Retrieving Meaning from Words

Using natural language parsing to learn from large quantities of text

Nathaniel Case & Eitan Romanoff

Tokenize

The first step is to take a block of text and split it into atomic tokens of words. These tokens may then be analyzed for which part of speech they belong to. This produces a list of tuples containing (token, tag)

```
>>> nltk.word_tokenize('NLTK is pretty awesome')
['NLTK', 'is', 'pretty', 'awesome']
>>> nltk.pos_tag(_)
[('NLTK', 'NN'), ('is', 'VBZ'), ('pretty', 'RB'), ('awesome', 'RB')]
```

Simplify

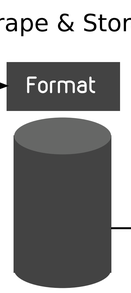
Tokens can be simplified according to what root the words belong to in a process known as stemming or lemmatization. Stemming assumes that all words which share a prefix

Classify

NLTK includes classifiers and other learning algorithms to make sense of your data

```
>>> classifier = nltk.NaiveBayesClassifier.train(training_set)
>>> classifier.show_most_informative_features(5)
Most Informative Features
          contains(just) = True           romney : obama  =      3.4 : 1.0
          contains(news) = True            obama : romney =      2.8 : 1.0
        contains(obama) = False           romney : obama  =      2.5 : 1.0
        contains(obama) = True             obama : romney =      2.3 : 1.0
        contains(think) = True           romney : obama  =      2.0 : 1.0
```

Mining Political Tweets



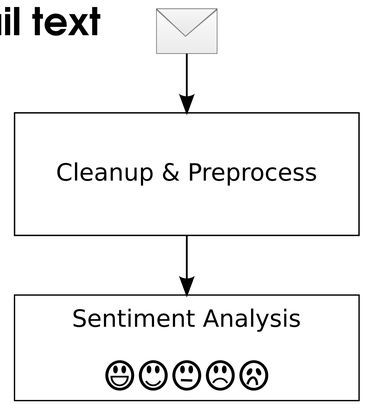
A two-hour debate session resulted in tens of millions of tweets, of which only 1% could be captured. Data is verified as correct JSON entries before the relevant parts are pulled out, formatted, and stored into a MongoDB database for later analysis.

Understanding email text

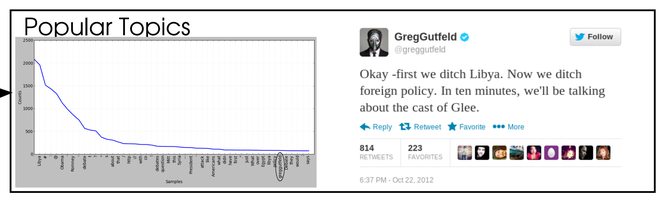
Emails sent to a local school district following an incident involving their students

Slang and misspellings commonly found, confusing NLTK

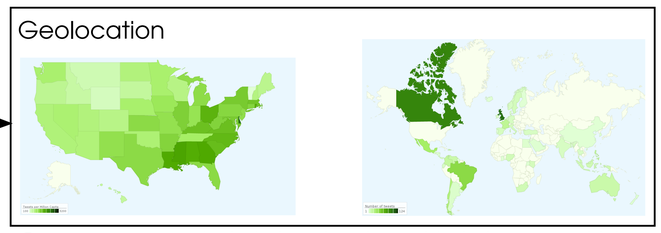
Many emails referenced youtube videos, which confuse most tokenizers



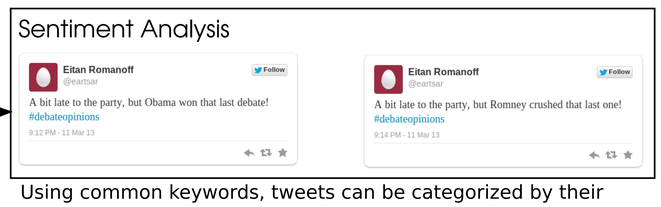
Visualizations



Simple frequency analysis can show off features of the data, including popular topics and influential users



Some users opt to include location data with their tweet, which allows us to visualize engagement in a geographic context



Using common keywords, tweets can be categorized by their content.

What is NLTK?

NLTK is a library for parsing and understanding natural language with Python.

What does that mean?

Tokenizing: Taking a string of words and splitting them into individual words or tokens.

Tagging: Taking a token in the context of the sentence to find its part of speech (verb, noun, adjective, etc.)

Classification: A process of training and applying a set of rules to a dataset so that a decision may be reached about their content.

mongoDB
{name: "mongo", type:"DB"}

FOSS@RIT

NLTK
http://nltk.org/