542 - GROUP 21

PYTEXTPREP

A Python package for preprocessing tweet data

Why Tweets?

LARGE USERBASE
206M Active Users

CURRENT TRENDS
Viral #HASHTAGS

CHARACTER LIMIT
Tweets tend to be straight to the point

EASY TO SCRAPE
Public API to retrieve tweets

Why PyTextPrep?

ONE SOLUTION
FOR YOUR TWEET DATA

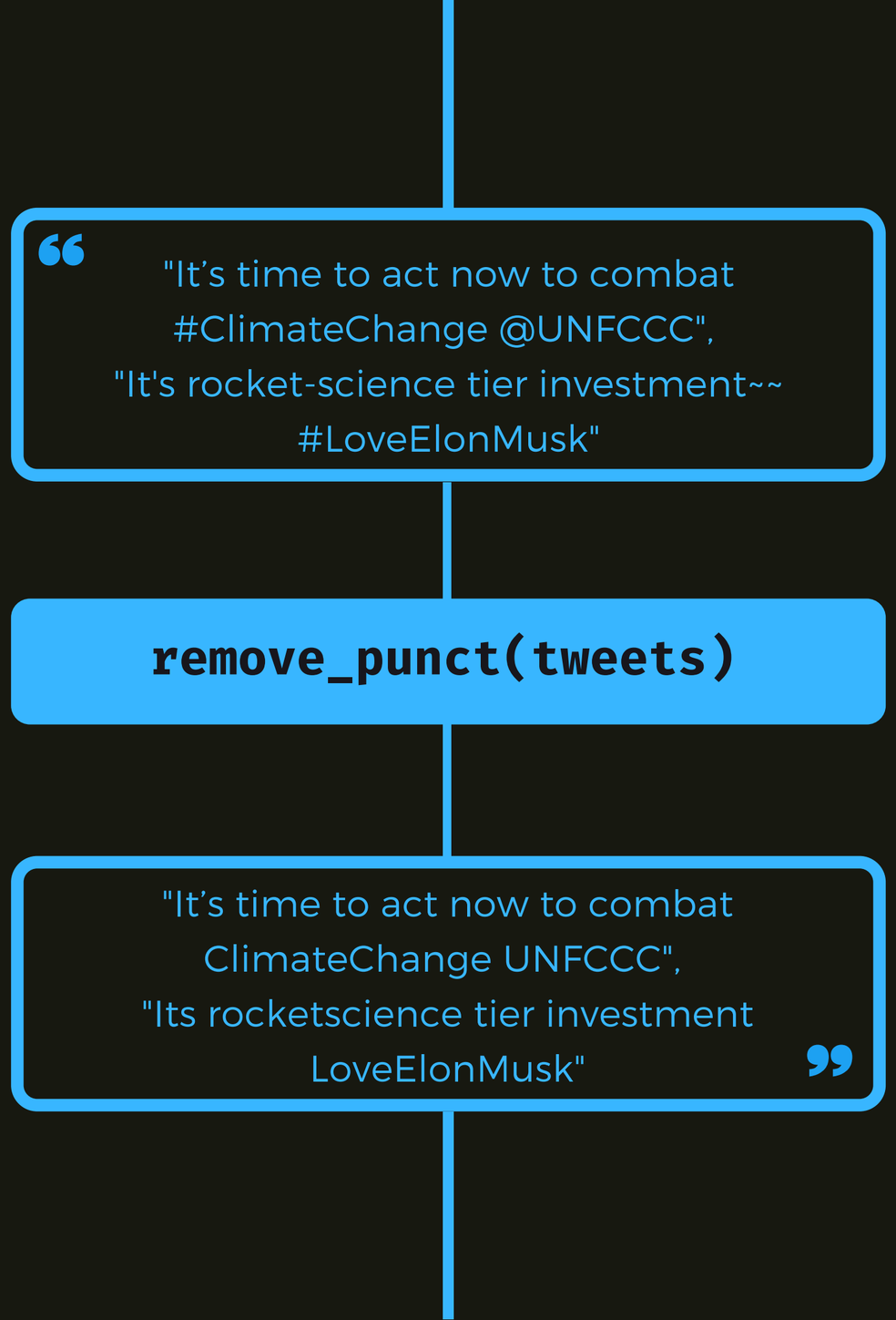
	PYTEXTPREP	OTHERS*
DATA CLEANING	✓	✓
FEATURE ENGINEERING	✓	✗
DATA VISUALIZATION	✓	✗

*compared to tweet_preprocessor

Data Cleaning

`remove_punct()`

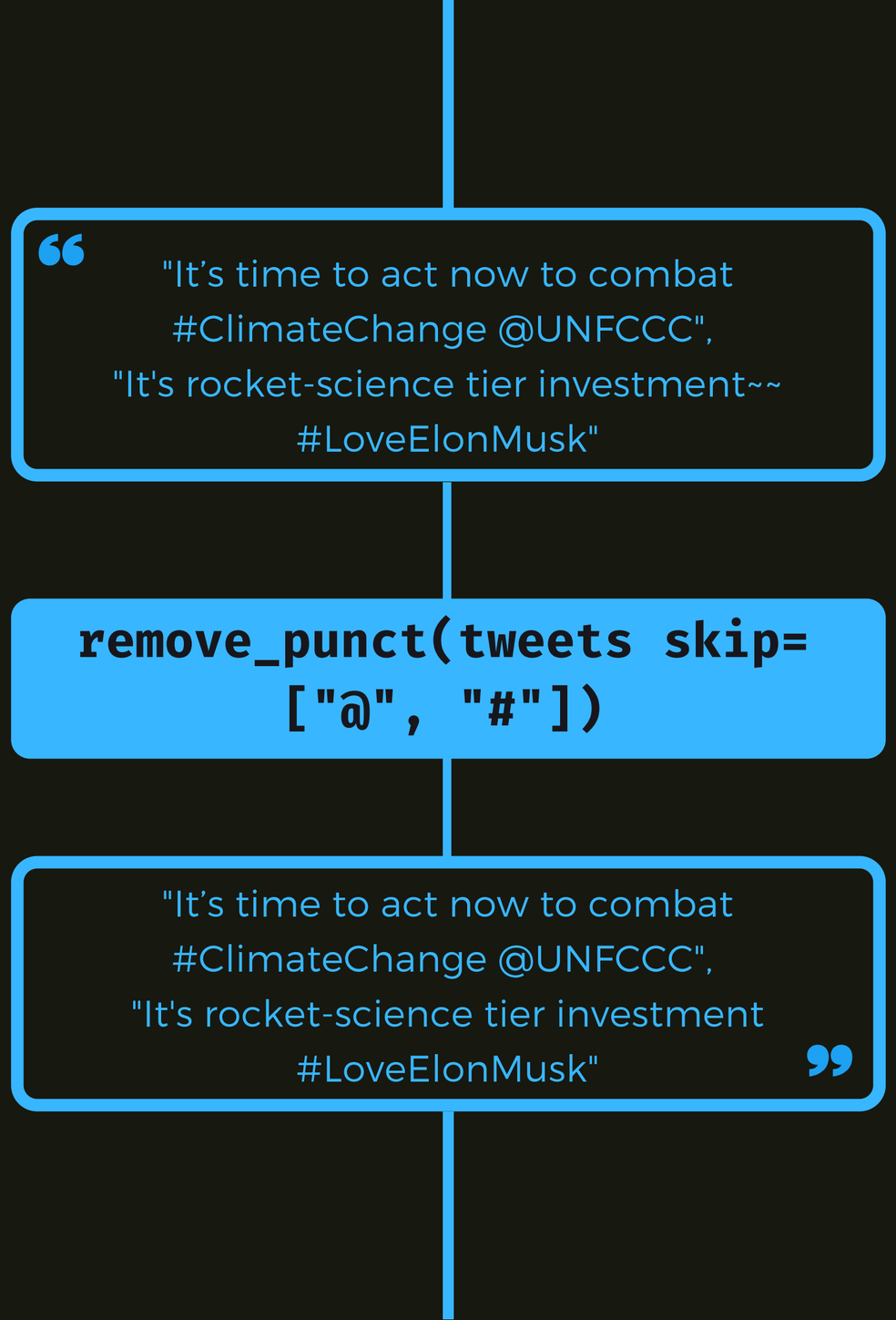
Easily remove punctuation from a list of tweets



Data Cleaning

`remove_punct()`

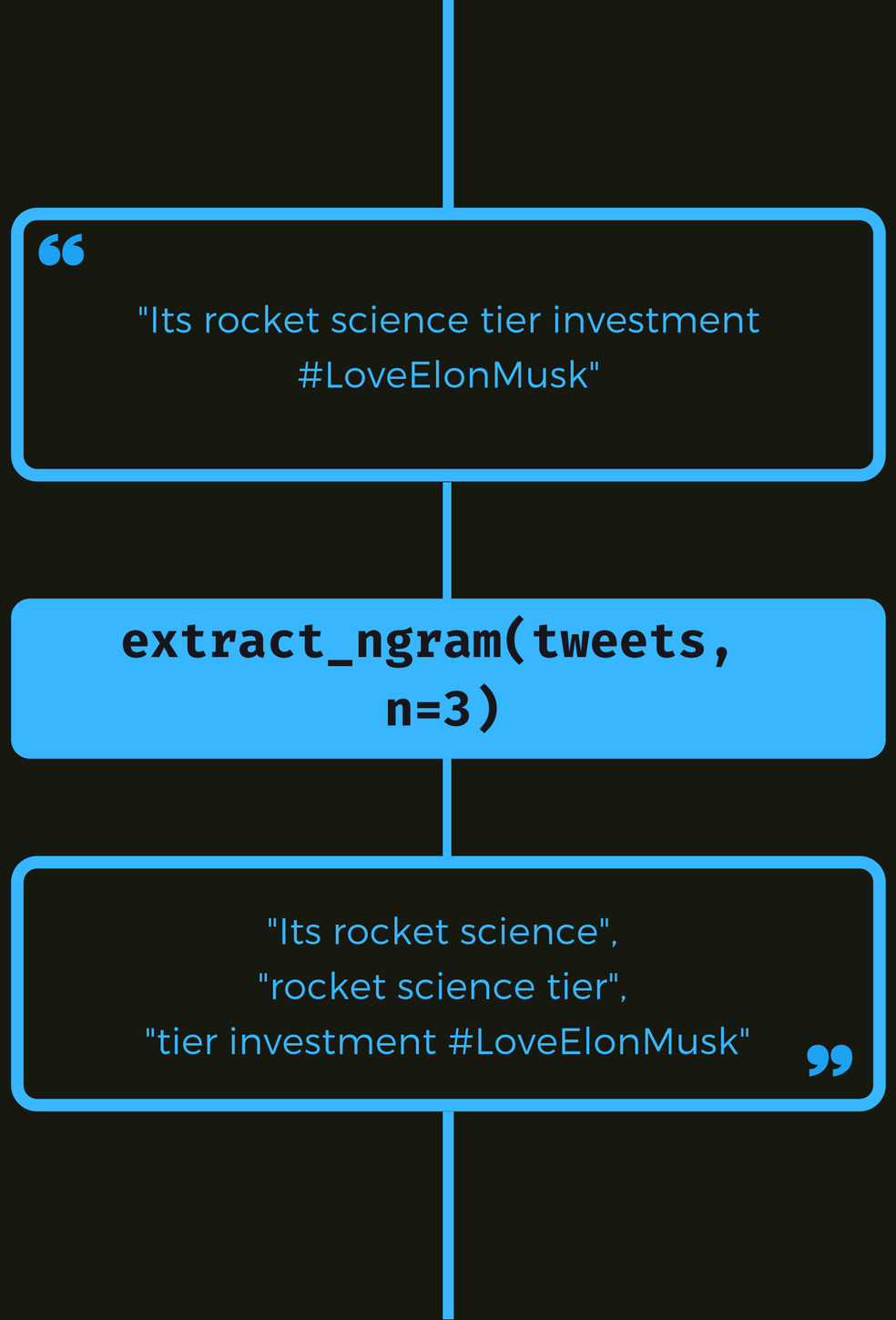
Use the skip parameter for selecting which elements should be maintained



Feature Engineering

`extract_ngram()`

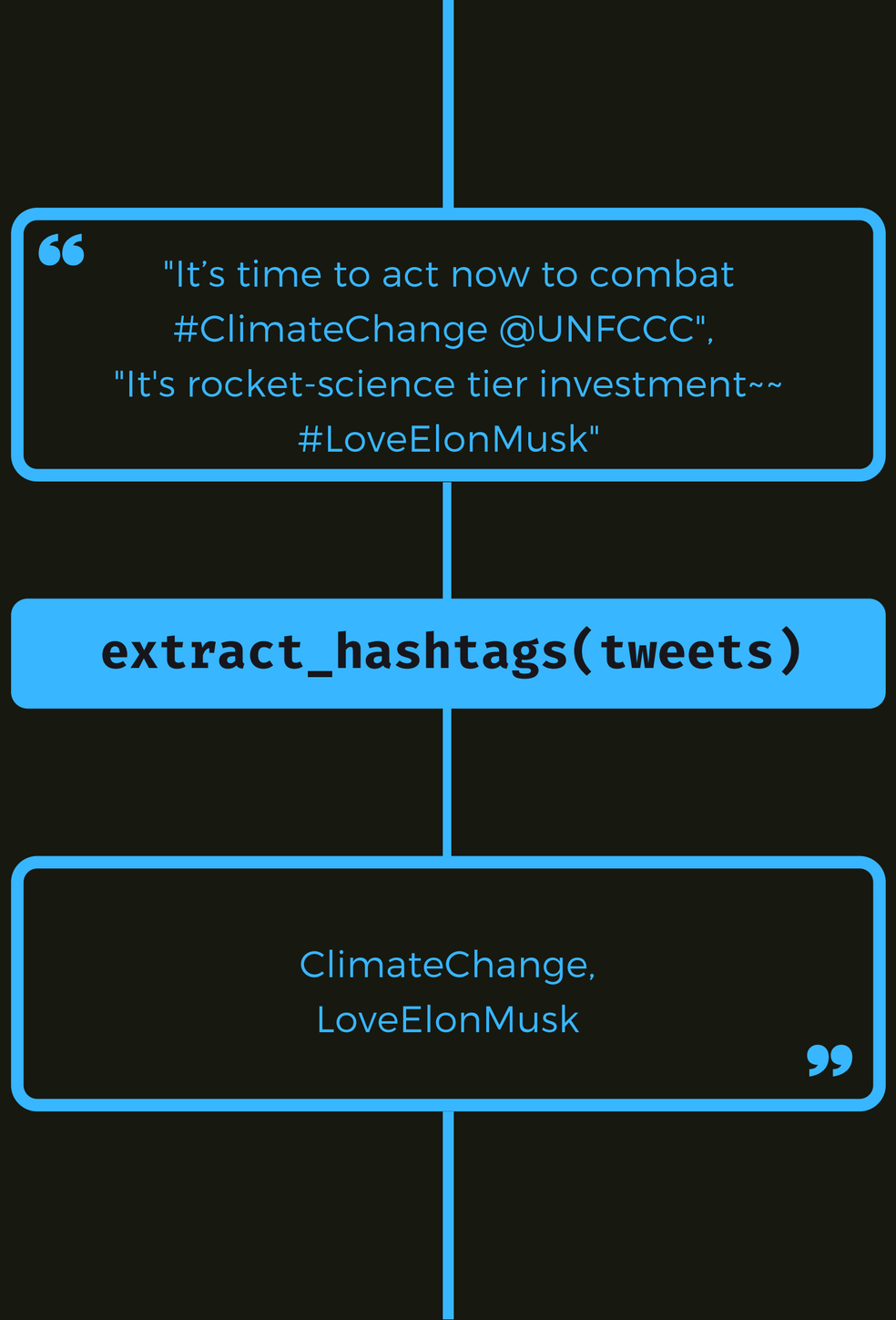
Specify the length of the n-grams to be created with the parameter n



Feature Engineering

`extract_hashtags()`

Extract all hashtags from a list of tweets



Data Visualization

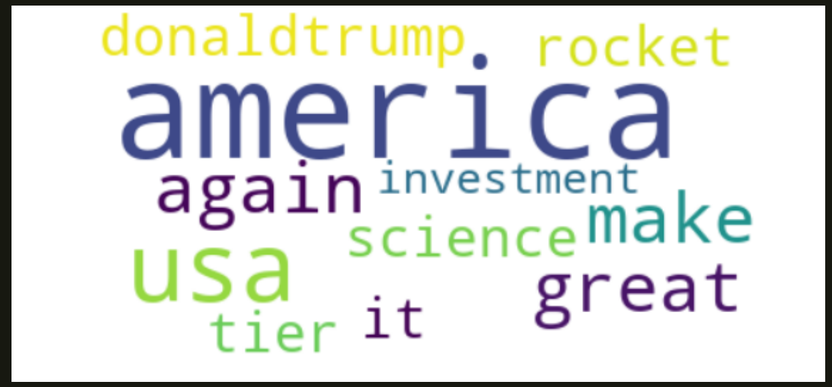
`generate_cloud()`

Generate word cloud

Supported types: words, hashtag, stopwords

"Make America Great Again!
@DonaldTrump #America",
"It's rocket-science tier investment~~",
"America America America #USA #USA"

```
generate_cloud(tweet_list,
type="words")
```



Installation

From GitHub

SOURCE CODE:

https://github.com/UBC-MDS/pytextprep

TO INSTALL:

pip install pytextprep

DOCUMENTATION:

https://pytextprep.readthedocs.io/en/latest/

Q&A