is cleaned, you are ready to practice and learn the

- 3. 2014 FIFA World Cup Prediction using Twitter Mining http://bit.ly/1kLeYSk

Learn Everything About Analytics

4. Text Mining Hack using Google API http://bit.ly/1LDPF6c

2 Step by Step guide to extract insights from free text

For more resources on analytics/data science, visit www.analyticsvidhya.com