TextDescriptives: A Python package for calculating a large variety of statistics from text

A PREPRINT

Lasse Hansen
Department of Affective Disorders - Psychiatry
Aarhus University Hospital
Aarhus

Kenneth Enevoldsen
Center for Humanities Computing
Aarhus University
Aarhus

January 4, 2023

ABSTRACT

TextDescriptives is a Python package for calculating a large variety of statistics from text. It is built on top of spaCy and can be easily integrated into existing workflows. The package has already been used for analysing the linguistic stability of clinical texts, creating features for predicting neuropsychiatric conditions, and analysing linguistic goals of primary school students. This paper describes the package and its features.

***K*eywords** Python • natural language processing • spacy • feature extraction

1 TextDescriptives: A Python package for calculating a large variety of statistics from text

2 Summary

Natural language processing (NLP) tasks often require a thorough understanding and description of the corpus. Document-level metrics can be used to identify low-quality data, assess outliers, or understand differences between groups. Further, text metrics have long been used in fields such as the digital humanities where e.g. metrics of text complexity are commonly used to analyse, understand and compare text corpora. However, extracting complex metrics can be an error-prone process and is rarely rigorously tested in research implementations. This can lead to subtle differences between implementations and reduces the reproducibility of scientific results.

`TextDescriptives` offers a simple and modular approach to extracting both simple and complex metrics from text. It achieves this by building on the `spaCy` framework (Honnibal et al. 2020). This means that `TextDescriptives` can easily be integrated into existing workflows while leveraging the efficiency and robustness of the `spaCy` library. The package has already been used for analysing the linguistic stability of clinical texts (Hansen et al. 2022), creating features for predicting neuropsychiatric conditions (**hansen_speaking_2022?**), and analysing linguistic goals of primary school students (Tannert 2023).

3 Statement of need

Computational text analysis is a broad term that refers to the process of analyzing and understanding text data. This often involves calculating a set of metrics that describe relevant properties of the data. Dependent on the task at hand, this can range from simple descriptive statistics related to e.g. word or sentence length to complex measures of text

complexity, coherence, or quality. This often requires drawing on multiple libraries and frameworks or writing custom code. This can be time-consuming and prone to bugs, especially with more complex metrics.

`TextDescriptives` seeks to unify the extraction of document-level metrics, in a modular fashion. The integration with spaCy allows the user to seamlessly integrate `TextDescriptives` in existing pipelines as well as giving the `TextDescriptives` package access to model-based metrics such as dependency graphs and part-of-speech tags. The ease of use and the variety of available metrics allows researchers and practitioners to extend the granularity of their analyses within a tested and validated framework.

Implementations of the majority of the metrics included in `TextDescriptives` exist, but none as feature complete. The `textstat` library (Ward 2022) implements the same readability metrics, however, each metric has to be extracted one at a time with no interface for multiple extractions. `spacy-readability` (Holtzscher 2019) adds readability metrics to `spaCy` pipelines, but does not work for new versions of `spaCy` (>=3.0.0). The `textacy` (DeWilde 2021) package has some overlap with `TextDescriptives`, but with a different focus. `TextDescriptives` focuses on document-level metrics, and includes a large number of metrics not included in `textacy` (dependency distance, coherence, (pseudo) perplexity, and quality), whereas `textacy` includes components for preprocessing, information extraction, and visualization that are outside the scope of `TextDescriptives`. What sets `TextDescriptives` apart is the easy access to document-level metrics through a simple user-facing API and exhaustive documentation.

4 Features & Functionality

`TextDescriptives` is a Python package and provides the following `spaCy` pipeline components: `textdescriptives.descriptive_stats`: Calculates the total number of tokens, number of unique tokens, number of characters, and the proportion of unique tokens, as well as the mean, median, and standard deviation of token length, sentence length, and the number of syllables per token. `textdescriptives.readability`: Calculates the Gunning-Fog index, the SMOG index, Flesch reading ease, Flesch-Kincaid grade, the Automated Readability Index, the Coleman-Liau index, the Lix score, and the Rix score. `textdescriptives.dependency_distance`: Calculates the mean and standard deviation of the dependency distance (the average distance between a word and its head word), and the mean and the standard deviation of the proportion adjacent dependency relations on the sentence level. `textdescriptives.pos_proportions`: Calculates the proportions of all part-of-speech tags in the documents. `textdescriptives.coherence`: Calculates the first- and second-order coherence of the document based on word embedding similarity between sentences. `textdescriptives.quality`: Calculates the text-quality metrics proposed in Rae et al. (2022) and Raffel et al. (2020). These measures can be used for filtering out low-quality text prior to model training or text analysis. These include heuristics such as the number of stop words, ratio of words containing alphabetic characters, proportion of lines ending with an ellipsis, proportion of lines starting with a bullet point, ratio of symbols to words, and whether the document contains a specified string (e.g. “lorem ipsum”), as well as repetitious text metrics such as the proportion of lines that are duplicates, the proportion of paragraphs in a document that are duplicates, the proportion of n-gram duplicates, and the proportion of characters in a document that are contained within the top n-grams.

All the components can be added to an existing `spaCy` pipeline with a single line of code, and jointly extracted to a dataframe or dictionary with a single call to `textdescriptives.extract_{df|dict}(doc)`.

5 Example Use Cases

Descriptive statistics can be used to summarize and understand data, such as by exploring patterns and relationships within the data, getting a better understanding of the data set, or identifying any changes in the distribution of the data. Readability metrics, which assess the clarity and ease of understanding of written text, have a variety of applications, including the design of educational materials and the improvement of legal or technical documents [DuBay (2004). Dependency distance can be used as a measure of language comprehension difficulty or of sentence complexity and has been used for analysing properties of natural language or for similar purposes as readability metrics (Gibson et al. 2019; Liu 2008). The proportions of different parts of speech in a document have been found to be predictive of certain mental disorders and can also be used to assess the quality and complexity of text (Tang et al. 2021). Semantic coherence, or the logical connection between sentences, has primarily been used in the field of computational psychiatry to predict the onset of psychosis or schizophrenia (Parola et al. 2022; Bedi et al. 2015), but it also has other applications in the digital humanities. Measures of text quality are useful cleaning and identifying low-quality data [Rae et al. (2022); Raffel et al. (2020).

6 Target Audience

The package is mainly targeted at NLP researchers and practitioners. In particular, researchers from fields new to NLP such as the digital humanities and social sciences as researchers might benefit from the readability metrics as well as the more complex, but highly useful, metrics such as coherence and dependency distance.

7 Acknowledgements

The authors thank the contributors of the package including Ludvig Olsen, for his work on the early versions of `TextDescriptives`, Martin Bernstorff for his work on the part-of-speech component, and Frida Hæstrup and Roberta Rocca for important fixes. The authors would also like to Dan Sattrup Nielsen for helpful reviews on early iterations of the text quality implementations.

References

Bedi, Gillinder, Facundo Carrillo, Guillermo A. Cecchi, Diego Fernández Slezak, Mariano Sigman, Natália B. Mota, Sidarta Ribeiro, Daniel C. Javitt, Mauro Copelli, and Cheryl M. Corcoran. 2015. "Automated Analysis of Free Speech Predicts Psychosis Onset in High-Risk Youths." *Npj Schizophrenia* 1 (1): 1–7. https://doi.org/10.1038/npjschz.2015.30.

DeWilde, Burton. 2021. *Textacy: NLP, Before and After spaCy* (version 0.12.0). https://github.com/chartbeat-labs/textacy.

DuBay, William H. 2004. "The Principles of Readability." *Online Submission*.

Gibson, Edward, Richard Futrell, Steven P. Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. 2019. "How Efficiency Shapes Human Language." *Trends in Cognitive Sciences* 23 (5): 389–407.

Hansen, Lasse, Kenneth Enevoldsen, Martin Bernstorff, Erik Perfalk, Andreas A. Danielsen, Kristoffer L. Nielbo, and Søren D. Østergaard. 2022. "Lexical Stability of Psychiatric Clinical Notes from Electronic Health Records over a Decade." *medRxiv*, January, 2022.09.05.22279610. https://doi.org/10.1101/2022.09.05.22279610.

Holtzscher, Michael. 2019. *Spacy-Readability: spaCy Pipeline Component for Adding Text Readability Meta Data to Doc Objects.* (version 1.4.1).

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. *spaCy: Industrial-Strength Natural Language Processing in Python*. https://doi.org/10.5281/zenodo.1212303.

Liu, Haitao. 2008. "Dependency Distance as a Metric of Language Comprehension Difficulty." *Journal of Cognitive Science* 9 (2): 159–91.

Parola, Alberto, Jessica Mary Lin, Arndis Simonsen, Vibeke Bliksted, Yuan Zhou, Huiling Wang, Lana Inoue, Katja Koelkebeck, and Riccardo Fusaroli. 2022. "Speech Disturbances in Schizophrenia: Assessing Cross-Linguistic Generalizability of NLP Automated Measures of Coherence." *Schizophrenia Research*, August. https://doi.org/10.1016/j.schres.2022.07.002.

Rae, Jack W., Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, et al. 2022. "Scaling Language Models: Methods, Analysis & Insights from Training Gopher." arXiv:2112.11446. arXiv. https://doi.org/10.48550/arXiv.2112.11446.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." *arXiv:1910.10683 [Cs, Stat]*, July. http://arxiv.org/abs/1910.10683.

Tang, Sunny X., Reno Kriz, Sunghye Cho, Suh Jung Park, Jenna Harowitz, Raquel E. Gur, Mahendra T. Bhati, Daniel H. Wolf, João Sedoc, and Mark Y. Liberman. 2021. "Natural Language Processing Methods Are Sensitive to Sub-Clinical Linguistic Differences in Schizophrenia Spectrum Disorders." *Npj Schizophrenia* 7 (1): 1–8. https://doi.org/10.1038/s41537-021-00154-3.

Tannert, Morten. 2023. "Skriftsproglig Udvikling i Grundskolens Danskfag." PhD thesis, Aarhus University.

Ward, Alex. 2022. *Textstat*. Textstat. https://github.com/textstat/textstat.