Tashaphyne: A Python package for Arabic Light Stemming

20 September 2022

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 and Garside 1999) ISRI stemmer (Taghva, Elkhoury, and Coombs 2005), Assem stemmer (Chelli 2019), and Farasa stemmer (Darwish and 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.

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 and Bouzoubaa 2015) is written in Java, Tashaphyne was translated to the Java programming language to enable its integration into the framework.

[1] https://pypi.org/project/tashaphyne/
[2] http://cltk.org

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 and Nouman 2019; Younes Jaafar et al. 2017; Y. Jaafar and Bouzoubaa 2015; ElDefrawy, El-Sonbaty, and Belal 2016, 2015b; Dahab, Ibrahim, and Al-Mutawa 2015). Root extraction and evaluation were studied by (ElDefrawy, El-Sonbaty, and Belal 2015a; ElDefrawy, Belal, and El-Sonbaty 2017).

Tashaphyne has been utilized for text categorization (Sallam, Mousa, and Hussein 2016; Hussein, Mousa, and Sallam 2016), classification (Muaad et al. 2022; Hijazi, Zeki, and Ismail 2022; Gharbat, Saadeh, and Al Fayez 2019; Naji, Ashour, and Alhanjouri 2017; El Mahdaouy, Gaussier, and El Alaoui 2016; Y. A. Alhaj et al. 2019), topic segmentation (Alahmadi, Wali, and Alzahrani 2022; Naili, Chaibi, and Ghezala 2018), and summarization (Etaiwi and Awajan 2022; Tanfouri and Jarray 2022; AlOudah et al. 2019).

It has been applied to social media analysis (Ameur et al. 2023; Almuqhim 2016; Bulbul, Kaplan, and Ismail 2018; Kumar et al. 2013; Kumar 2015), sentiment analysis (Alqahtani, Al-Twairesh, and Alsanad 2023; Mouaad et al. 2023; Oussous, Lahcen, and Belfkih 2019; Oussous et al. 2020; AlYasiri and Al-Azawei 2019; Saud S. Alotaibi and Anderson 2016; Saud Saleh Alotaibi 2015; AlTwairesh, Al-Khalifa, and Al-Salman 2014; Oraby, El-Sonbaty, and El-Nasr 2013; A. Shoukry and Rafea 2012; A. M. Shoukry 2013; AlAyyoub et al. 2018), and tweet classification (F. Alhaj et al. 2022; E. A. Abozinadah and Jones Jr 2016; E. Abozinadah 2017; Brahimi, Touahria, and Tari 2016; Mourad, Scholer, and Sanderson 2017).

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

Acknowledgements

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

References

Abdul Salam, Mohamed Abd AND Hassan, 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, Ehab. 2017. "Detecting Abusive Arabic Language Twitter Accounts Using a Multidimensional Analysis Model." PhD thesis, George Mason University.

Abozinadah, Ehab A, and James H Jones Jr. 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, Dimah, Arwa Wali, and Sarah Alzahrani. 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): 2651–65. https://doi.org/https://doi.org/10.1016/j.jksuci.2022.03.026.

AlAyyoub, Mahmoud, Abed Allah Khamaiseh, Yaser Jararweh, and Mohammed N Al-Kabi. 2018. "A Comprehensive Survey of Arabic Sentiment Analysis." *Information Processing & Management*. https://doi.org/10.1016/j.ipm.2018.07.006.

Albukhitan, Saeed, Tarek Helmy, and Ahmed Alnazer. 2017. "Arabic Ontology Learning Using Deep Learning." In *Proceedings of the International Conference on Web Intelligence*, 1138–42. ACM. https://doi.org/10.1145/3106426.3109052.

Alhaj, Fatima, Ali Al-Haj, Ahmad Sharieh, and Riad Jabri. 2022. "Improving Arabic Cognitive Distortion Classification in Twitter Using BERTopic." *International Journal of Advanced Computer Science and Applications* 13 (1): 854–60. https://doi.org/10.14569/IJACSA.2022.0130199.

Alhaj, Yousif A, Jianwen Xiang, Dongdong Zhao, Mohammed AA Al-Qaness, Mohamed Abd Elaziz, and Abdelghani Dahou. 2019. "A Study of the Effects of Stemming Strategies on Arabic Document Classification." *IEEE Access* 7: 32664–71. https://doi.org/10.1109/access.2019.2903331.

Al-Khatib, Ra'ed M, Taha Zerrouki, Mohammed M Abu Shquier, Amar Balla, and Asef Al-Khateeb. 2021. "A New Enhanced Arabic Light Stemmer for IR in Medical Documents." *CMC-COMPUTERS MATERIALS & CONTINUA* 68 (1): 1255–69. https://doi.org/10.32604/cmc.2021.016155.

Almuqhim, Fahad. 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, and Charles W Anderson. 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, Abrar, Kholoud Al Bassam, Heba Kurdi, and Shiroq Al-Megren. 2019. "Wajeez: An Extractive Automatic Arabic Text Summarisation System." In *International Conference on Human-Computer Interaction*, 3–14. Springer. https://doi.org/10.1007/978-3-030-21902-4_1.

Alqahtani, Yathrib, Nora Al-Twairesh, and Ahmed Alsanad. 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, Nora, Hend Al-Khalifa, and AbdulMalik Al-Salman. 2014. "Subjectivity and Sentiment Analysis of Arabic: Trends and Challenges." In *Computer Systems and Applications (AICCSA), 2014 IEEE/ACS 11th International Conference on*, 148–55. IEEE. https://doi.org/10.1109/aiccsa.2014.7073192.

AlYasiri, Essam Kazem, and Ahmed Al-Azawei. 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, Hanen, Amal Rekik, Salma Jamoussi, and Abdelmajid Ben Hamadou. 2023. "ChildProtect: A Parental Control Application for Tracking Hostile Surfing Content." *Entertainment Computing* 44: 100517. https://doi.org/https://doi.org/10.1016/j.entcom.2022.100517.

Atoum, Jalal Omer, and Mais Nouman. 2019. "Sentiment Analysis of Arabic Jordanian Dialect Tweets." *International Journal of Advanced Computer Science and Applications* 10 (2): 256–62. https://doi.org/10.14569/ijacsa.2019.0100234.

Brahimi, Belgacem, Mohamed Touahria, and Abdelkamel Tari. 2016. "Data and Text Mining Techniques for Classifying Arabic Tweet Polarity." *Journal of Digital Information Management* 14 (1).

Bulbul, Abdullah, Cagri Kaplan, and Salah Haj Ismail. 2018. "Social Media Based Analysis of Refugees in Turkey." In *Proceedings of the First International Workshop on Analysis of Broad Dynamic Topics over Social Media: BroDyn*. Vol. 18.

Chelli, Assem. 2019. "Assem's Arabic Stemmers Based on Snowball Framework." https://arabicstemmer.com.

Dahab, Mohamed Y, Al Ibrahim, and Rihab Al-Mutawa. 2015. "A Comparative

Study on Arabic Stemmers." *International Journal of Computer Applications* 125 (8). https://doi.org/10.5120/ijca2015906129.

Darwish, Kareem, and Hamdy Mubarak. 2016. "Farasa: A New Fast and Accurate Arabic Word Segmenter." In *The International Conference on Language Resources and Evaluation LREC'10*.

El Mahdaouy, Abdelkader, Eric Gaussier, and Said Ouatik El Alaoui. 2016. "Arabic Text Classification Based on Word and Document Embeddings." In *International Conference on Advanced Intelligent Systems and Informatics*, 32–41. Springer. https://doi.org/10.1007/978-3-319-48308-5_4.

ElDefrawy, Mahmoud, Nahla A Belal, and Yasser El-Sonbaty. 2017. "An Efficient Rank Based Arabic Root Extractor." In *Intelligent Systems Conference (IntelliSys), 2017*, 870–78. IEEE. https://doi.org/10.1109/intellisys.2017.8324232.

ElDefrawy, Mahmoud, Yasser El-Sonbaty, and Nahla Belal. 2015a. "Enhancing Root Extractors Using Light Stemmers." In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation: Posters*, 157–66.

ElDefrawy, Mahmoud, Yasser El-Sonbaty, and Nahla A Belal. 2015b. "Cbas: Context Based Arabic Stemmer." *arXiv Preprint arXiv:1611.00027*. https://doi.org/10.5121/ijnlc.2015.4301.

———. 2016. "A Rule-Based Subject-Correlated Arabic Stemmer." *Arabian Journal for Science and Engineering* 41 (8): 2883–91. https://doi.org/10.1007/s13369-016-2029-2.

Etaiwi, Wael, and Arafat Awajan. 2022. "SemG-TS: Abstractive Arabic Text Summarization Using Semantic Graph Embedding." *Mathematics* 10 (18): 3225. https://doi.org/10.3390/math10183225.

Ezzeldin, Ahmed Magdy. 2014. "Answer Selection and Validation for Arabic Questions." PhD thesis, Arab Academy for Science.

Ezzeldin, Ahmed Magdy, Yasser El-Sonbaty, and Mohamed Hamed Kholief. 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, Mohammad, Heba Saadeh, and Reem Q Al Fayez. 2019. "Discovering the Applicability of Classification Algorithms with Arabic Poetry." In *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, 453–58. IEEE. https://doi.org/10.1109/jeeit.2019.8717387.

Hijazi, Musab Mustafa, Akram Zeki, and Amelia Ismail. 2022. "A Review Study on Arabic Text Classification." In *2022 International Arab Conference on Information Technology (ACIT)*, 1–13. https://doi.org/10.1109/ACIT57182.2022.9994124.

Hussein, Mahmoud, Hamdy M Mousa, and Rouhia M Sallam. 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, and K Bouzoubaa. 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, Driss Namly, Karim Bouzoubaa, and Abdellah Yousfi. 2017. "Enhancing Arabic Stemming Process Using Resources and Benchmarking Tools." *Journal of King Saud University-Computer and Information Sciences* 29 (2): 164–70. https://doi.org/10.1016/j.jksuci.2016.11.010.

Johnson, Kyle. 2014. "CLTK: The Classical Language Toolkit." https://github.com/cltk/cltk.

Khoja, Shereen, and Roger Garside. 1999. "Stemming Arabic Text." *Lancaster, UK, Computing Department, Lancaster University.*

Kumar, Shamanth. 2015. *Social Media Analytics for Crisis Response.* Arizona State University.

Kumar, Shamanth, Fred Morstatter, Reza Zafarani, and Huan Liu. 2013. "Whom Should i Follow?: Identifying Relevant Users During Crises." *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, 139–47. https://doi.org/10.1145/2481492.2481507.

Kuppevelt, Dafne van, EG Patrick Bos, A Melle Lyklema, Umar Ryad, Christian R Lange, and Janneke M van der Zwaan. 2018. "Bridging the Gap: Digital Humanities and the Arabic-Islamic Corpus." In *DH*, 682.

Mortaja, Mohammed MS. 2017. "Developing Interactive Cross Lingual Information Retrieval Tool." PhD thesis, The Islamic University–Gaza.

Mouaad, Errami, Mohamed Amine Ouassil, Rabia Rachidi, Bouchaib Cherradi, Soufiane Hamida, and Abdelhadi Raihani. 2023. "Sentiment Analysis on Moroccan Dialect Based on ML and Social Media Content Detection." *International Journal of Advanced Computer Science and Applications* 14 (April): 315–25. https://doi.org/10.14569/IJACSA.2023.0140347.

Mourad, Ahmed, Falk Scholer, and Mark Sanderson. 2017. "Language Influences on Tweeter Geolocation." In *European Conference on Information Retrieval*, 331–42. Springer. https://doi.org/10.1007/978-3-319-56608-5_26.

Muaad, Abdullah Y, Hanumanthappa Jayappa Davanagere, DS Guru, JV Bibal Benifa, Channabasava Chola, Hussain AlSalman, Abdu H Gumaei, and Mugahed A Al-antari. 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, Marwa, Anja Habacha Chaibi, and Henda Hajjami Ben Ghezala. 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, Hamza A, Wesam M Ashour, and Mohammed A Alhanjouri. 2017. "A New Model in Arabic Text Classification Using BPSO/REP-Tree." *Journal of Engineering Research and Technology* 4 (1).

Oraby, Shereen, Yasser El-Sonbaty, and Mohamad Abou El-Nasr. 2013. "Exploring the Effects of Word Roots for Arabic Sentiment Analysis." In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 471–79.

Oussous, Ahmed, Fatima-Zahra Benjelloun, Ayoub Ait Lahcen, and Samir

Belfkih. 2020. “ASA: A Framework for Arabic Sentiment Analysis.” *Journal of Information Science* 46 (4): 544–59. https://doi.org/10.1177/0165551519849516.

Oussous, Ahmed, Ayoub Ait Lahcen, and Samir Belfkih. 2019. “Impact of Text Pre-Processing and Ensemble Learning on Arabic Sentiment Analysis.” In *Proceedings of the 2nd International Conference on Networking, Information Systems & Security*, 65. ACM. https://doi.org/10.1145/3320326.3320399.

S, Sakthi Vel, and Priya R. 2022. “Text Pre-Processing Methods on Cross Language Information Retrieval.” In *2022 International Conference on Connected Systems & Intelligence (CSI)*, 1–5. https://doi.org/10.1109/CSI54720.2022.9923952.

Sallam, Rouhia M, Hamdy M Mousa, and Mahmoud Hussein. 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, Amira Magdy. 2013. “ARABIC Sentence Level Sentiment Analysis.” PhD thesis, The American University in Cairo.

Shoukry, Amira, and Ahmed Rafea. 2012. “Preprocessing Egyptian Dialect Tweets for Sentiment Mining.” In *The Fourth Workshop on Computational Approaches to Arabic Script-Based Languages*, 47.

Taghva, Kazem, Rania Elkhoury, and Jeffrey Coombs. 2005. “Arabic Stemming Without a Root Dictionary.” In *Information Technology: Coding and Computing, 2005. ITCC 2005. International Conference on*, 1:152–57. IEEE. https://doi.org/10.1109/itcc.2005.90.

Tanfouri, Imen, and Fethi Jarray. 2022. “Genetic Algorithm and Latent Semantic Analysis Based Documents Summarization Technique.” In, 223–27. https://doi.org/10.5220/0011585700003335.

Zerrouki, Taha. 2020. “Towards an Open Platform for Arabic Language Processing.” PhD, Ecole Nationale Supérieure d’Informatique ESI, Algiers, Algeria.

———. 2022a. “Mishkal Arabic Text Vocalization Software.” *GitHub Repository*. GitHub. https://github.com/linuxscout/mishkal.

———. 2022b. “Qalsadi Arabic Morphological Analyzer and Lemmatizer for Python.” *GitHub Repository*. GitHub. https://github.com/linuxscout/qalsadi.

———. 2023. “PyArabic: A Python Package for Arabic Text.” *Journal of Open Source Software* 8 (84): 4886. https://doi.org/10.21105/joss.04886.