# Natural Language Processing with Python & nltk Cheat Sheet by murenei

A quick reference guide for basic (and more advanced) natural language processing tasks in Python, using mostly nltk (the Natural Language Toolkit package), including POS tagging, lemmatizing, sentence parsing and text classification.

## Handling Text

| | |
|---|---|
| `text='Some words'` | assign string |
| `list(text)` | Split text into character tokens |
| `set(text)` | Unique tokens |
| `len(text)` | Number of characters |

## Accessing corpora and lexical resources

| | |
|---|---|
| `from nltk.corpus import brown` | import CorpusReader object |
| `brown.words(text_id)` | Returns pretokenised document as list of words |
| `brown.fileids()` | Lists docs in Brown corpus |
| `brown.categories()` | Lists categories in Brown corpus |

## Tokenization

| | |
|---|---|
| `text.split(" ")` | Split by space |
| `nltk.word_tokenizer(text)` | nltk in-built word tokenizer |
| `nltk.sent_tokenize(doc)` | nltk in-built sentence tokenizer |

## Sentence Parsing

| | |
|---|---|
| `g=nltk.data.load('grammar.cfg')` | Load a grammar from a file |
| `g=nltk.CFG.fromstring("""...""")` | Manually define grammar |
| `parser=nltk.ChartParser(g)` | Create a parser out of the grammar |
| `trees=parser.parse_all(text)` | |
| `for tree in trees: ... print tree` | |
| `from nltk.corpus import treebank` | |
| `treebank.parsed_sents('wsj_0001.mrg')` | Treebank parsed sentences |

## Text Classification

`from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer`

| | |
|---|---|
| `vect=CountVectorizer().fit(X_train)` | Fit bag of words model to data |
| `vect.get_feature_names()` | Get features |
| `vect.transform(X_train)` | Convert to doc-term matrix |

| | |
|---|---|
| input="List listed lists listing listings" | Different suffixes |
| words=input.lower().split(' ') | Normalize (lower-case) words |
| porter=nltk.PorterStemmer | Initialise Stemmer |
| [porter.stem(t) for t in words] | Create list of stems |
| WNL=nltk.WordNetLemmatizer() | Initialise WordNet lemmatizer |
| [WNL.lemmatize(t) for t in words] | Use the lemmatizer |

## Part of Speech (POS) Tagging

| | |
|---|---|
| nltk.help.upenn_tagset('MD') | Lookup definition for a POS tag |
| nltk.pos_tag(words) | nltk in-built POS tagger |
| | <use an alternative tagger to illustrate ambiguity> |

| | |
|---|---|
| g="NP: {<DT>?<JJ>*<NN>}" | Regex chunk grammar |
| cp=nltk.RegexpParser(g) | Parse grammar |
| ch=cp.parse(pos_sent) | Parse tagged sent. using grammar |
| print(ch) | Show chunks |
| ch.draw() | Show chunks in IOB tree |
| cp.evaluate(test_sents) | Evaluate against test doc |

```
sents=nltk.corpus.treebank.tagged_sents()
```

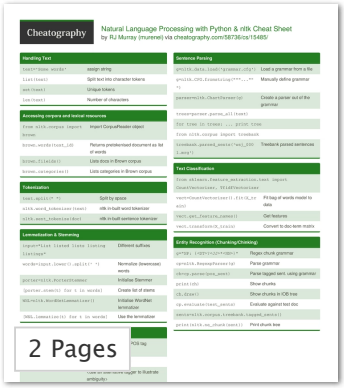
| | |
|---|---|
| print(nltk.ne_chunk(sent)) | Print chunk tree |

## RegEx with Pandas & Named Groups

```
df=pd.DataFrame(time_sents, columns=['text'])
df['text'].str.split().str.len()
df['text'].str.contains('word')
df['text'].str.count(r'\d')
df['text'].str.findall(r'\d')
df['text'].str.replace(r'\w+day\b', '???')
df['text'].str.replace(r'(\w)', lambda x: x.groups()[0][:3])
df['text'].str.extract(r'(\d?\d):(\d\d)')
df['text'].str.extractall(r'((\d?\d):(\d\d) ?([ap]m))')
df['text'].str.extractall(r'(?P<digits>\d)')
```

python nltk nlp

## Download the Natural Language Processing with Python & nltk Cheat Sheet




PDF (recommended)

PDF (2 pages)

Alternative Downloads

PDF (black and white)

LaTeX

## Comments

No comments yet. Add yours below!

## Created By

murenei

https://tutify.com.au

## Add a Comment

Your Name

Your Email Address

Your Comment

Post Your Comment

# Metadata

Languages: English

Published: 28th May, 2018

Last Updated: 29th May, 2018

Rated: 5 stars based on 6 ratings

# Favourited By

# Related Cheat Sheets

Python 3 Cheat Sheet
by Finxter

NLP For Arabic Cheat Sheet

# More Cheat Sheets by murenei

★★★★☆

Network Analysis with Python and NetworkX Cheat Sheet

## Latest Cheat Sheet

1 Page

☆☆☆☆☆ (0)

### AP Physics Formulas (Kinematic) Cheat Sheet

Some physics formulas that will be useful in kinematics. Not a truly complete list of formulas though, as some things are missing.I can't think of any more formulas for this cheat sheet though, so suggestions on what to add would be helpful.

ReSummit
23 Oct 20
physics

## Random Cheat Sheet

1 Page

★★★★★ (2)

### FREQUENTLY USED DX CODES Cheat Sheet

List of diagnosis frequently used on admission to rehab and long term care facilities.

charlesnurse
9 May 13, updated 12 May 16
nursing, coding, nurse, admission, icd-9 and 3 more ...

## About Cheatography

## Behind the Scenes

## Recent Activity

references in 25 languages for everything from travel to maths!

right here:

DaveChild SpaceDuck

Cheatography

Sheet.
1 day 9 hours ago

ReSummit published AP Physics Formulas (Kinematic).
1 day 11 hours ago

cjdvslee updated UTS.
1 day 18 hours ago