Natural Language

Toolkit

@victorneo

Natural Language Processing

"the process of a computer **extracting meaningful information from natural language input** and/or producing natural language output"

Dan Jurafsky

Language Technology

mostly solved

Spam detection

Let's go to Agra! ✓

Buy V1AGRA ... ✗

Part-of-speech (POS) tagging

ADJ	ADJ	NOUN	VERB	ADV
Colorless	green	ideas	sleep	furiously.

Named entity recognition (NER)

PERSON	ORG	LOC

Einstein met with UN officials in Princeton

making good progress

Sentiment analysis

Best roast chicken in San Francisco!

The waiter ignored us for 20 minutes.

Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my ***mouse***.

Parsing

I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕...

The 13th Shanghai International Film Festival...

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

still really hard

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

→ Economy is good

Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

Getting started with NLTK

NLTK

Open source Python modules, linguistic data and documentation for research and development in natural language processing and text analytics, with distributions for Windows, Mac OSX and Linux.

Natural Language Toolkit

Search this site

Navigation

- **NLTK Home**
- Book
- Code
- Contribute
- Courses
- Data
- Documentation
- Download
- FAQ
- Forums
- Getting Started
- HOWTO
- News Archive
- Projects
- Quotes
- Teaching

Software License

Apache License 2.0

Website License

Creative Commons Attribution Noncommercial No Derivative Works 3.0 US License

Feedback

This site is maintained by Steven Bird.

NLTK Home

Open source Python modules, linguistic data and documentation for research and development in natural language processing and text analytics, with distributions for Windows, Mac OSX and Linux.

- News - NLTK development has moved to GitHub [October 2011], Version 2.0.1rc1 released [April 2011], NLTK Cookbook by Jacob Perkins [December 2010], NLTK book in third printing [November 2010], Japanese translation of NLTK book published [November 2010]
- Code - functionality provided by NLTK in over 100,000 lines of Python code, distributed under the Apache License
- Data - ~60 corpora, grammar collections, and trained models that come with NLTK
- Quotes - what people have said about NLTK

Getting Started

- Documentation - book, articles, guides, reviews
- Download - instructions for downloading and installing Python and NLTK on all platforms
- Getting Started - simple things to try, including NLTK's demonstrations
- Subscribe - sign up for important announcements - approx 1 post per month

Getting Help

- FAQ - answers to frequently asked questions
- HOWTO - guides for a variety of NLTK packages, including many examples
- User forum - mailing list for discussion amongst NLTK users
- Chatroom - #nltk on irc.freenode.net (not often staffed)

Software

- API Documentation - complete documentation of all NLTK modules
- Source - browse the Python source code
- github - the home of the NLTK development work (submit a feature request or bug report)
- People - the NLTK development team

Education and Research

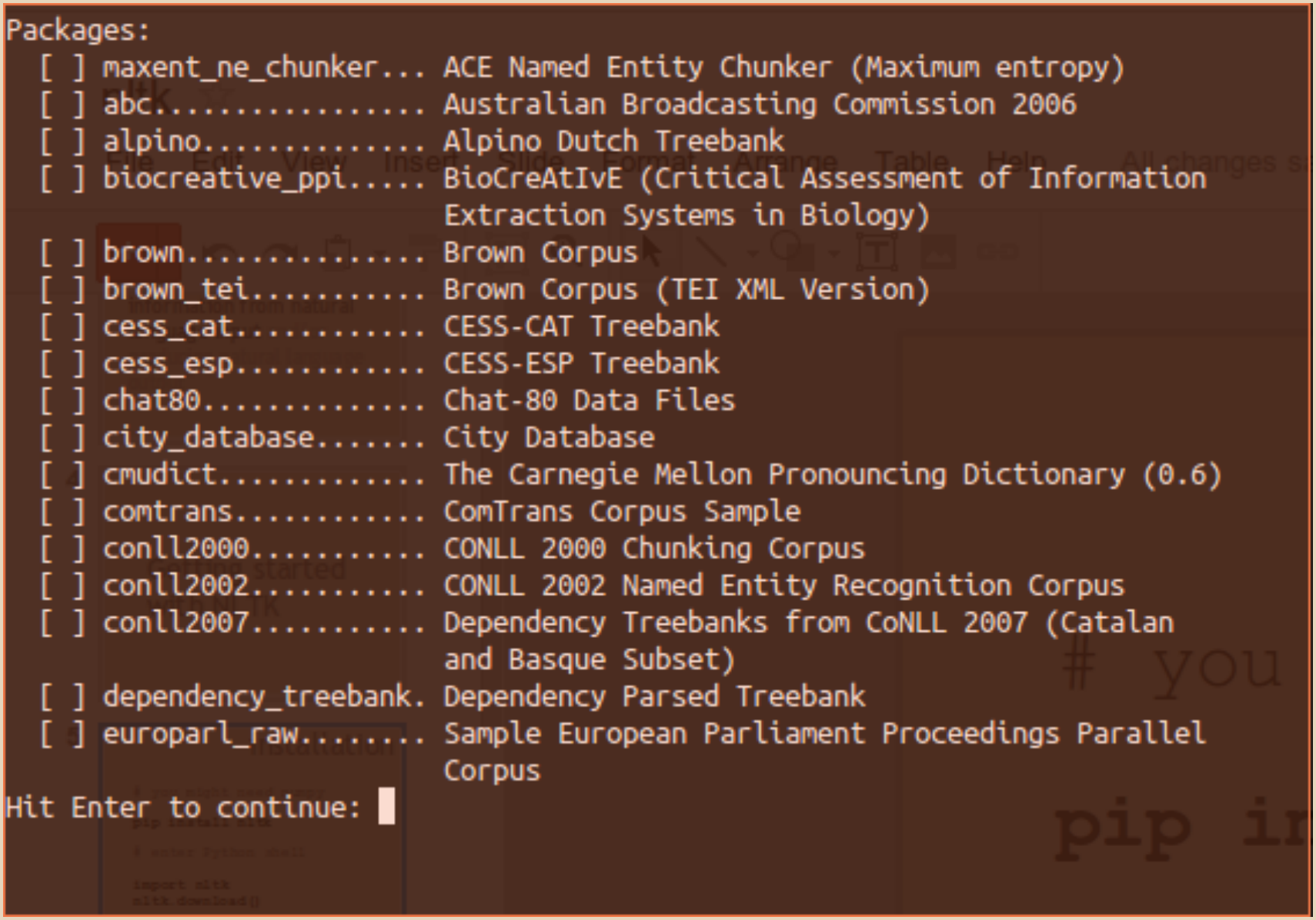
installation

```
# you might need numpy

pip install nltk

# enter Python shell

import nltk
nltk.download()
```


Packages:
[ ] maxent_ne_chunker... ACE Named Entity Chunker (Maximum entropy)
[ ] abc..... Australian Broadcasting Commission 2006
[ ] alpino..... Alpino Dutch Treebank
[ ] biocreative_ppi..... BioCreAtIvE (Critical Assessment of Information Extraction Systems in Biology)
[ ] brown..... Brown Corpus
[ ] brown_tei..... Brown Corpus (TEI XML Version)
[ ] cess_cat..... CESS-CAT Treebank
[ ] cess_esp..... CESS-ESP Treebank
[ ] chat80..... Chat-80 Data Files
[ ] city_database..... City Database
[ ] cmudict..... The Carnegie Mellon Pronouncing Dictionary (0.6)
[ ] comtrans..... ComTrans Corpus Sample
[ ] conll2000..... CONLL 2000 Chunking Corpus
[ ] conll2002..... CONLL 2002 Named Entity Recognition Corpus
[ ] conll2007..... Dependency Treebanks from CoNLL 2007 (Catalan and Basque Subset)
[ ] dependency_treebank. Dependency Parsed Treebank
[ ] europarl_raw..... Sample European Parliament Proceedings Parallel Corpus
Hit Enter to continue:

packages

```
# For Part of Speech tagging
maxent_treebank_pos_tagger

# Get a list of stopwords
stopwords

# Brown corpus to play around
brown
```

Preparing data / corpus

tokens

NLTK works on Tokens, for example,
"Hello World!" will be tokenized to:

```
['Hello', 'World', '!']
```

The built-in tokenizer for most use cases:

```
nltk.word_tokenize("Hello World!")
```

text

processing HTML text:

```
raw = nltk.clean_html(html_text)
tokens = nltk.word_tokenize(raw)
text = nltk.Text(tokens)
```

Use BeautifulSoup for preprocessing of the HTML text to discard unnecessary data.

Part-of-speech tagging

pos tagging

```
text = "Run away!"
nltk.word_tokenize(text)
nltk.pos_tag(tokens)

[('Run', 'NNP'),
 ('away', 'RB'),
 ('!', '.')]
```

pos tagging

```
[('Run', 'NNP'),
 ('away', 'RB'),
 ('!', '.')]
```

NNP: Proper Noun, Singular
RB : Adverb

http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

pos tagging

"The sailor dogs the barmaid."

```
[('The', 'DT'), ('sailor',
'NN'), ('dogs', 'NNS'), ('the',
'DT'), ('barmaid', 'NN'), ('.',
'.')]
```

Sentiment Analysis

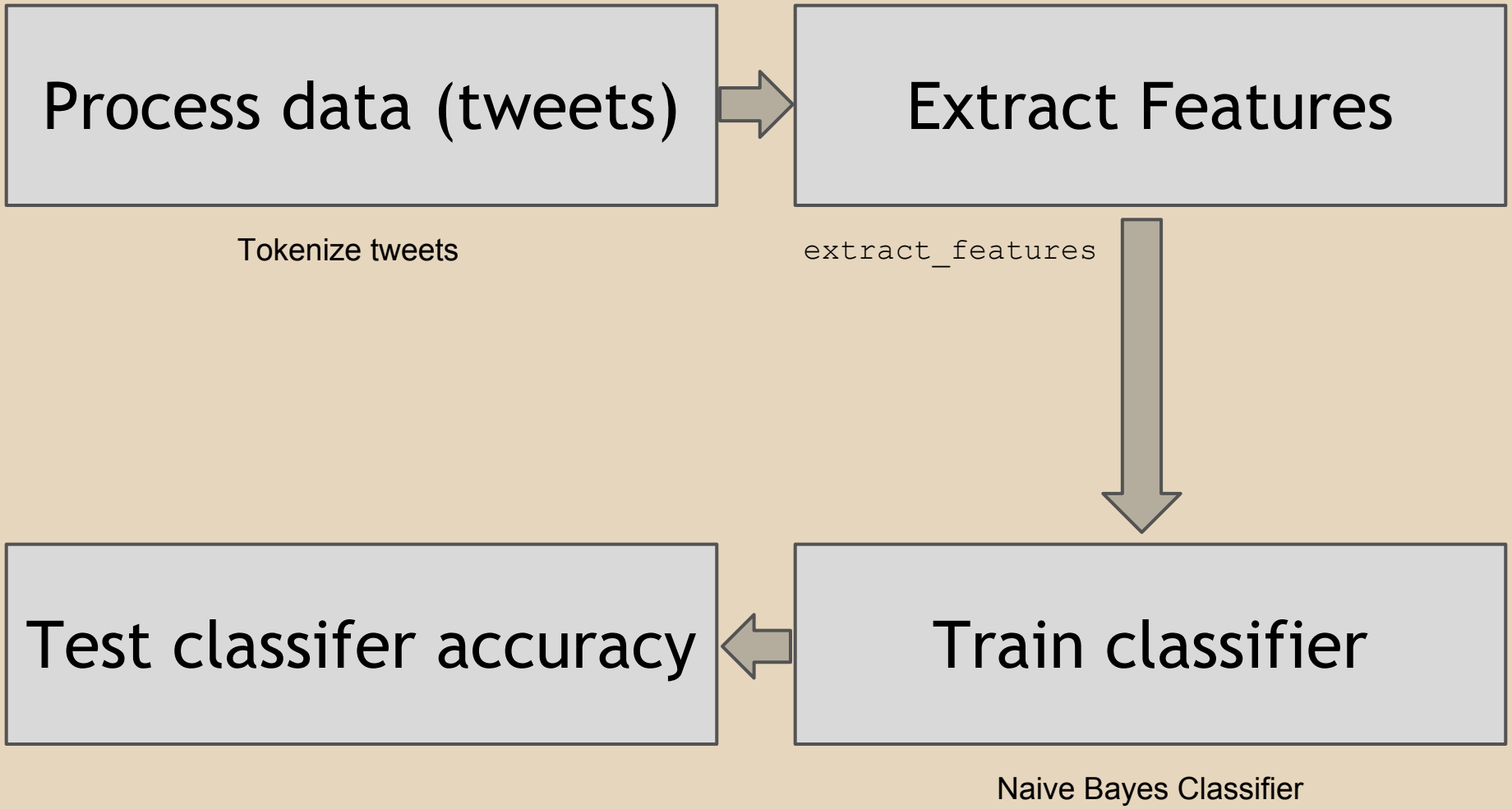
Code:
http://bit.ly/GLu2Q9

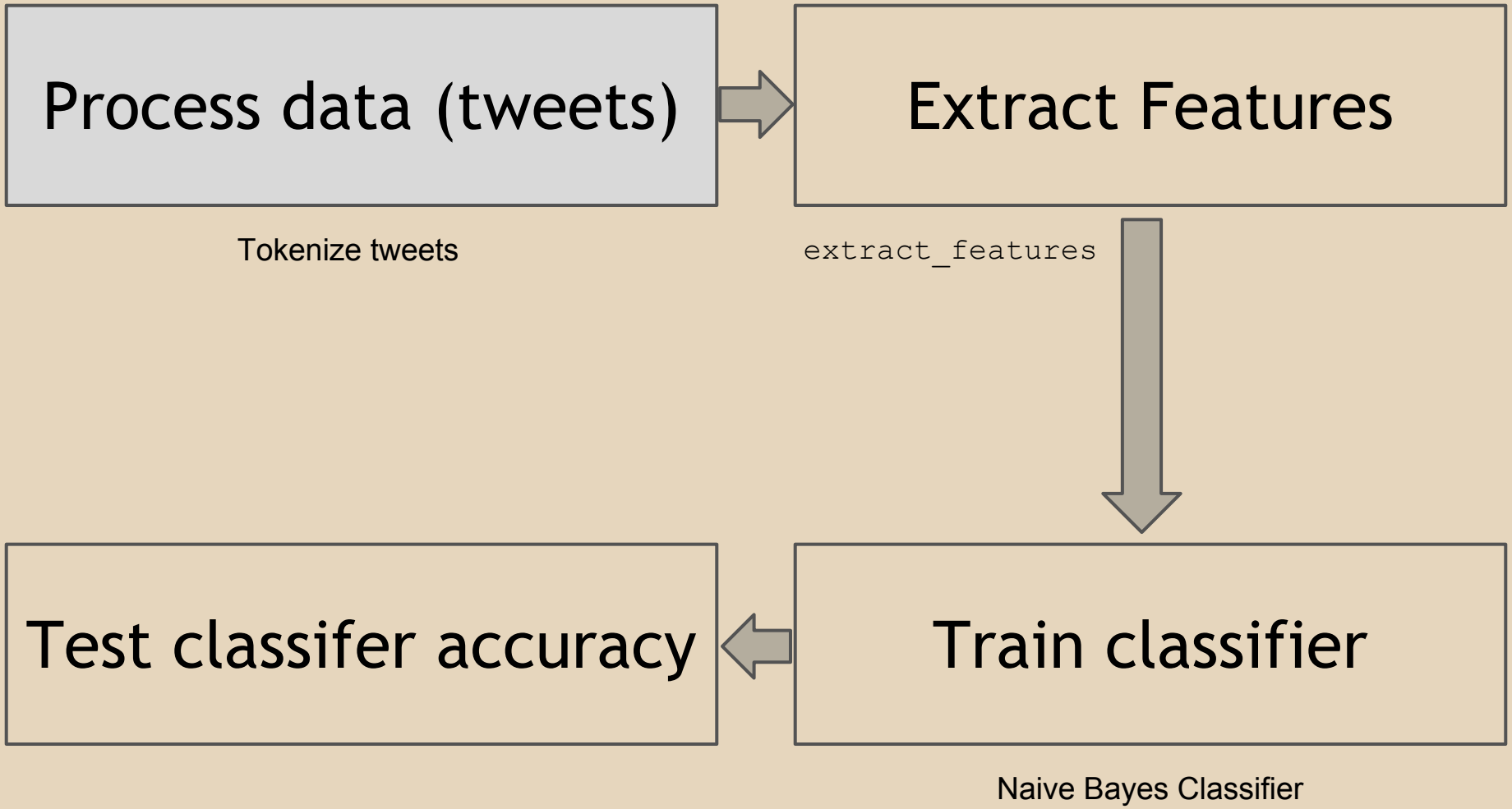
Differentiate between "happy" and "sad" tweets.

Teach the classifier the "features" of happy & sad tweets and test how good it is.

Happy:
"Looking through old pics and realizing everything happens for a reason. So happy with where I am right now"

Sad:
"So sad I have 8 AM class tomorrow"


Process data (tweets)
Extract Features
Tokenize tweets
extract_features
Test classifer accuracy
Train classifier
Naive Bayes Classifier


Process data (tweets)
Extract Features
Tokenize tweets
extract_features
Test classifer accuracy
Train classifier
Naive Bayes Classifier

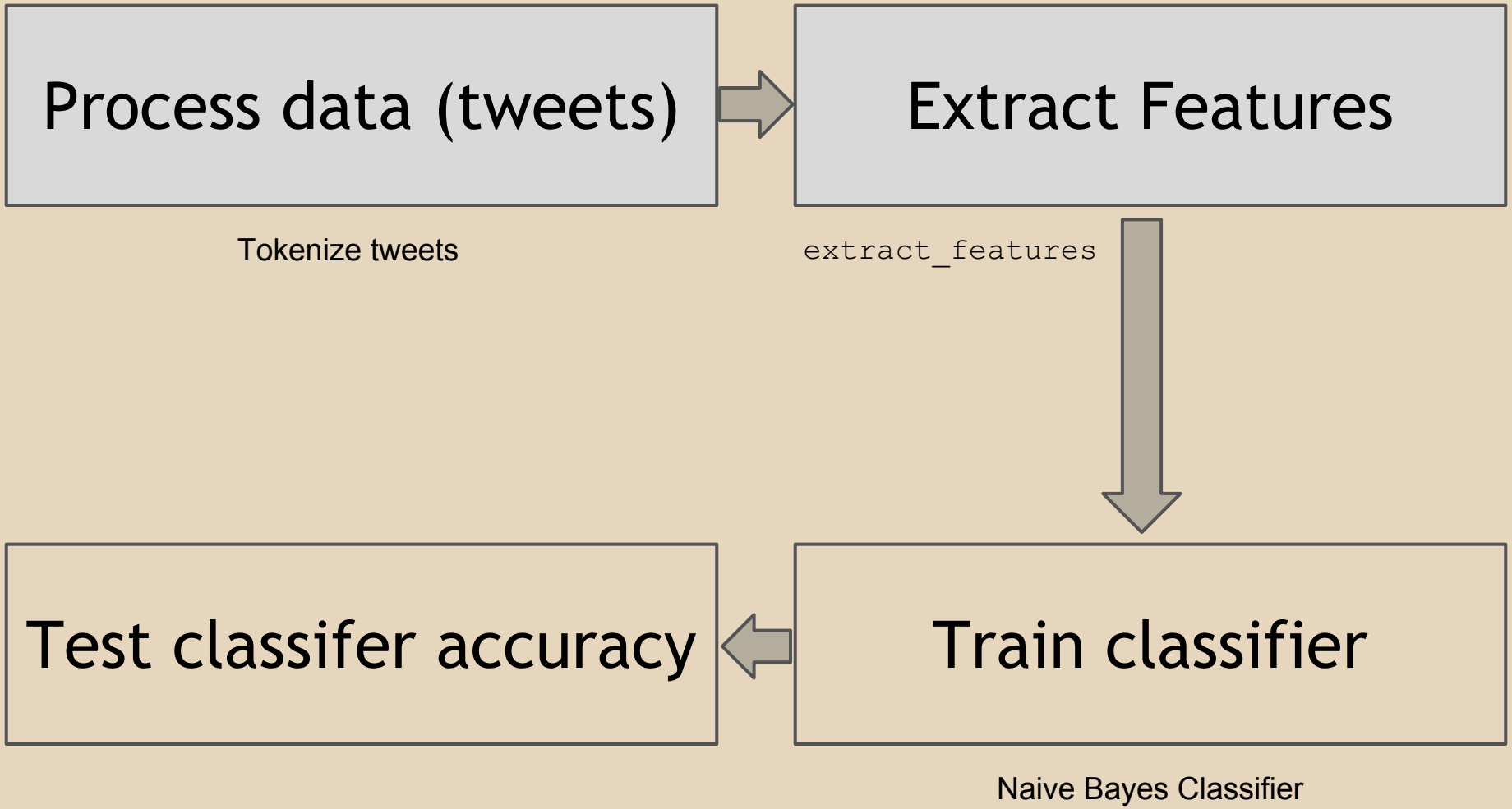
happy.txt
sad.txt

} training data

happy_test.txt
sad_test.txt

} testing data

Tweets obtained from Twitter Search API


Process data (tweets)
Extract Features
Tokenize tweets
extract_features
Test classifer accuracy
Train classifier
Naive Bayes Classifier

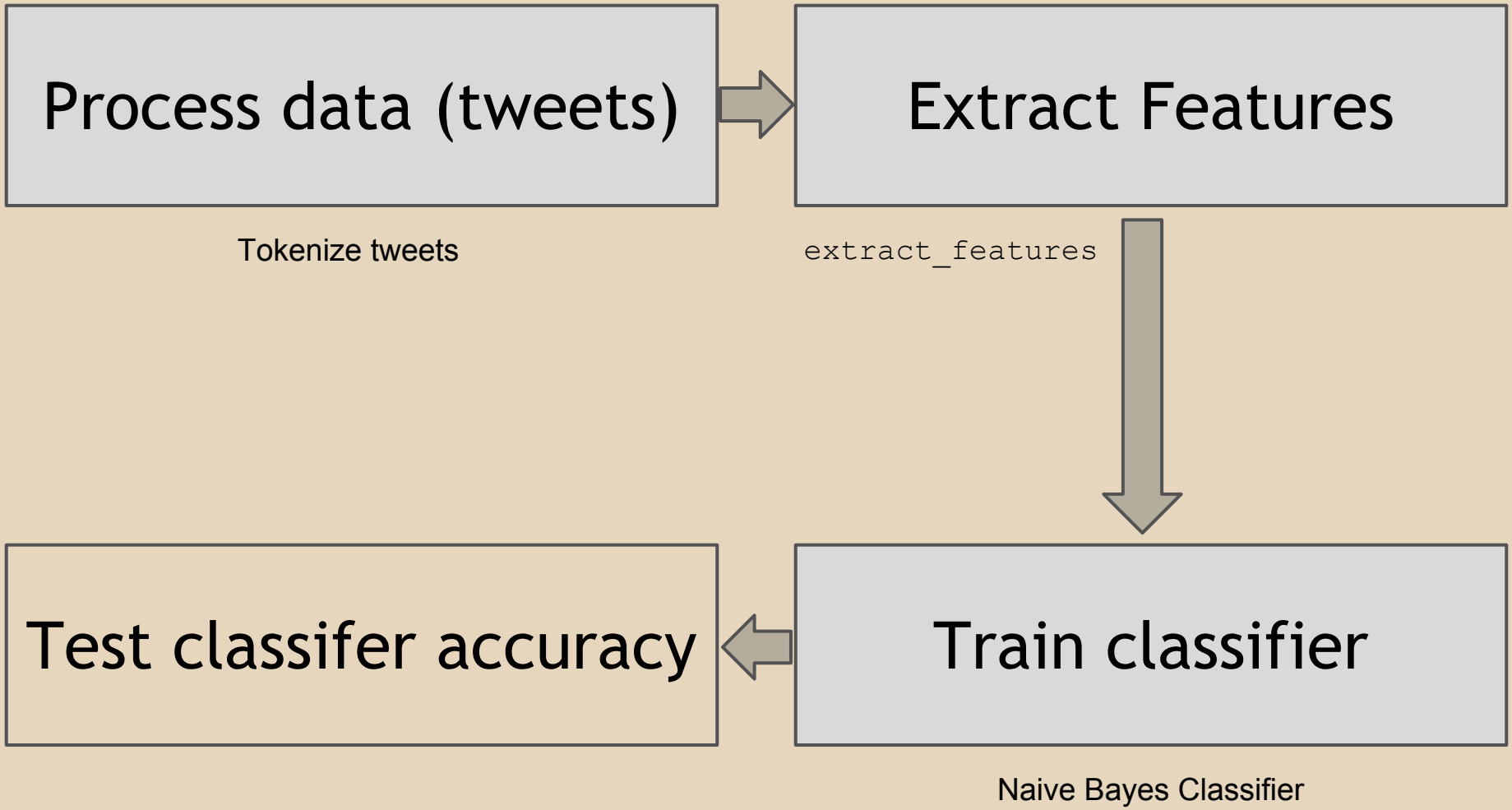
features

Happy tweets usually contain the following words:
"am happy", "great day" etc.

Sad tweets usually contain the following:
"not happy", "am sad" etc.

```
{'contains(not)': False,
 'contains(view)': False,
 'contains(best)': False,
 'contains(excited)': False,
 'contains(morning)': False,
 'contains(about)': False,
 'contains(horrible)': True,
 'contains(like)': False,
    ...
}
```

output of `extract_features()`

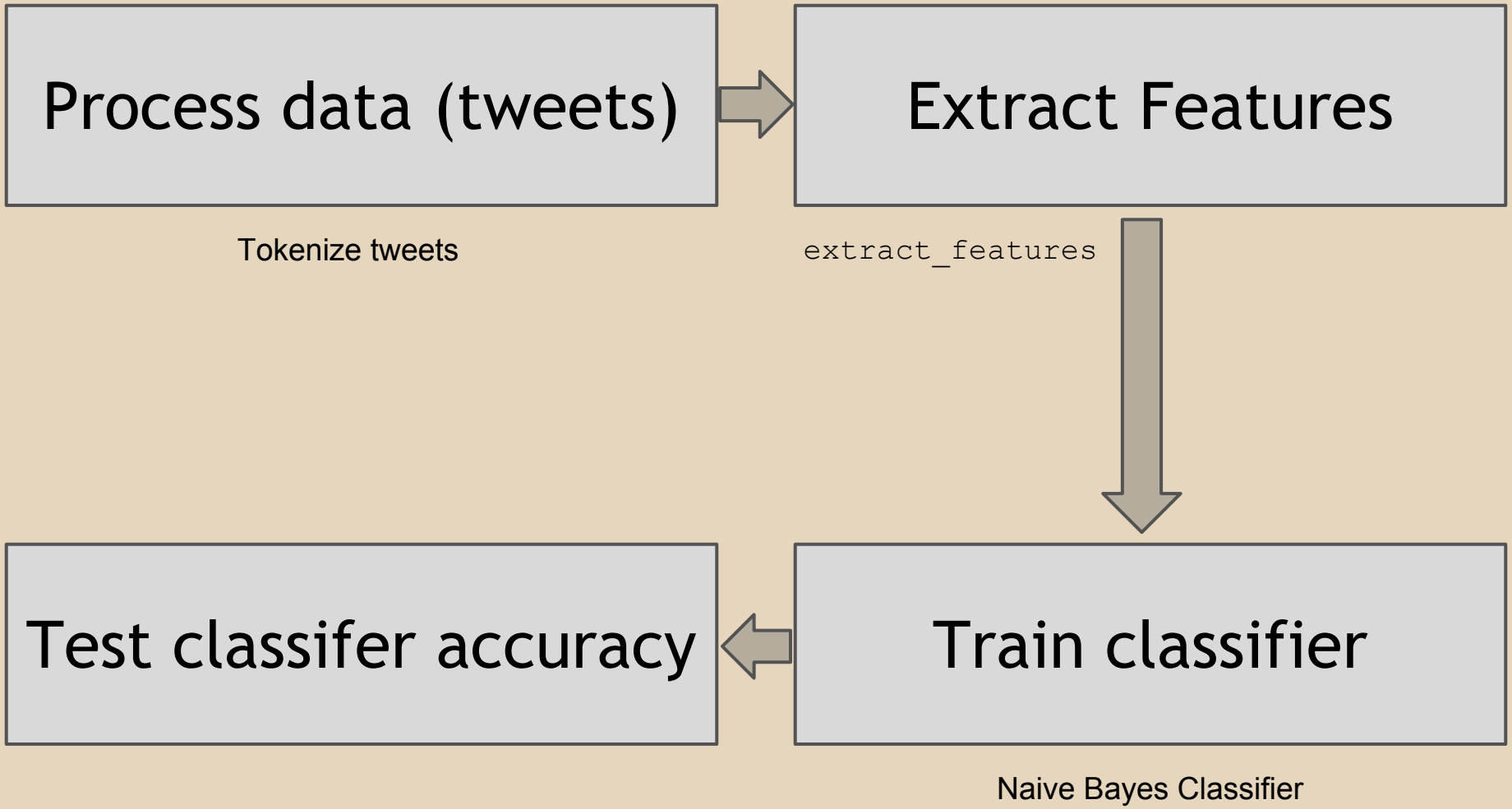

Process data (tweets)
Extract Features
Tokenize tweets
extract_features
Test classifer accuracy
Train classifier
Naive Bayes Classifier

training classifer

```
training_set = \
   nltk.classify.util.\
   apply_features(extract_features, tweets)

classifier = \ NaiveBayesClassifier.train
(training_set)
```

training the classifer


Process data (tweets)
Extract Features
Tokenize tweets
extract_features
Test classifer accuracy
Train classifier
Naive Bayes Classifier

testing classifer

```
def classify_tweet(tweet):
    return \
     classifier.classify(extract_features
(tweet))
```

```
$ python classification.py
Total accuracy: 90.00% (18/20)
```

18 tweets got classified correctly.

Where to go from here.

Navigation

NLTK Home
Book
Code
Contribute
Courses
Data
Documentation
Download
FAQ
Forums
Getting Started
HOWTO
News Archive
Projects
Quotes
Teaching

Software License

Apache License 2.0

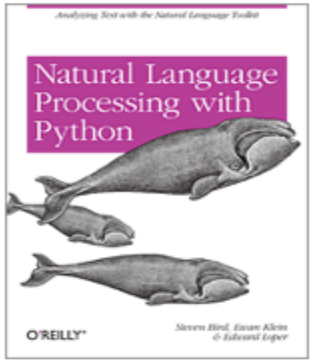
Website License

Creative Commons
Attribution
Noncommercial
No Derivative Works
3.0 US License

Feedback

This site is maintained by Steven Bird.

Book



Natural Language Processing with Python
--- Analyzing Text with the Natural Language Toolkit

Steven Bird, Ewan Klein, and Edward Loper

O'Reilly Media, 2009 | Sellers and prices | Request inspection copy

0. Preface (extras)
1. Language Processing and Python (extras)
2. Accessing Text Corpora and Lexical Resources (extras)
3. Processing Raw Text
4. Writing Structured Programs (extras)
5. Categorizing and Tagging Words
6. Learning to Classify Text (extras)
7. Extracting Information from Text
8. Analyzing Sentence Structure (extras)
9. Building Feature Based Grammars
10. Analyzing the Meaning of Sentences (extras)
11. Managing Linguistic Data
12. Afterword: Facing the Language Challenge

Bibliography
Term Index

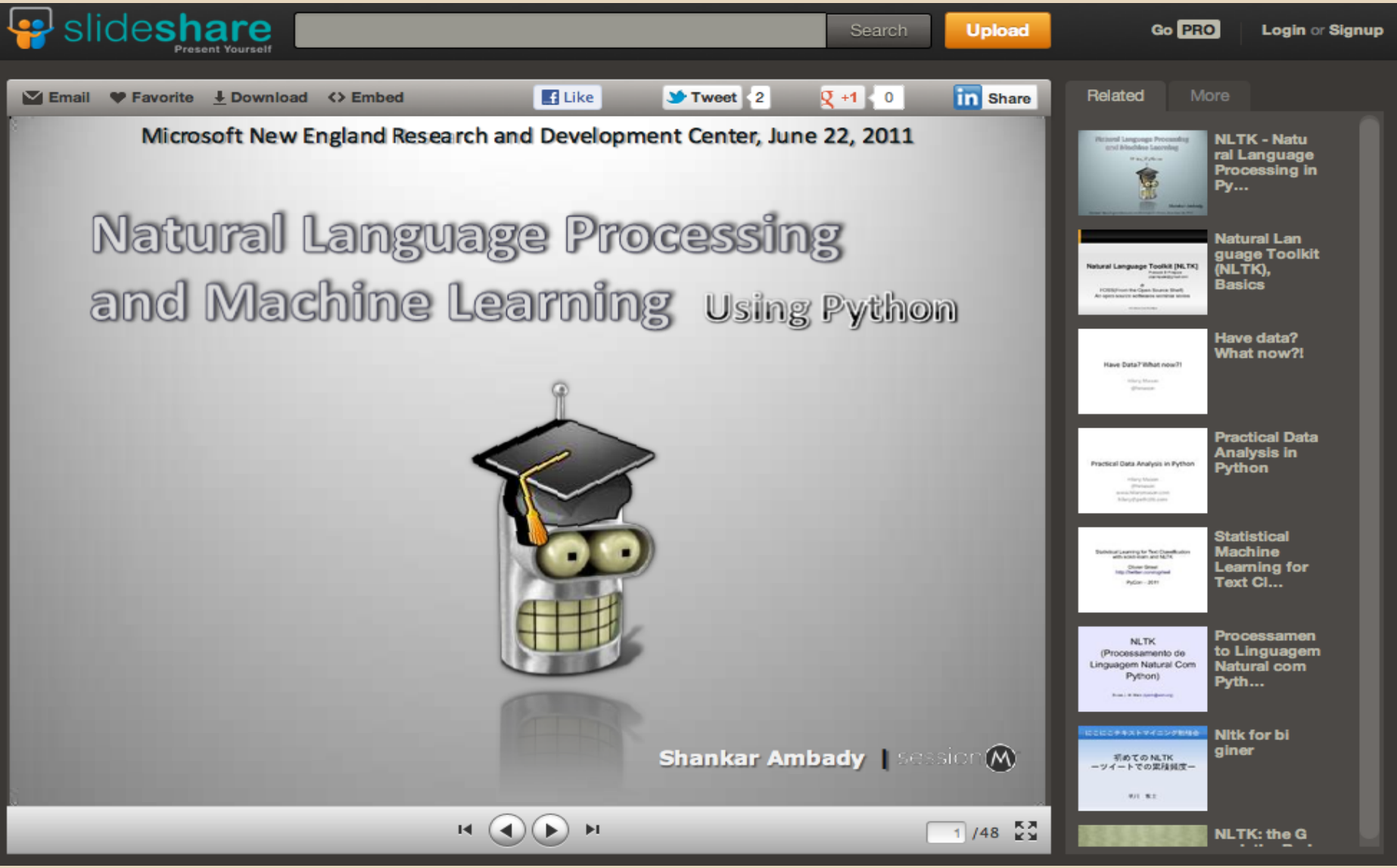
Errata (corrected here, and in second printing of book, available in January 2010)

Translations: Book (jp), Prefácio (pt), Przedmowa (pl)

Reviews: LanguageLog, Amazon.com, Slashdot.org, Dr Dobbs

http://www.nltk.org/book

https://class.coursera.org/nlp/auth/welcome



http://www.slideshare.net/shanbady/nltk-boston-text-analytics

```
[('Thank', 'NNP'),
 ('you', 'PRP'),
 ('.', '.')]
```

@victorneo