The Journal of Open Source Software

Tashaphyne: A Python package for Arabic Light Stemming

Taha Zerrouki [1¶]

1 Bouira University, Bouira, Algeria ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Samuel Forbes

Reviewers:

- @SamHames
- @kikarimullah

Submitted: 28 April 2023
Published: unpublished

License
Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC BY 4.0)

Summary

Stemming is an important task in natural language processing that involves reducing a word to its root form, or stem. In many cases, stemming can significantly improve the accuracy and efficiency of text analysis tasks such as information retrieval, text classification, and sentiment analysis. For the Arabic language, which has a rich morphology with a large number of prefixes and suffixes, stemming is particularly challenging. Tashaphyne provides an effective solution to this challenge, making it a valuable tool for researchers and practitioners working with Arabic text data.

Tashaphyne is a Python package that provides a comprehensive light stemmer and segmentor for the Arabic language. It stands out among other stemmers for its ability to perform stemming and root extraction simultaneously, unlike the Khoja stemmer, ISRI stemmer, Assem stemmer, and Farasa stemmer. Tashaphyne uses a modified finite state automaton that generates all possible segmentations, making it an extremely flexible tool for customizing stemmers without changing the code. Furthermore, Tashaphyne comes with default prefixes and suffixes, and allows for the use of customized lists to handle more complex aspects of stemming. Overall, Tashaphyne is an important contribution to the open-source community for Arabic language processing.

Statement of need

The Arabic language has a complex morphology with a rich system of prefixes, suffixes, and infixes. As a result, stemming Arabic text is a challenging task that requires specialized tools. While there are several Arabic stemmers available, they often have limitations in terms of accuracy and flexibility. Tashaphyne addresses these limitations by providing a comprehensive light stemmer and segmentor that performs stemming and root extraction simultaneously, generating all possible segmentations.

Tashaphyne is a light stemmer and segmentor in Arabic. It mostly supports light stemming (the removal of prefixes and suffixes) and provides all conceivable segmentations. Tahsphyne is a stem-based finite state automaton that extracts affixes (prefixes and suffixes) from a predefined list. It extracts and provides all possible affixations and configurations that result from a given word. Unlike the Khoja stemmer (Khoja & Garside, 1999) ISRI stemmer (Taghva et al., 2005), Assem stemmer (Chelli, 2019), and Farasa stemmer (Darwish & Mubarak, 2016), it can do both stemming and root extraction.

Tashaphyne also supports modifiable prefixes and suffixes, making it a highly adaptable tool for building customized stemmers without altering the code in any way.

Tashaphyne can be found at PyPi.org index [1]., it's available as demo on Mishkal, choose Tools/Analysis and as source code on Github.

[1] https://pypi.org/project/tashaphyne/

DRAFT

Tashaphyne contains two important submodules: stemming and normalizing. Normalizing text is an important preprocessing step in natural language processing that involves transforming text data into a standardized format. Normalization of Arabic text involves several sub-tasks, including removing diacritics (Zerrouki, 2023), normalizing characters, and removing ligatures. These sub-tasks are essential for improving the accuracy of downstream tasks such as text classification, named entity recognition, and sentiment analysis. Tashaphyne, with its ability to perform light stemming and segmenting, can also assist in normalizing Arabic text, further highlighting its importance in Arabic language processing

Tashaphyne has been developed within "Adawat", an open-source framework for processing Arabic texts developed as part of a PhD research project (Zerrouki, 2020). Adawat includes several tools, including Mishkal (Zerrouki, 2022a) for restoring Arabic text diacritics and Qalsadi (Zerrouki, 2022b) for Arabic morphology analysis, both of which rely on Tashaphyne's functionalities. In another project, we worked on applying the stemming algorithm to tackle the information retrieval problem in medical documents. (Al-Khatib et al., 2021).

Another framework that has incorporated Tashaphyne is the Classical Language Toolkit (CLTK [2] (Johnson, 2014)), which provides natural language processing support for ancient, classical, and medieval Eurasian languages. CLTK uses Tashaphyne for several tasks, including corpus importer, tokenization, text conversion, and transliteration for classical Arabic (Johnson, 2014) (like the orthography of the Quran).

The SAFAR framework, a comprehensive toolkit for Arabic natural language processing, has also incorporated Tashaphyne as part of its stemmers. However, as SAFAR (Y. Jaafar & Bouzoubaa, 2015) is written in Java, Tashaphyne was translated to the Java programming language to enable its integration into the framework.

Tashaphyne is a powerful Python package designed to facilitate natural language processing tasks, with a particular focus on Arabic text preprocessing. Its numerous features make it a valuable tool for researchers and developers alike. Tashaphyne provides support for light stemming of Arabic words, root extraction, and word segmentation. It also includes a default list of Arabic affixes and allows users to customize their own stemmer options and data. Furthermore, Tashaphyne supports data-independent stemming, making it highly versatile and adaptable to a wide range of use cases.

In terms of applications, Tashaphyne is ideal for stemming Arabic text, which is a crucial step in many natural language processing tasks. It is also useful for text classification and categorization, sentiment analysis, and named entity recognition. Tashaphyne has already been used in numerous scientific publications, demonstrating its reliability and effectiveness in a variety of real-world applications. With its comprehensive set of features and wide range of potential applications, Tashaphyne is an indispensable tool for anyone working with Arabic text data.

Mention

Tashaphyne has been widely used as a tool in various natural language processing tasks by researchers. Stemming development and evaluation have been explored by (Atoum & Nouman, 2019; Dahab et al., 2015; ElDefrawy et al., 2015b, 2016; Younes Jaafar et al., 2017; Y. Jaafar & Bouzoubaa, 2015). Root extraction and evaluation were studied by (ElDefrawy et al., 2015a, 2017).

Tashaphyne has been utilized for text categorization (Hussein et al., 2016; Sallam et al., 2016), classification (Y. A. Alhaj et al., 2019; El Mahdaouy et al., 2016; Gharbat et al., 2019; Hijazi et al., 2022; Muaad et al., 2022; Naji et al., 2017), topic segmentation (Alahmadi et al., 2022; Naili et al., 2018), and summarization (AlOudah et al., 2019; Etaiwi & Awajan, 2022; Tanfouri & Jarray, 2022).

[2] http://cltk.org

DRAFT

It has been applied to social media analysis (Almuqhim, 2016; Ameur et al., 2023; Bulbul et al., 2018; Kumar et al., 2013; Kumar, 2015), sentiment analysis (AlAyyoub et al., 2018; Saud Saleh Alotaibi, 2015; Saud S. Alotaibi & Anderson, 2016; Alqahtani et al., 2023; AlTwairesh et al., 2014; AlYasiri & Al-Azawei, 2019; Mouaad et al., 2023; Oraby et al., 2013; Oussous et al., 2019, 2020; A. M. Shoukry, 2013; A. Shoukry & Rafea, 2012), and tweet classification (E. Abozinadah, 2017; E. A. Abozinadah & Jones Jr, 2016; F. Alhaj et al., 2022; Brahimi et al., 2016; Mourad et al., 2017).

Tashaphyne has also been utilized for building resources such as corpora (Kuppevelt et al., 2018) and ontologies (Albukhitan et al., 2017), question answering (Abdul Salam, 2022; Ezzeldin, 2014; Ezzeldin et al., 2015), and information retrieval (Al-Khatib et al., 2021; Mortaja, 2017; S & R, 2022).

Acknowledgements

We gratefully acknowledge the contributions of Tashaphyne light stemmer, and Arabeyes.org during the project's inception.

References

Abdul Salam, M. A. A. H., Mustafa AND El-Fatah. (2022). Automatic grading for arabic short answer questions using optimized deep learning model. *PLOS ONE*, *17*(8), 1–41. https://doi.org/10.1371/journal.pone.0272269

Abozinadah, E. (2017). *Detecting abusive arabic language twitter accounts using a multidimensional analysis model* [PhD thesis]. George Mason University.

Abozinadah, E. A., & Jones Jr, J. H. (2016). Improved microblog classification for detecting abusive arabic twitter accounts. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, *6*(6), 17–28. https://doi.org/10.5121/ijdkp.2016.6602

Alahmadi, D., Wali, A., & Alzahrani, S. (2022). TAAM: Topic-aware abstractive arabic text summarisation using deep recurrent neural networks. *Journal of King Saud University - Computer and Information Sciences*, *34*(6, Part A), 2651–2665. https://doi.org/10.1016/j.jksuci.2022.03.026

AlAyyoub, M., Khamaiseh, A. A., Jararweh, Y., & Al-Kabi, M. N. (2018). A comprehensive survey of arabic sentiment analysis. *Information Processing & Management*. https://doi.org/10.1016/j.ipm.2018.07.006

Albukhitan, S., Helmy, T., & Alnazer, A. (2017). Arabic ontology learning using deep learning. *Proceedings of the International Conference on Web Intelligence*, 1138–1142. https://doi.org/10.1145/3106426.3109052

Alhaj, F., Al-Haj, A., Sharieh, A., & Jabri, R. (2022). Improving arabic cognitive distortion classification in twitter using BERTopic. *International Journal of Advanced Computer Science and Applications*, *13*(1), 854–860. https://doi.org/10.14569/IJACSA.2022.0130199

Alhaj, Y. A., Xiang, J., Zhao, D., Al-Qaness, M. A., Elaziz, M. A., & Dahou, A. (2019). A study of the effects of stemming strategies on arabic document classification. *IEEE Access*, *7*, 32664–32671. https://doi.org/10.1109/access.2019.2903331

Al-Khatib, R. M., Zerrouki, T., Abu Shquier, M. M., Balla, A., & Al-Khateeb, A. (2021). A new enhanced arabic light stemmer for IR in medical documents. *CMC-COMPUTERS MATERIALS & CONTINUA*, *68*(1), 1255–1269. https://doi.org/10.32604/cmc.2021.016155

DRAFT

Almuqhim, F. (2016). *Strategies for sentiment analysis and classification of non english tweets* [PhD thesis]. Rochester Institute of Technology.

Alotaibi, Saud Saleh. (2015). *Sentiment analysis in the arabic language using machine learning* [PhD thesis]. Colorado State University. Libraries.

Alotaibi, Saud S., & Anderson, C. W. (2016). Extending the knowledge of the arabic sentiment classification using a foreign external lexical source. *International Journal on Natural Language Computing*, *5*(3), 1–11. https://doi.org/10.5121/ijnlc.2016.5301

AlOudah, A., Al Bassam, K., Kurdi, H., & Al-Megren, S. (2019). Wajeez: An extractive automatic arabic text summarisation system. *International Conference on Human-Computer Interaction*, 3–14. https://doi.org/10.1007/978-3-030-21902-4_1

Alqahtani, Y., Al-Twairesh, N., & Alsanad, A. (2023). A comparative study of effective domain adaptation approaches for arabic sentiment classification. *Applied Sciences*, *13*(3), 1387. https://doi.org/10.3390/app13031387

AlTwairesh, N., Al-Khalifa, H., & Al-Salman, A. (2014). Subjectivity and sentiment analysis of arabic: Trends and challenges. *Computer Systems and Applications (AICCSA), 2014 IEEE/ACS 11th International Conference on*, 148–155. https://doi.org/10.1109/aiccsa.2014.7073192

AlYasiri, E. K., & Al-Azawei, A. (2019). Improving arabic sentiment analysis on social media: A comparative study on applying different pre-processing techniques. *COMPUSOFT, An International Journal of Advanced Computer Technology*, *8*(6).

Ameur, H., Rekik, A., Jamoussi, S., & Hamadou, A. B. (2023). ChildProtect: A parental control application for tracking hostile surfing content. *Entertainment Computing*, *44*, 100517. https://doi.org/10.1016/j.entcom.2022.100517

Atoum, J. O., & Nouman, M. (2019). Sentiment analysis of arabic jordanian dialect tweets. *International Journal of Advanced Computer Science and Applications*, *10*(2), 256–262. https://doi.org/10.14569/ijacsa.2019.0100234

Brahimi, B., Touahria, M., & Tari, A. (2016). Data and text mining techniques for classifying arabic tweet polarity. *Journal of Digital Information Management*, *14*(1).

Bulbul, A., Kaplan, C., & Ismail, S. H. (2018). Social media based analysis of refugees in turkey. *Proceedings of the First International Workshop on Analysis of Broad Dynamic Topics over Social Media: BroDyn*, *18*.

Chelli, A. (2019). *Assem's arabic stemmers based on snowball framework*. https://arabicstemmer.com

Dahab, M. Y., Ibrahim, A., & Al-Mutawa, R. (2015). A comparative study on arabic stemmers. *International Journal of Computer Applications*, *125*(8). https://doi.org/10.5120/ijca2015906129

Darwish, K., & Mubarak, H. (2016). Farasa: A new fast and accurate arabic word segmenter. *The International Conference on Language Resources and Evaluation LREC'10*.

El Mahdaouy, A., Gaussier, E., & El Alaoui, S. O. (2016). Arabic text classification based on word and document embeddings. *International Conference on Advanced Intelligent Systems and Informatics*, 32–41. https://doi.org/10.1007/978-3-319-48308-5_4

ElDefrawy, M., Belal, N. A., & El-Sonbaty, Y. (2017). An efficient rank based arabic root extractor. *Intelligent Systems Conference (IntelliSys), 2017*, 870–878. https://doi.org/10.1109/intellisys.2017.8324232

ElDefrawy, M., El-Sonbaty, Y., & Belal, N. (2015a). Enhancing root extractors using light stemmers. *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation: Posters*, 157–166.

DRAFT

ElDefrawy, M., El-Sonbaty, Y., & Belal, N. A. (2015b). Cbas: Context based arabic stemmer. *arXiv Preprint arXiv:1611.00027*. https://doi.org/10.5121/ijnlc.2015.4301

ElDefrawy, M., El-Sonbaty, Y., & Belal, N. A. (2016). A rule-based subject-correlated arabic stemmer. *Arabian Journal for Science and Engineering*, *41*(8), 2883–2891. https://doi.org/10.1007/s13369-016-2029-2

Etaiwi, W., & Awajan, A. (2022). SemG-TS: Abstractive arabic text summarization using semantic graph embedding. *Mathematics*, *10*(18), 3225. https://doi.org/10.3390/math10183225

Ezzeldin, A. M. (2014). *Answer selection and validation for arabic questions* [PhD thesis]. Arab Academy for Science.

Ezzeldin, A. M., El-Sonbaty, Y., & Kholief, M. H. (2015). Exploring the effects of root expansion, sentence splitting and ontology on arabic answer selection. *Natural Language Processing and Cognitive Science: Proceedings*, *2014*, 273. https://doi.org/10.1515/9781501501289.273

Gharbat, M., Saadeh, H., & Al Fayez, R. Q. (2019). Discovering the applicability of classification algorithms with arabic poetry. *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, 453–458. https://doi.org/10.1109/jeeit.2019.8717387

Hijazi, M. M., Zeki, A., & Ismail, A. (2022). A review study on arabic text classification. *2022 International Arab Conference on Information Technology (ACIT)*, 1–13. https://doi.org/10.1109/ACIT57182.2022.9994124

Hussein, M., Mousa, H. M., & Sallam, R. M. (2016). Arabic text categorization using mixed words. *I.J. Information Technology and Computer Science*, *11*, 74–81. https://doi.org/10.5815/ijitcs.2016.11.09

Jaafar, Y., & Bouzoubaa, K. (2015). Arabic Natural Language Processing from Software Engineering to Complex Pipeline. *2015 First International Conference on Arabic Computational Linguistics (ACLing)*, 29–36. https://doi.org/10.1109/ACLing.2015.11

Jaafar, Younes, Namly, D., Bouzoubaa, K., & Yousfi, A. (2017). Enhancing arabic stemming process using resources and benchmarking tools. *Journal of King Saud University-Computer and Information Sciences*, *29*(2), 164–170. https://doi.org/10.1016/j.jksuci.2016.11.010

Johnson, K. (2014). *CLTK: The classical language toolkit*. https://github.com/cltk/cltk.

Khoja, S., & Garside, R. (1999). Stemming arabic text. *Lancaster, UK, Computing Department, Lancaster University*.

Kumar, S. (2015). *Social media analytics for crisis response*. Arizona State University.

Kumar, S., Morstatter, F., Zafarani, R., & Liu, H. (2013). Whom should i follow?: Identifying relevant users during crises. *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, 139–147. https://doi.org/10.1145/2481492.2481507

Kuppevelt, D. van, Bos, E. P., Lyklema, A. M., Ryad, U., Lange, C. R., & Zwaan, J. M. van der. (2018). Bridging the gap: Digital humanities and the arabic-islamic corpus. *DH*, 682.

Mortaja, M. M. (2017). *Developing interactive cross lingual information retrieval tool* [PhD thesis]. The Islamic University–Gaza.

Mouaad, E., Ouassil, M. A., Rachidi, R., Cherradi, B., Hamida, S., & Raihani, A. (2023). Sentiment analysis on moroccan dialect based on ML and social media content detection. *International Journal of Advanced Computer Science and Applications*, *14*, 315–325. https://doi.org/10.14569/IJACSA.2023.0140347

Mourad, A., Scholer, F., & Sanderson, M. (2017). Language influences on tweeter geolocation. *European Conference on Information Retrieval*, 331–342. https://doi.org/10.1007/

DRAFT

978-3-319-56608-5_26

Muaad, A. Y., Davanagere, H. J., Guru, D., Benifa, J. B., Chola, C., AlSalman, H., Gumaei, A. H., & Al-antari, M. A. (2022). Arabic document classification: Performance investigation of preprocessing and representation techniques. *Mathematical Problems in Engineering*, *2022*, 1–16. https://doi.org/10.1155/2022/3720358

Naili, M., Chaibi, A. H., & Ghezala, H. H. B. (2018). The contribution of stemming and semantics in arabic topic segmentation. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, *17*(2), 12. https://doi.org/10.1145/3152464

Naji, H. A., Ashour, W. M., & Alhanjouri, M. A. (2017). A new model in arabic text classification using BPSO/REP-tree. *Journal of Engineering Research and Technology*, *4*(1).

Oraby, S., El-Sonbaty, Y., & El-Nasr, M. A. (2013). Exploring the effects of word roots for arabic sentiment analysis. *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 471–479.

Oussous, A., Benjelloun, F.-Z., Lahcen, A. A., & Belfkih, S. (2020). ASA: A framework for arabic sentiment analysis. *Journal of Information Science*, *46*(4), 544–559. https://doi.org/10.1177/0165551519849516

Oussous, A., Lahcen, A. A., & Belfkih, S. (2019). Impact of text pre-processing and ensemble learning on arabic sentiment analysis. *Proceedings of the 2nd International Conference on Networking, Information Systems & Security*, 65. https://doi.org/10.1145/3320326.3320399

S, S. V., & R, P. (2022). Text pre-processing methods on cross language information retrieval. *2022 International Conference on Connected Systems & Intelligence (CSI)*, 1–5. https://doi.org/10.1109/CSI54720.2022.9923952

Sallam, R. M., Mousa, H. M., & Hussein, M. (2016). Improving arabic text categorization using normalization and stemming techniques. *International Journal of Computer Applications*, *135*(2), 38–43. https://doi.org/10.5120/ijca2016908328

Shoukry, A. M. (2013). *ARABIC Sentence Level Sentiment Analysis* [PhD thesis]. The American University in Cairo.

Shoukry, A., & Rafea, A. (2012). Preprocessing egyptian dialect tweets for sentiment mining. *The Fourth Workshop on Computational Approaches to Arabic Script-Based Languages*, 47.

Taghva, K., Elkhoury, R., & Coombs, J. (2005). Arabic stemming without a root dictionary. *Information Technology: Coding and Computing, 2005. ITCC 2005. International Conference on*, *1*, 152–157. https://doi.org/10.1109/itcc.2005.90

Tanfouri, I., & Jarray, F. (2022). *Genetic algorithm and latent semantic analysis based documents summarization technique*. 223–227. https://doi.org/10.5220/0011585700003335

Zerrouki, T. (2020). *Towards an open platform for arabic language processing* (p. 39) [PhD]. Ecole Nationale Supérieure d'Informatique ESI, Algiers, Algeria.

Zerrouki, T. (2022a). Mishkal arabic text vocalization software. In *GitHub repository*. GitHub. https://github.com/linuxscout/mishkal

Zerrouki, T. (2022b). Qalsadi arabic morphological analyzer and lemmatizer for python. In *GitHub repository*. GitHub. https://github.com/linuxscout/qalsadi

Zerrouki, T. (2023). PyArabic: A python package for arabic text. *Journal of Open Source Software*, *8*(84), 4886. https://doi.org/10.21105/joss.04886

DRAFT