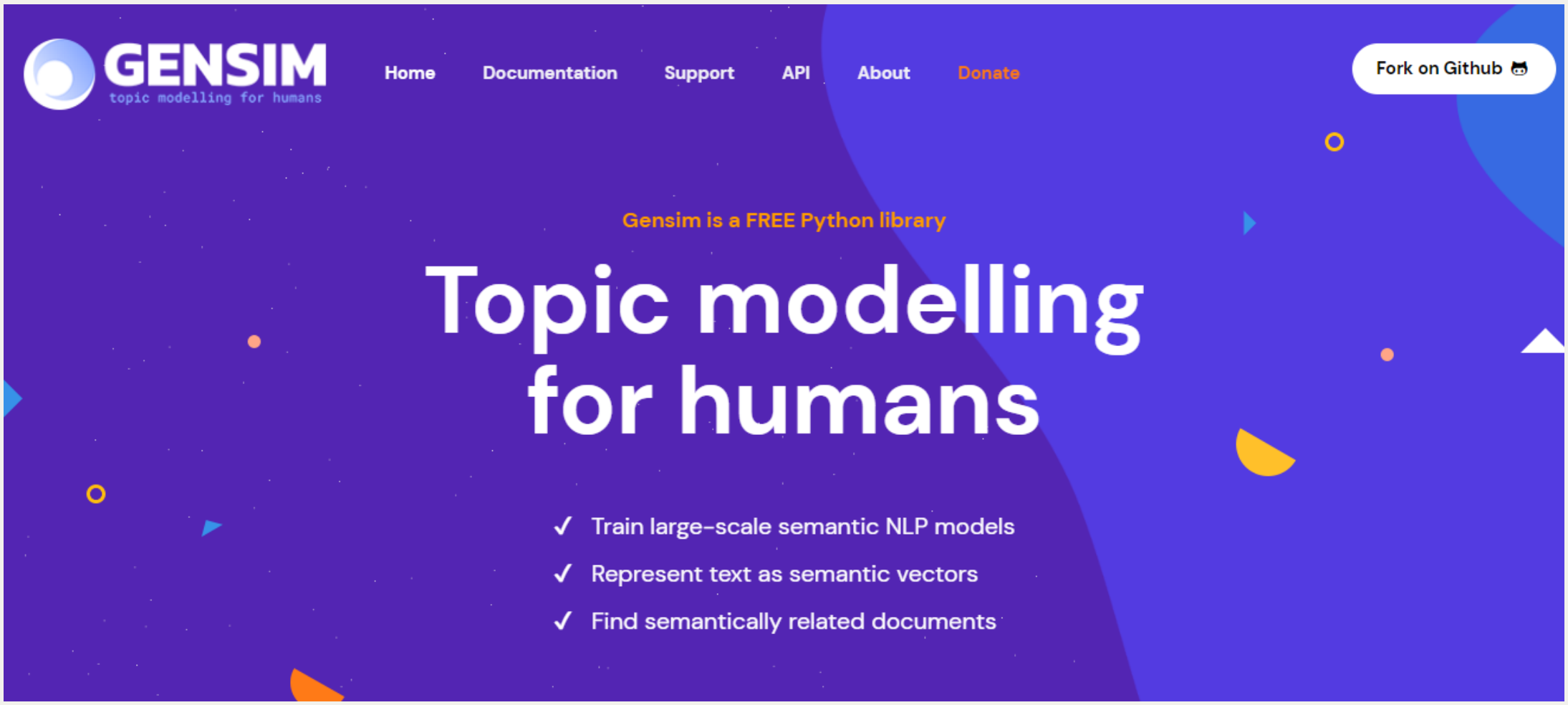
BEST TOPIC MODELING PYTHON LIBRARIES TO USE IN 2022

READ MORE

mdena

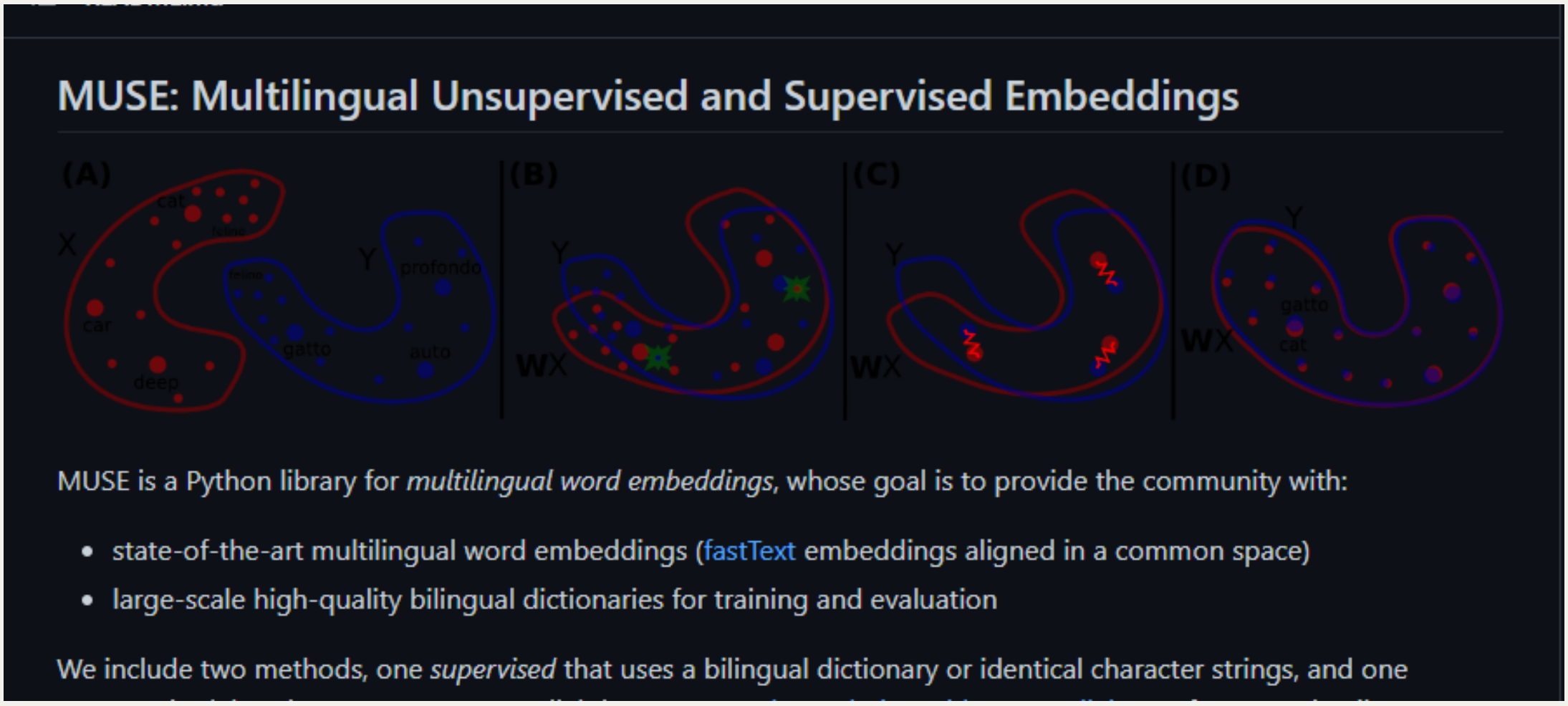
GENSIM



https://radimrehurek.com/gensim/

mdena

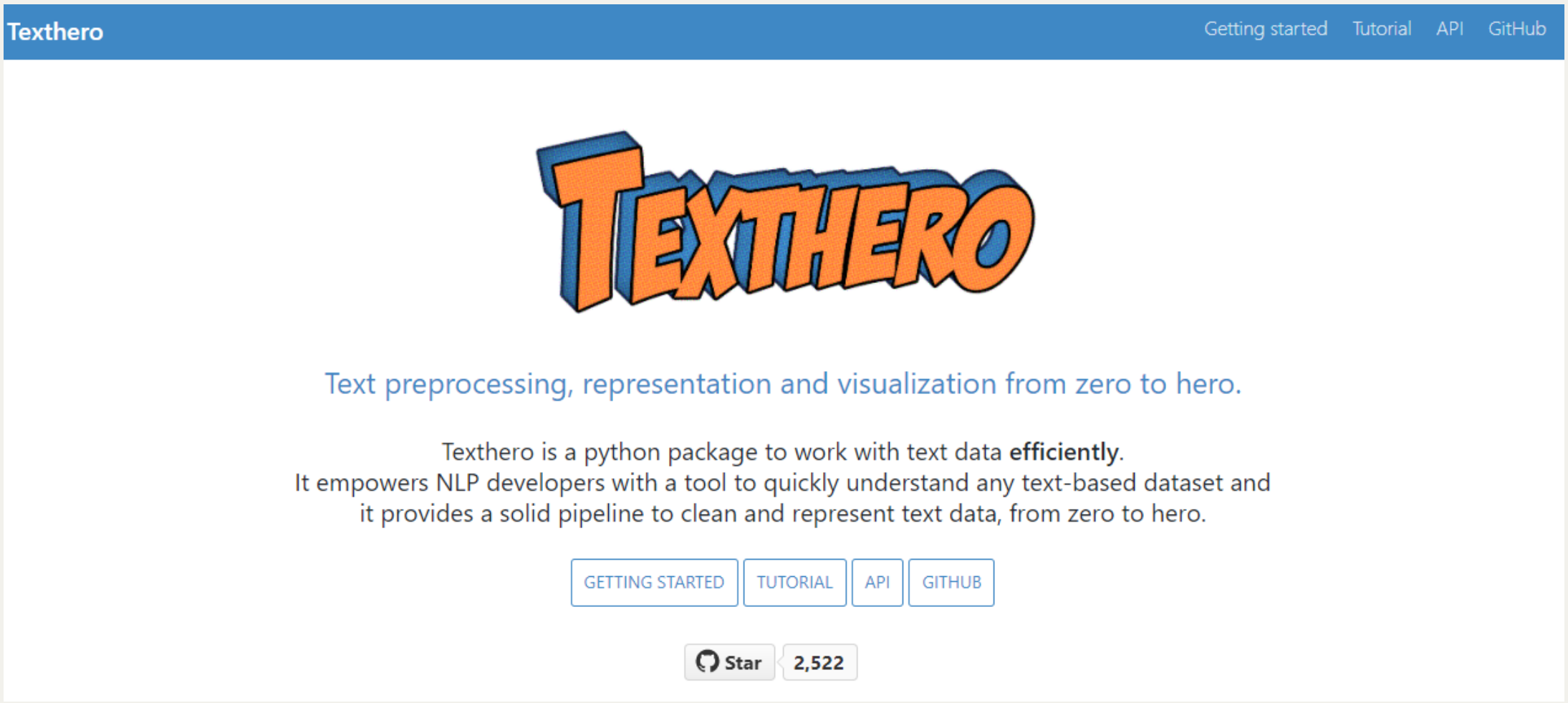
MUSE



https://github.com/facebookresearch/MUSE

mdena

TEXTHERO



https://texthero.org/

mdena

BERTOPIC

README.md

python v3.7+ build passing docs Passing pypi v0.10.0 license MIT arXiv 2203.05794

BERTopic

BERTopic is a topic modeling technique that leverages 🤗 transformers and c-TF-IDF to create dense clusters allowing for easily interpretable topics whilst keeping important words in the topic descriptions.

BERTopic supports **guided**, (semi-) **supervised**, and **dynamic** topic modeling. It even supports visualizations similar to LDAvis!

Corresponding medium posts can be found here and here. For a more detailed overview, you can read the paper.

https://github.com/MaartenGr/BERTopic

mdena

SCATTERTEXT

README.md

build passing pypi v0.1.6 GITTER join chat Follow 466

Scattertext 0.1.6

A tool for finding distinguishing terms in corpora and displaying them in an interactive HTML scatter plot. Points corresponding to terms are selectively labeled so that they don't overlap with other labels or points.

Below is an example of using Scattertext to create visualize terms used in 2012 American political conventions. The 2,000 most party-associated unigrams are displayed as points in the scatter plot. Their x- and y- axes are the dense ranks of their usage by Republican and Democratic speakers respectively.

https://github.com/JasonKessler/scattertext

mdena

LDA

lda: Topic modeling with latent Dirichlet Allocation

lda implements latent Dirichlet allocation (LDA) using collapsed Gibbs sampling. lda is fast and can be installed without a compiler on Linux and macOS.

The interface follows conventions found in scikit-learn. The following demonstrates how to inspect a model of a subset of the Reuters news dataset. (The input below, `X`, is a document-term matrix.)

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

ETM

README.rst

Contextualized Topic Models

pypi v2.3.0 | Python package passing | docs passing | contributors 9 | License MIT | downloads 153k | Open in Colab | Medium | youtube video

Contextualized Topic Models (CTM) are a family of topic models that use pre-trained representations of language (e.g., BERT) to support topic modeling. See the papers for details:

- Bianchi, F., Terragni, S., & Hovy, D. (2021). Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence. ACL. https://aclanthology.org/2021.acl-short.96/
- Bianchi, F., Terragni, S., Hovy, D., Nozza, D., & Fersini, E. (2021). Cross-lingual Contextualized Topic Models with Zero-shot Learning. EACL. https://www.aclweb.org/anthology/2021.eacl-main.143/

https://github.com/MilaNLProc/contextualized-topic-models

mdena

To learn more about NLP and NLP project examples, go to the link below:

omdena.com/blog

Omdena