PATTERN

* The Pattern library is a multipurpose library capable of handling the following tasks:
* Natural Language Processing: Performing tasks such as tokenization, stemming, POS tagging, sentiment analysis, etc.
* Data Mining: It contains APIs to mine data from sites like Twitter, Facebook, Wikipedia, etc.
* Machine Learning: Contains machine learning models such as SVM, KNN, and perceptron, which can be used for classification, regression, and clustering tasks.

**Installing the Library**

pip install pattern/conda install -c asmeurer pattern

**Tokenizing, POS Tagging, and Chunking**

* In the NLTK and spaCy libraries, we have a separate function for tokenizing, POS tagging, and finding noun phrases in text documents. On the other hand, in the Pattern library there is the all-in-one parse method that takes a text string as an input parameter and returns corresponding tokens in the string, along with the POS tag.
* The parse method also tells us if a token is a noun phrase or verb phrase, or subject or object. You can also retrieve lemmatized tokens by setting lemmata parameter to True
* The syntax of the parse method along with the default values for different parameters is as follows:

parse(string,

tokenize=True, # Split punctuation marks from words?

tags=True, # Parse part-of-speech tags? (NN, JJ, ...)

chunks=True, # Parse chunks? (NP, VP, PNP, ...)

relations=False, # Parse chunk relations? (-SBJ, -OBJ, ...)

lemmata=False, # Parse lemmata? (ate => eat)

encoding='utf-8', # Input string encoding.

tagset=None # Penn Treebank II (default) or UNIVERSAL. )

**Parse method**

* To use the parse method, you have to import the en module from the pattern library.
* The en module contains English language NLP functions. If you use the pprint method to print the output of the parse method on the console
* split method on the object returned by the parse method, the output will be a list of sentences, where each sentence is a list of tokens and each token is a list of words, along with the tags associated with the words.

**Pluralizing and Singularizing the Tokens**

The  pluralize and singularize methods are used to convert singular words to plurals and vice versa, respectively.

#### Converting Adjective to Comparative and Superlative Degrees

You can retrieve comparative and superlative degrees of an adjective using comparative and superlative function. For instance, the comparative degree of good is better and the superlative degree of good is best.

#### Finding N-Grams

* N-Grams refer to "n" combination of words in a sentence. For instance, for the sentence "He goes to hospital", 2-grams would be (He goes), (goes to) and (to hospital). N-Grams can play a crucial role in text classification and language modeling.
* In the Pattern library, the ngram method is used to find the all the n-grams in a text string. The first parameter to the ngram method is the text string. The number of n-grams is passed to the n parameter of the method.

#### Finding Sentiments

* Sentiment refers to an opinion or feeling towards a certain thing. The Pattern library offers functionality to find sentiment from a text string.
* In Pattern, the sentiment object is used to find the polarity (positivity or negativity) of a text along with its subjectivity.
* Depending upon the most commonly occurring positive (good, best, excellent, etc.) and negative (bad, awful, pathetic, etc.) adjectives, a sentiment score between 1 and -1 is assigned to the text. This sentiment score is also called the polarity.
* In addition to the sentiment score, subjectivity is also returned. The subjectivity value can be between 0 and 1. Subjectivity quantifies the amount of personal opinion and factual information contained in the text. The higher subjectivity means that the text contains personal opinion rather than factual information.
* (0.75, 0.8)
* The sentence "This is an excellent movie to watch. I really love it" has a sentiment of 0.75, which shows that it is highly positive. Similarly, the subjectivity of 0.8 refers to the fact that the sentence is a personal opinion of the user.

#### Checking if a Statement is a Fact/Modality

#### The modality function from the Pattern library can be used to find the degree of certainty in the text string. The modality function returns a value between -1 to 1. For facts, the modality function returns a value greater than 0.5.Similarly, for a sentence which is not certain, the value returned by the modality method is around 0.0.

#### Spelling Corrections

#### The suggest method can be used to find if a word is spelled correctly or not. The suggest method returns 1 if a word is 100% correctly spelled. Otherwise the suggest method returns the possible corrections for the word along with their probability of correctness.

#### Suggest function is used for spelling corrections but it is more than that. It not only checks the spelling it also gives you suggestions of what might be the correct word with their probabilities. This function also distinguishes pattern from other libraries.

#### Working with Numbers

#### The Pattern library contains functions that can be used to convert numbers in the form of text strings into their numeric counterparts and vice versa. To convert from text to numeric representation the number function is used. Similarly to convert back from numbers to their corresponding text representation the numerals function is used.for numerals function we have to provide the integer value that we want our number to be rounded-off to.

**Quantify**

The quantify function is used to get a word count estimation of the items in the list, which provides a phrase for referring to the group. If a list has 3-8 similar items, the quantify function will quantify it to "several". Two items are quantified to a "couple".

Quantify function is used to provide a word count estimation of the words given.

**Pattern Library Functions for Data Mining**

The web module of the Pattern library is used for web mining tasks.

**Accessing Web Pages**

The URL object is used to retrieve contents from the web pages. It has several methods that can be used to open a webpage, download the contents from a webpage and read a webpage. You can directly use the download method to download the HTML contents of any webpage.

**Finding URLs within Text**

You can use the findurl method to extract URLs from text strings.

**Making Asynchronous Requests for Web pages**

* Web pages can be very large and it can take quite a bit of time download the complete contents of the webpage, which can block a user from performing any other task on the application until the complete webpage is downloaded.
* However, the web module of the Pattern library contains a function asynchronous, which downloads contents of a webpage in a parallel manner. The asynchronous method runs in the background so that the user can interact with the application while the webpage is being downloaded.

**Getting Search Engine Results with APIs**

The pattern library contains Search Engine class which is derived by the classes that can be used to connect to call API's of different search engines and websites such as Google, Bing, Face book, Wikipedia, Twitter, etc. The Search Engine object construct accepts three parameters:

* license: The developer license key for the corresponding search engine or website
* throttle: Corresponds to the time difference between successive request to the server
* langauge: Specifies the language for the results

The search method of the SearchEngine class is used to make a request to search engine for certain search query. The search method can take the following parameters:

* query: The search string
* type: The type of data you want to search, it can take three values: SEARCH, NEWS and IMAGE.
* start: The page from which you want to start the search
* count: The number of results per page.

The search engine classes that inherit the SearchEngine class along with its search method are: Google, Bing, Twitter, Facebook, Wikipedia, and Flickr.

The search query returns objects for each item. The result object can then be used to retrieve the information about the searched result. The attributes of the result object are url, title, text, language, author, date.

The process is similar for Bing search engine, you only have to replace the Bing class with Google in the script above.

# Converting HTML Data to Plain Text

The download method of the URL class returns data in the form of HTML. However, if you want to do a semantic analysis of the text, for instance, sentiment classification, you need data cleaned data without HTML tags. You can clean the data with the plaintext method. The method takes as a parameter, the HTML content returned by the download method, and returns cleaned text.

# Parsing PDF Documents

The Pattern library contains PDF object that can be used to parse a PDF document. PDF (Portable Document Format) is a cross platform file which contains images, texts, and fonts in a stand-alone document.

# Clearing the Cache

The results returned by the methods such as SearchEngine.search() and URL.download() are, by default, stored in the local cache. To clear the cache after downloading an HTML document, we can use clear method of the cache class

The Pattern library is one of the most useful natural language processing libraries in Python. Although it is not as well-known as spaCy or NLTK, it contains functionalities such as finding superlatives and comparatives, and fact and opinion detection which distinguishes it from the other NLP libraries.

The whole package consists of six main modules:

* pattern.web: A toolkit that includes APIs for various Web services, including Google, Gmail, Bing, Twitter Wikipedia and Flickr. It has its own HTML parser and Web Crawler.
* pattern.table: A module for working with tabular data, used for storing data from the pattern.web module.
* pattern.en: A natural language processing toolkit for English.pattern.search: A module containing a search algorithm.
* pattern.vector: A module containing various tools for analyzing the text of a document.
* pattern.graph: A module for data visualization using Canvas

# Flickr Mining

Flickr is an American image hosting and video hosting service, as well as an online community. Pattern can be used to extract data from Flickr.

# Pattern Modules Pattern includes a lot of modules:

* [pattern.web](http://www.clips.ua.ac.be/pages/pattern-web)
* [pattern.db](http://www.clips.ua.ac.be/pages/pattern-db)
* [pattern.en](http://www.clips.ua.ac.be/pages/pattern-en) | [es](http://www.clips.ua.ac.be/pages/pattern-es) | [de](http://www.clips.ua.ac.be/pages/pattern-de) | [fr](http://www.clips.ua.ac.be/pages/pattern-fr) | [it](http://www.clips.ua.ac.be/pages/pattern-it) | [nl](http://www.clips.ua.ac.be/pages/pattern-nl)
* [pattern.search](http://www.clips.ua.ac.be/pages/pattern-search)
* [pattern.vector](http://www.clips.ua.ac.be/pages/pattern-vector)
* [pattern.graph](http://www.clips.ua.ac.be/pages/pattern-graph)
* [pattern.metrics](http://www.clips.ua.ac.be/pages/pattern-metrics)
* [canvas.js](http://www.clips.ua.ac.be/pages/pattern-canvas)

**Text Analysis Module pattern.en**

The pattern.en module can be used for English [text processing](http://textprocessing.org/), like word tokenize, sentence tokenize(sentence segmentation), pos tagging (part-of-speech tagger), english grammar(Indefinite article, Pluralization + singularization, Comparative + superlative), [sentiment analysis](http://sentimentanalysis.info/), Parser(identifies nouns, adjectives..), WordNet interface and etc.

[**Word Tokenization**](http://textanalysisonline.com/pattern-word-tokenize)

The tokenize methods in pattern.en supports word tokenization and sentence tokenization for english text

[**POS Tagging**](http://textanalysisonline.com/pattern-pos-tagging)

The tag method from pattern.en can be used POS Tagging for english text, the part-of-speech tags is followed Penn Treebank II tag set, can be found the details here: [Penn Treebank II tag set](http://www.clips.ua.ac.be/pages/mbsp-tags)

[**Indefinite Article**](http://textanalysisonline.com/pattern-indefinite-article)

In pattern.en module, you can use the referenced method to get the indefinite article