Open source libraries and tools data scientists in marketing should check out

Janet Wagner

Zylotech™

Customer Data & Analytics Blog

Janet Wagner, on October 18, 2018 | 3 minute read

We recently published a blog post that highlights how Zylotech can help data scientists in the marketing technology industry boost their productivity by automating many data preparation tasks. Our data scientists have

discovered many open source libraries and tools that they have over time extended and customized to support marketing analytics and marketing segmentation specifically. In fact, many marketing technology platforms available today are built not only with proprietary technologies but also open source software. So, today we thought we would highlight a few open source libraries and tools that data scientists in the marketing technology industry may find especially helpful.

Please note that almost all of these tools are written in Python.

Dask

Python is a single threaded language, and Dask is a library for parallel computing in Python. Dask has two main parts (collections and schedulers), and parallel computations are represented with task graphs. The task graphs allow users to create complex algorithms and handle scenarios that cannot be easily managed using the common paradigm of "map/filter/groupby." Dask includes a set of APIs so that the library can be easily integrated with a variety of data science tools such as Scikit-Learn, Pandas, and Numpy.

dateparser

dateparser provides modules for parsing localized dates in nearly any string format commonly found on web pages. Among the many capabilities of dateparser are the generic parsing of dates in more than 200 language locales, generic parsing of relative dates, search for and extract dates from long strings of text, and support for non-Gregorian calendar systems.

Gensim

Gensim is a library than can be used for analyzing plain-text documents for semantic structure, retrieving semantically similar documents, and document indexing. Gensim features include (but are not limited to) memory-independent algorithms (regarding corpus size), distributed computing, and
multicore implementations of popular algorithms. Among the algorithm implementations are Latent Semantic Analysis (LSA/LSI/SVD), Random Projections (RP), and Word2vec deep learning. The library also includes a trivial streaming API for plugging in your own input corpus/data stream and a trivial
transformation API for extending Gensim with other Vector Space algorithms.

TextBlob

TextBlob is a library that simplifies text processing. The library includes an API that allows common natural language processing (NLP) tasks to be enabled in applications and models. Among the many TextBlob features are noun phrase extraction, part-of-speech tagging, sentiment analysis, and n-grams. The documentation says that “TextBlob aims to provide access to common text-processing operations through a familiar interface. You can treat TextBlob objects as if they were Python strings that learned how to do Natural Language Processing.”

A few more libraries and tools to check out:

EmailHarvester
EmailHarvester is a tool that can be used to extract Domain email addresses from web search engines like Google, Bing, and Baidu.

Selenium Grid

Selenium Grid is a smart proxy server that aims to make it easy to run tests in parallel on multiple machines.

Sherlock

Sherlock is a tool that extracts interesting information about redditors based on their submissions and comments.

theHarvester

theHarvester is a tool that can be used to retrieve e-mails, subdomains, and names from many public sources e.g. search engines and pgp key servers.

Janet Wagner is a Zylotech contributing writer.

If you liked this post, check out our other blog post on the CDP is not just another customer database.

Subscribe to our blog newsletter for all things customer data and analytics, AI and marketing, sent monthly.