نور-استم نسخهٔ ۱. یک مجموعه‌داده معیار برای ارزیابی ریشه‌یاب‌های عربی

ازل العصواد[۱]، بهروز مینایی بیدگلی*[۱]، محمدابراهیم شناسا[۲]، حبیب سریانی[۲]، سیدعلی حسینی[۱]
[۱]دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران، ایران،
[۲] آزمایشگاه هوش مصنوعی پژوهشکده علوم اسلامی و انسانی دیجیتال (نور)، قم، ایران

چکیده

ریشه‌یابی مرحله اصلی چندین فرایند پردازشی مانند متن‌کاوی، بازیابی اطلاعات و پردازش زبان طبیعی‌است. ابزارهای تشخیص میانوند واژگان عربی با چالش‌های زیادی روبه‌رو هستند که بیشتر ناشی از ماهیت پیچیده واژگان این زبان و سبک‌های نوشتاری متفاوت آن‌ها است. تا جایی که ما می‌دانیم، هیچ مجموعه‌دادهٔ ریشه‌یابیِ معیاری وجود ندارد که طیف گسترده‌ای از چالش‌های ریشه‌یابی را پوشش دهد؛ بنابراین، ما توسعه یک مجموعه‌داده برای ارزیابی پایداری ریشه‌یاب‌ها را در چنین موقعیت‌های چالش برانگیزی ارزشمند می‌دانیم. این مقاله، نور-استم، یک مجموعه‌داده معیار با سبک‌های نوشتاری مختلف را برای ارزیابی ابزارهای تشخیص میانوند (استم) عربی معرفی می‌کند. جهت تأیید عملکرد این دادگان، عملکرد سه ریشه‌یاب عربی (لایت ۱۰، NLTK و تاشفین) مورد ارزیابی قرار گرفته‌است. نتایج نشان می‌دهد که سنجهٔ اف در ریشه‌یاب تاشفین بهتر از سایر ریشه‌یاب‌ها است که این موضوع در پژوهش‌های مرتبط نیز مشاهده شده‌است.

واژگان کلیدی: دادگان معیار، ریشه‌یاب، نور-استم، میانوند، استخراج اطلاعات

Noor –stem v.1 A Benchmark Dataset for Evaluating the Arabic Stemmers

Azal Alaswad[1], Behrouz Minaei-Bidgoli[1*], Mohammad E. Shenassa[2], Habib Seryani[2], Sayyed Ali Hossayni[1]
Schoolf of Computer Engineering, Iran University of Science and Technology, Tehran, Iran
Artificial Intelligence Laboratory of Digital Humanities and Islamic Sciences Research Institute (Noor), Qom, Iran

Abstract

The main task of the tokenization is to divide the sentences of the text into its constituent units and remove punctuation marks (dots, commas, etc.). Each unit is a continuous lexical or grammatical writing chain that is an independent semantic unit. Tokenization occurs at the word level and the extracted units can be used as input to other components such as stemmer. Stemming is the main step of several processing tasks such as text mining, information retrieval, and natural language processing .

Arabic stemmers face many challenges, mostly caused by the complex nature of Arabic words and their different writing styles. To our knowledge, there is no gold stemming dataset, which contains a wide variety of different possible stemming challenges, so that, stemmers face numerous and different possible real-world challenges to stem the words. Thus, we find it valuable to develop a dataset for evaluating the sustainability of stemmers in such a variety of challenging situations. In this paper, we introduce Noor-Stem, a benchmark dataset with various writing styles for the evaluation of Arabic stemmers. We use two thousand Arabic words in this dataset. We choose the words from different sources such as holy Quran as well as the Arabic websites and assign them to two groups of human experts to determine the correct stem for each word. The first chosen collection of words includes non-

* نویسندهٔ عهده‌دار مکاتبات
* Corresponding author

• تاریخ ارسال مقاله:۱۴۰۱/۸/۱۱ • تاریخ پذیرش: ۱۴۰۲/۱۰/۲۴ • تاریخ انتشار: ۱۴۰۳/۵/۱۰ • نوع مطالعه: پژوهشی

[DOI: 10.61186/jsdp.21.1.101]

repetitive words of the Quran according to their morphological structure. This collection, with more than 16,000 words, is completely by its Quranic usage, labeling only the words stems. The necessity of morphological analysis in Quranic texts as an example of the index of classical Arabic texts has given rise to this evaluation. The second word collection includes 10 thousand words from the non-repetitive words of the text data in general classic Arabic texts. Out of more than 2,600,000 non-repetitive words, considering that the dataset is going to be gold and each stem must be labeled/ensured by a couple of experts, 10,000 words are chosen, regarding the comprehensive and unique patterns to fully measure the length. The variety of patterns can face each stemmer with a serious challenge to demonstrate its performance in various processes. We evaluate the performance of three Arabic stemmers (Light 10, NLTK and Tashaphyne) on this dataset. The results show that the F-measure of Tashaphyne is better than the other stemmers, which re-proves the superiority of this stemmer in this type of problem, as well.

Keywords: Benchmark Dataset, Stemmer, Noor-Stem, Infixes, Information Retrieval

۱- مقدمه

ریشه‌یابی[1] یا استخراج استم، یک الگوریتم محاسباتی است که واژگان را با تهی‌کردن آن‌ها از پیشوندها و پسوندها به یک شکل مشترک باز می‌گرداند که گاه با معادل فارسی «پیراسته»، «میانوند» یا «هسته» نیز مورد اشاره قرار می‌گیرد. از این الگوریتم در زبان عربی نیز با تعبیر «التجذیع» یاد می‌شود که فرایندی متفاوت از پیداکردن ریشه[2] به معنای حروف اصلی واژه است؛ برای مثال واژگان عربی «یعلمون»، «سیعلمان»، «یعلمن»، «فیعلمونه» و «أیعلم؟» که همگی از ریشه «ع ل م» هستند، دارای استم مشترک «یعلم» هستند (سریانی و هاشمی ۱۳۹۹). ازاین‌رو ریشه‌یابی فرآیند فشرده‌سازی واژگان به اشکال اصلی آن‌ها، با حذف بخش‌های غیر اصلی واژه‌است.

ریشه‌یابی اندازه شاخص را کاهش می‌دهد؛ زیرا از مزیت کاهش فضای ذخیره‌سازی با حذف واژگان اضافی برخوردار است [۱]. این کار باعث بهبود نتایج می‌شود و احتمال تطابق برای جستجوی واژگان را افزایش می‌دهد. ریشه‌یابی یک فرآیند مهم در بسیاری از وظایف پردازش زبان طبیعی مانند بازیابی اطلاعات، موتورهای جستجوی وب، پاسخ به سؤال، طبقه‌بندی متون، ساخت خودکار لغت نامه، خلاصه‌سازی متن و غیره است [۲]. برای توسعه اصول ریشه‌یابی، باید تمام قواعد زبان را درک کنیم. ریشه‌یابی در طیف وسیعی از زبان‌ها مانند عربی، انگلیسی، فرانسوی، چینی، آلمانی، فرانسوی، ایتالیایی استفاده می‌شود [۳]. زبان عربی یکی از زبان‌های رایج در جهان است [۴]. این زبان متعلق به شاخه زبان‌های سامی است که خانواده‌ای از گروه زبان‌های آسیایی-آفریقایی محسوب می‌شود [۵].

ریشه‌یاب‌های عربی با چالش‌های مهمی روبه‌رو هستند که در بخش بعدی به‌اختصار آن‌ها را برشمرده و معرفی می‌کنیم. فرآیند ریشه‌یابی در زبان عربی دشوارتر از زبان‌های دیگر مانند انگلیسی است؛ زیرا واژگان زبان عربی دارای ساختاری با دو جنسیت مذکر و مؤنث و سه نوع تعداد مفرد، مثنی و جمع هستند. یک جمله عربی دارای دو نوع است: جمله اسمی و جمله فعلی. ساختار یک اسم عربی براساس جنسیت، عدد و اعراب دستوری آن در دو فرض معرب (مرفوع، منصوب و مجرور) و مبنی تعیین می‌شود [۶].

۱-۱- چالش‌های ریشه‌یاب‌های عربی

ریشه‌یاب‌های عربی با مجموعه‌ای از چالش‌ها مواجه هستند که در ادامه به معرفی و بررسی آن‌ها می‌پردازیم.

۱-۱-۱- حذف نادرست

واژگانی که حاوی حروف ابتدایی شبیه به برخی پیشوندها یا حروف پایانی شبیه به برخی پسوندها هستند، در معرض حذف نادرست قرار می‌گیرند. برای مثال، لایت ۱۰ [۷] با درنظرگرفتن "و + ال" به عنوان یک پیشوند و "ه" به‌عنوان یک پسوند، واژه "والده" را به یک استم غلط تبدیل می‌کند؛ درحالی‌که "وال" بخشی از استم درست است. این مسأله همچنین در مورد واژه "ألم" و دیگر موارد مشابه نیز وجود دارد.

۱-۱-۲- تک‌جواب بودن

بیشتر ریشه‌یاب‌های عربی تنها یک راه حل را به‌عنوان خروجی استم خود ارائه می‌دهند؛ درحالی‌که قواعد دستوری تجزیه واژگان عربی بر وجود بیش از یک استم دلالت دارد. به‌عنوان مثال، واژه "لهم" را می‌توان به چهار شکل مختلف در نظر گرفت: فعل "لَهِمَ" (او بلعید)، اسم

[1] Stemming
[2] Root

[DOI: 10.61186/jsdp.21.1.101]

"لَهِمٌ " (پرخور) و فعل "هَمَّ" (او غمگین شد) و (ل به‌عنوان پیشوند) و ضمیر "هُم" (آن ها). برای رسیدن به راه حل صحیح، ریشه‌یاب باید تمام حالت‌های صحیح واژه را در نظر بگیرد [۸].

۳-۱-۱- تغییرات واژگان

قوانین پیچیده‌ای در زبان عربی مثل ادغام، اعلال و تخفیف وجود دارد که منجر به تغییر کامل واژه در ترکیب واژگان عربی و ساخت اشکال جدید می‌شود. این موضوع در زبان عربی بیشتر از زبان‌های دیگر دیده می‌شود. برای مثال، در زمان گذشته فعل "نام " (خوابید) از ریشه "ن و م" هنگامی که برای مخاطب به کار می‌رود، ساختار آن به‌طور کامل تغییر کرده و با توجه به قواعد افعال معتل، به "نمتَ" (خوابیدی) یا "نمتُما " (خوابیدید) (برای دونفر) تغییر می‌یابد. چنین تغییراتی در واژه باعث می‌شود تا استم واژه به‌سختی پیدا شود.

۴-۱-۱- ابهام ناشی از حذف حرکت‌ها[1]

حذف حروف اول و آخر بدون درنظرگرفتن حرکت‌ها منجر به ساخت یک ریشه غلط خواهد شد؛ برای مثال، تجزیه و تحلیل ریخت‌شناسی واژه "لِعَینٍ" (برای چشم) بدون درنظرگرفتن حرکت حروف منجر به ساخت "ل" به‌عنوان پیشوند حذف شده و "عین" به عنوان ریشه خواهد شد؛ درحالی‌که این واژه ممکن است به شکل "لَعینٌ" (ملعون) باشد که در آن "ل" حرف اصلی واژه‌است. برای مثالی دیگر، برای واژه "بِکر" به معنای تازه ممکن است یک ریشه‌یاب به اشتباه حرف "ب" را به‌عنوان پیشوند حذف کرده و استم "کر" را ارائه دهد.

۵-۱-۱- ساختارهای بی‌قاعده

برخی تغییرات ساختاری که در واژگان عربی ایجادشده‌اند، از قاعده خاصی پیروی نمی‌کنند. یک مثال واضح از این ساختارها را می‌توان در شکل جمع بیش‌تر واژگان عربی مشاهده کرد. در جمع قاعده‌مند، حروف خاص "ین، ون، ات" به واژگان اضافه می‌شوند و شکل جمع این واژگان با چند تغییر ایجاد می‌شود؛ مانند "قارئون" (خوانندگان) و "قارئات" (خوانندگان زن) که فرم‌های جمع "قارئ " و "قارئهٔ " هستند. ریشه‌یاب‌های لایت می‌توانند با حذف "ون" و "ات" به این واژگان دست یابند، اما در اسم‌های جمع مکسر، ساختار واژه کاملا تغییر می‌کند و ریشه‌یاب‌ها اغلب قادر به تشخیص صحیح استم نیستند. به عنوان مثال، واژه‌های "نساء " (زنان) و "ابواب" (درها) که جمع واژه‌های "امرأة" (زن) و "باب" (در) هستند را نمی‌توان با حذف حروف یا پیروی از هر قانون خاص دیگری به شکل مفرد آن‌ها تغییر داد (سریانی، هاشمی ۱۳۹۹).

۶-۱-۱- استفاده از صفحه کلید غیر عربی

امروزه منابع متنی عربی زیادی وجود دارد که در آن‌ها واژگان با صفحه کلید غیر عربی مانند صفحه کلید فارسی نوشته می‌شوند. به عنوان مثال در کیبورد فارسی تنها یک نوع (یاء) وجود دارد، در حالی که در زبان عربی دو نوع از آن وجود دارد؛ بنابراین، برخی از واژگان با نوع (یاء) اشتباه نوشته می‌شوند که باعث می‌شود ریشه‌یاب نتواند ریشه صحیح را استخراج کند. همین مورد در مورد زبان فارسی (حرف کاف) صادق است که دو نسخه مشابه عربی برای این حرف وجود دارد.

۲-۱- دستاورد و ساختار

در این بخش، جنبه‌های نوآورانه این مقاله و همچنین ساختار کلی آن را شرح می‌دهیم:

۱-۲-۱- دستاورد این مقاله

هدف این مقاله پیشنهاد "مجموعه‌داده ریشه‌یابی-نور" به‌عنوان یک مجموعه‌داده، حاوی حروف عربی و غیرعربی برای ارزیابی پایداری ریشه‌یاب‌های عربی در چالش‌های ذکر شده‌است. ما جنبه‌های آماری مجموعه‌داده ریشه‌یابی-نور را مورد بحث قرار می‌دهیم و عملکرد هر ریشه‌یاب را ارزیابی می‌کنیم و آن را بر روی مجموعه‌داده آزمایش می‌کنیم.

۲-۱-۱- ساختار مقاله

بقیه این مقاله به‌صورت زیر سازماندهی شده‌است: ما روش‌های ریشه‌یابی و مجموعه‌داده‌های موجود را در بخش ۲ معرفی می‌کنیم؛ سپس، مجموعه‌داده پیشنهادی را در بخش ۳ توضیح می‌دهیم. بخش ۴ مجموعه‌ای از ریشه یاب‌ها را به‌عنوان روش‌های معیار برای مقایسه نتایج آن‌ها با یکدیگر معرفی می‌کند. در بخش ۵، نتایج را ارایه و مورد بحث قرار می‌دهیم و درنهایت، در بخش ۶ مقاله را با یک نتیجه‌گیری به پایان می‌رسانیم.

[1] Diacritics

[DOI: 10.61186/jsdp.21.1.101]

۳- پیشینهٔ پژوهش

در این بخش، ما پژوهش‌های انجام‌شده مربوط به ریشه‌یابی عربی و برخی از مجموعه‌داده‌های موجود را که برای ارزیابی ریشه‌یاب‌های عربی طراحی شده‌اند، توضیح می‌دهیم.

۲-۱- روش‌های ریشه‌یابی موجود

پژوهش‌گران مطالعات گسترده‌ای را روی کارایی تکنیک‌های ریشه‌یابی در طبقه‌بندی متن عربی و ارزیابی دقت‌ها انجام داده‌اند. نویسندگان یک مقاله [۹]دقت طبقه‌بندی را با استفاده از دو ریشه‌یاب پرطرفدار خوجهٔ (Khoja) و لایت (Light) با تکیه بر تکنیک استخراج واژگان مقایسه کرده و ثابت کردند که ریشه‌یاب لایت دقت بالاتری را با استفاده از طبقه‌بند بهینه‌سازی حداقلی ترتیبی[1] ارائه‌می‌دهد. نویسندهٔ دیگری [۱۰]یک روش جدید برای ریشه‌یابی واژگان عربی، با توسعه تکنیک ریشه‌یابی لایت ارائه داده که به فهرست‌های ازپیش‌تعریف‌شده و منظم از پسوندها و پیشوندها وابسته‌است. این مقاله فراسا (Farasa) را معرفی می‌کند که از SVM (ماشین‌های بردار پشتیبان) برای رتبه‌بندی استفاده می‌کند. کار فراسا شامل بخش‌بندی واژه به پیشوند + ریشه + پسوند است [۱۱].

نویسندگان [۱۲]اثر وظایف پیش‌پردازش را با اثرات ترکیب آن‌ها بر طبقه‌بندی اسناد عربی مورد مطالعه قرار دادند و نشان دادند که وظایف پیش‌پردازش شامل ریشه‌یابی لایت، نرمال‌سازی و حذف واژگان بی‌اثر[2]است. مطالعه [۱۳]یک روش جدید را برای تشخیص ریشه با استفاده از یک نسخه توسعه‌یافته از ریشه‌یاب خوجهٔ ارائه‌می‌دهد. سه مرحله پردازش اصلی برای تولید ریشه عربی از واژگان به کار می‌رود. مرحلهٔ ۱ مسئول حذف پیشوندها و پسوندها است؛ مرحلهٔ ۲ خروجی را با منابع یا اشکال استاندارد واژگان مقایسه می‌کند و مرحلهٔ ۳ به اصلاح استم استخراج‌شده می‌پردازد. مقاله [۱۴] اثرات تکنیک‌های ریشه‌یابی را بر طبقه‌بندی اسناد عربی بررسی می‌کند. وظایف پیش‌پردازش مانند نرمال‌سازی، حذف واژگان بی‌اثر و تکنیک‌های ریشه‌یابی نیز در آن گنجانده شده‌است. مطالعه پژوهشی [۱۵] اثرات تکنیک‌های ریشه‌یابی لایت را بر استخراج ویژگی‌هایی مانند جعبه واژگان[3] و TF - IDF برای طبقه‌بندی اسناد عربی مطالعه می‌کند.

نویسندگان [۱۶] با استفاده از چندین طبقه‌بند باناظر[4]، تأثیر ریشه‌یابی را بر طبقه‌بندی نمونه ترجمه‌ها و تفاسیر قرآن کریم بررسی و تحلیل کردند. این چارچوب طبقه‌بندی شامل مراحل پیش‌پردازش متن، استخراج ویژگی و طبقه‌بندی متن است. همچنین، ریشه‌یاب Sastrawi برای انجام عملیات ریشه‌یابی در مرحله پیش‌پردازش متن استفاده می‌شود. این مقاله روش جدیدی را برای استخراج استم عربی به نام بازیابی اطلاعات ریخت‌شناسی عربی (AMIR) معرفی و سپس عملکرد آن را با فاراسا و LUCENE از نظر میانگین دقت مقایسه می‌کند[۱۷].

۲-۲- دادگان‌های موجود

دادگان‌های متعددی برای ریشه‌یابی عربی وجود دارد که به‌اختصاربه شرح زیر توضیح می‌دهیم:

دادگان CBAS. مجموعه‌داده CBAS از مقالات موجود در روزنامه‌های عمان استخراج شده‌است. این مجموعه‌داده شامل ۲۰۲۹۱ مقاله از موضوعات مختلف مانند فرهنگ و ورزش است [۱۸].

دادگان LABR (نقد کتاب‌های عربی در مقیاس بزرگ). مجموعه LABR شامل بیش از ۶۳۰۰۰ نقد کتاب به زبان عربی است که هر یک از آن‌ها از ۱ تا ۵ ستاره رتبه‌بندی شده‌اند. این مجموعه‌داده در دو حوزهٔ طبقه‌بندی قطبیت احساسات[5] و ریشه‌یابی مورد استفاده قرار گرفته‌است. این مجموعه‌داده در دو حالت تقسیم‌بندی متعادل و نامتعادل بررسی شده‌است. برای دسترسی به این داده‌ها می‌توان به سایت Goodreads.com مراجعه نمود [۱۹].

مجموعه‌داده ICA (مجموعه‌داده بین‌المللی عربی). مجموعه ICA شامل (۱۰۰ میلیون) واژه است. مجموعه‌ای از نمونه‌ها از طیف گسترده‌ای از منابع انتخاب شده و سطح گسترده‌ای از زبان عربی را پوشش می‌دهند [۲۰].

مجموعه‌داده مخزن درخت وابستگی عربی ARL. داده‌های منبع در این مجموعه‌داده شامل اخبار عربی و برنامه‌های پخش‌شده توسط LDC از شبکه‌های خبری و بخش‌های مختلف هستند. فایل‌ها در قالب یک تب ۱۱ ستونی جداشده با یک یا چند خط خالی بین جملات مرتب شده‌اند. همه فایل‌ها دارای کدگذاری UTF-8 هستند [۲۱].

[1] Sequential Minimal Optimization
[2] Stop words
[3] Bag-of-words
[4] Supervised Classifier
[5] Sentiment Polarity

[DOI: 10.61186/jsdp.21.1.101]

۳- معرفی دادگان نور-استم

بررسی عملکرد هر ماشین تحلیل ریخت‌شناسی نیازمند اندازه‌گیری و مقایسه آن با یک مجموعه‌داده معیار است که شامل انواع واژگان مورد استفاده در زمینه‌های مختلف علمی است. با توجه به بررسی ما، مجموعه‌داده‌های معیار موجود شامل انواع کاملی از چالش‌های ریخت‌شناسی برای ریشه‌یاب‌ها نیستند؛ بنابراین، معرفی یک مجموعه‌داده معیار جدید، شامل چالش‌های احتمالی مختلف برای ریشه‌یاب‌ها ارزشمند است، به‌طوری که بتواند ریشه‌یاب‌های عربی مختلف را ارزیابی کند و اثربخشی آن‌ها را بر روی چنین مجموعه‌داده جامعی نشان دهد. در این بخش، ابتدا ساختار این مجموعه‌داده را توضیح می‌دهیم و سپس به ویژگی‌های آماری آن اشاره می‌کنیم. ما به‌طورتقریبی از دو هزار واژه عربی در این مجموعه‌داده استفاده و همچنین واژگان را از منابع مختلف مانند قرآن کریم و پایگاه‌های اینترنتی عربی انتخاب می‌کنیم و آن‌ها را به دو گروه از متخصصان انسانی ارجاع می‌دهیم تا استم صحیح هر واژه را تعیین کنند.

نخستین مجموعه انتخاب‌شده از واژگان شامل واژگان غیر تکراری قرآن با توجه به ساختار ریخت‌شناسی آن‌ها است. این مجموعه، با بیش از شانزده‌هزار واژه، به‌طورکامل مطابق با کاربرد قرآنی آن است و تنها استم واژگان را برچسب‌گذاری می‌کند. ضرورت تحلیل ریخت‌شناسی در متون قرآنی به‌عنوان نمونه‌ای شاخص از متون کلاسیک عربی، اعتبار این ارزیابی را افزایش داده‌است.

مجموعه واژگان دوم شامل ۱۰ هزار واژه از واژگان غیر تکراری داده‌های متنی در متون کلاسیک عربی‌است. از بین بیش از ۲،۶۰۰،۰۰۰ واژه غیر تکراری، (هر ریشه باید توسط چند متخصص برچسب گذاری / تضمین شود)، ده هزار واژه با توجه به الگوهای جامع و منحصربه‌فرد انتخاب می‌شوند. تنوع الگوها می‌تواند هر ریشه‌یابی را با چالشی جدی مواجه کند تا عملکرد آن را در فرایندهای مختلف نشان دهد. به‌عنوان چند نمونه از این الگوهای خاص می‌توان به اسم‌های جمع مکسر، حذف ضمایر، واژگان مضاعف و معتل و نسخه‌های مختلف نوشتاری قرآن کریم و ... اشاره کرد. جدول (۱) شامل این الگوهای خاص‌است. برای هر یک از ۳۴ الگوی زیر، یک فهرست تصادفی در مجموعه‌داده معیار تشکیل شده‌است.

(جدول-۱). تنوع الگوها

(Table-1). The variety of patterns

ردیف	الگوها
۱	یک مولفه تصادفی مطلق از کل داده‌ها
۲	واژگانی که مساوی یا کمتر از ۵ بار تکرار شده‌اند، به منظور ارزیابی ریشه‌یاب‌ها در واژگان کم تکرار (واژگان دیگر داده معیار همه بیش از ۵ بار ظاهر شده اند)
۳	اسم‌های دخیل یا معرب که از زبان دیگری وارد شده و علم نیستند
۴	اسم‌های دخیل یا معرب که از زبان دیگری وارد شده و علم هم هستند
۵	اسم‌های علمی که اختصاص به زبان عربی دارند
۶	واژگان مبنی یعنی واژگانی که حرکت آخر ثابت است
۷	واژه‌های رایج غیر عربی که توسط نویسندگان مشهور ابداع شده اند، مانند «اللابشرط» که توسط نویسنده بسیار معروف شیخ انصاری ساخته شده‌است.
۸	حالت‌های نوشتاری خاص واژگان قرآنی که با دستور زبان کلاسیک متفاوت هستند (مثلاً «بقیت الله» که شکل کلاسیک آن «بقیهٔ اَللّه» است).
۹	واژگانی که در زبان عربی به آن‌ها اسم فعل اطلاق می‌شود؛ مانند «هیهات»، «آمین» و «سرعان» و «ویلک».
۱۰	واژگانی که از عربی به فارسی وارد شده و در کتاب‌ها استعمال شده‌است
۱۱	واژگانی با ریشه ۳ حرفی که حرف وسط آنها «و» یا «ی» است (معتل اجوف)
۱۲	واژگانی با ریشه ۳ حرفی که حرف آخر آنها «و» یا «ی» است (معتل ناقص)
۱۳	واژگانی با ریشه ۳ حرفی که حرف اول آنها «و» یا «ی» است (معتل مثال)
۱۴	واژگانی با ریشه سه حرفی که حداقل دو حرف ریشه "و" یا "ی" است (معتل لفیف)
۱۵	واژگانی با ریشه ۳ حرفی با دو حرف یکسان در کنار هم (واژگان مضاعف)
۱۶	واژگانی با ریشه ۳ حرفی با حداقل یک "ء" (واژگان مهموز)
۱۷	واژگانی که بیش از سه حرف ریشه دارند (واژگان رباعی)
۱۸	واژگانی که حداقل دارای یک «أ» یا «إ» هستند ولی ریشه آنها شامل «ء» نباشد.
۱۹	واژگانی دارای حداقل یک "ئ" یا "ؤ" یا "ء" که ریشه آنها شامل «ء» نباشد.
۲۰	واژگانی دارای حداقل یک «ی» یا «و» که ریشه آنها شامل «و» و «ی» نباشد.
۲۱	واژگان با پیشوند «ل» یا «ک» یا «ب» یا «ف» یا «س» که ریشه آنها شامل این حروف نمی‌شود.
۲۲	واژگان با دو پیشوند از فهرست «ل» یا «ک» یا «ب» یا «ف» یا «س» که ریشه آن شامل هیچ یک از آنها نمی‌شود.
۲۳	واژگانی با پسوند «ک» یا «ه» یا «و» یا «ن» یا «ی» یا «م» یا «ا» که ریشه آن‌ها شامل این حروف نیست.
۲۴	واژگانی با دو پسوند از فهرست «ت» یا «ک» یا «ه» یا «و» یا «ن» یا «ی» یا «م» یا «ا» که ریشه آن‌ها شامل هیچ یک از آنها نمی‌شود.
۲۵	واژگانی که حرف آخر آنها «ة» مربوطه‌است.
۲۶	واژگانی که دو حرف آخر آنها «یة»است.

[DOI: 10.61186/jsdp.21.1.101]

کلماتی که حرف آخر آنها «ی» است و در ریشه حروف آنها «و» و «ی» وجود ندارد (واژگان منقوص)	۲۷
واژگان با پسوند «ان» یا «ات» یا «ون» یا «ین»	۲۸
واژگانی که «ی» یا «و» یا «ء» در ریشه آنها وجود نداشته و کمتر از ۳۰ بار در دیتا تکرار شده‌اند.	۲۹
واژگانی که برای آن‌ها بیش از دو ریشه محتمل باشد؛ مانند «بکر» و «یعدون»	۳۰
واژگانی با اشکال صرفی بسیار پیچیده که توسط یک زبانشناس پیشنهاد شده‌اند	۳۱
واژگانی که یکی از واژگان موجود در داده جمع مکسر را در درون خود دارند؛ مانند «بکتبهم»	۳۲
واژگانی که یکی از واژگان موجود در داده‌های ثبتی غیر از جمع مکسر را در درون خود دارند؛ مانند «ربّنا»	۳۳
واژگانی که ریشه‌یابی خودکار آنها (از طریق آنالیزورهای صرفی) اغلب با شکست مواجه می‌شود	۳۴

پس از بررسی خروجی، تشخیص دو گروه به‌صورت زیر توزیع شده‌است:

جدول (۲) ویژگی‌های آماری دادگان نور-استم را از دیدگاه ویژگی‌های ریخت‌شناسی نشان‌می‌دهد.

(جدول-۲). آمار دادگان نور-استم

(Table-2).the statistics of the Noor-stem dataset

میانگین طول واژگان (حرف)	تعداد لغات با پیشوند	تعداد فعل ها	تعداد اسم ها
۶	۸۲۱	۲۲۰	۴۱۳

بنابراین، در مجموع ۱۸۴۷ واژه جمع آوری شده‌است، درحالی‌که ۷۶ واژه دارای قضاوت‌های متفاوتی بین گروه‌ها هستند. این واژگان با قضاوت متفاوت از مجموعه‌داده حذف می‌شوند و واژگان باقی مانده به عنوان مجموعه‌داده معیار شامل ۱۷۷۱ واژه پیشنهاد می‌شوند.

۴-روش‌های معیار و شاخص‌های ارزیابی

در این بخش به معرفی روش‌های معیار و شاخص‌های ارزیابی استفاده‌شده بر روی دادگان معیار می‌پردازیم.

۴-۱- روش‌های معیار

ما مجموعه‌ای از ریشه‌یاب‌های شناخته‌شده را انتخاب کرده و در مقایسه خود از آن‌ها استفاده می‌کنیم. ریشه‌یاب‌های استفاده شده در این مقایسه در ادامه آمده‌اند.

۴-۱-۱-تاشفین

ریشه‌یاب تاشفین ابتدا واژگان را نرمال سازی کرده و حرکت‌ها و مدها [1] را از واژگان ورودی حذف می‌کند. سپس، ورودی را با استفاده از یک لیست جستجوی ترکیبی برای سطوح مختلف ریشه‌یابی، بخش‌بندی و ریشه‌یابی می‌کند [۲۲].

۴-۱-۲- لایت-۱۰

یکی از ویژگی‌های اصلی لایت ۱۰، ریشه‌یابی زیررشته-هایی است که از یک سو اغلب به عنوان پیشوندها / پسوندها یافت می‌شوند و از سوی دیگر در ابتدای / انتهای ریشه‌ها استفاده نمی‌شوند. لایت ۱۰ پیشوندهایی مانند (وال، اَل، فَال، بَالَ، کال، وَ) و پسوندهایی مانند (ها، اَن، اَت، وَن، یَهُ، یَه، هُ، هَ، یَ) را حذف می‌کند [۲۳].

۴-۲- شاخص‌های ارزیابی

برای ارزیابی سه ریشه‌یاب ما از معیار دقت[2] (معادله ۱)، فراخوانی[3] (معادله ۲) و سنجهٔ اف[4] (معادله ۳) استفاده می‌کنیم [۲۴].

$$\text{Precision} = \frac{\text{correct}}{\text{correct} + \text{incorrect}} \quad (۱)$$

$$\text{Recall} = \frac{\text{correct}}{\text{correct} + \text{no} - \text{stem}} \quad (۲)$$

$$\text{F} - \text{measure} = \frac{2 * \text{Precision} * \text{Rrecall}}{\text{Precision} + \text{Recall}} \quad (۳)$$

که در آن *correct* تعداد ریشه‌های صحیح، *incorrect* تعداد ریشه‌های ناصحیح و *no-stem* تعداد ریشه‌های صحیح است که به‌وسیلهٔ ریشه‌یاب تشخیص داده نشده‌اند.

۵- نتایج و بحث

ما سه ریشه‌یاب مختلف را بر روی مجموعه‌داده ریشه‌یابی-نور اعمال می‌کنیم (ریشه‌یاب تاشفین، لایت ۱۰، NLTK). نتایج آزمایش‌های ذکرشده در جدول (۴) گزارش شده‌اند. همان‌طورکه مشاهده می‌شود، سنجهٔ اف ریشه‌یاب تاشفین برتری معناداری نسبت به دیگر نمونه‌ها دارد. سنجهٔ اف تاشفین ۵۲ درصداست. این دقت به دست آمده بسیار کم‌تر از عملکرد این ریشه‌یاب در سایر مجموعه‌های داده است. دلیل آن، نسبت زیاد واژه‌های دشوار در مجموعه‌داده نور است؛ چنانچه مجموعه‌داده نور

[1] Elongation
[2] Precision
[3] Recall
[4] F-Measure

[DOI: 10.61186/jsdp.21.1.101]

شامل واژه‌های عربی استاندارد مدرن (MSA) و عربی کلاسیک (CA)است. مثلا در این مقاله [۲۵] دقت آزمایش لایت ۱۰ و بر روی دیتاست بیشتر از دیتاسیت نور استم بوده‌است. در این مقاله از روزنامه های اینترنتی عربی که فقط شامل واژه‌های عربی استاندارد مدرن (MSA) است استفاده می‌کنند. نتایج این مقاله در جدول (۳) قابل مشاهده‌است.

(جدول-۳).مقایسه عملکرد ریشه‌یاب‌های عربی در مقاله [۲۵]

(Table-3). Comparison of the performance of Arabic in other articles stemmers

ریشه‌یاب‌ها	دقت	فراخوانی	سنجهٔ اف
P- stemmer	0.90	0.94	0.92
کوجا	0.78	0.80	0.79
لایت ۱۰	0.83	0.83	0.82

دلیل آن وجود واژگان بسیار سخت در دیتاست نوراست. مثلا کلماتی مانند بقیهٔ الله, آیهٔ الله, عصبصب (روز فوق العاده گرم) و الجلجال (کسی با صدای قوی) از این دست هستند.

(جدول-۴).مقایسه عملکرد ریشه‌یاب‌های عربی

(Table-4).Comparison of the performance of Arabic stemmers

ریشه‌یاب‌ها	دقت	فراخوانی	سنجهٔ اف
تاشفین	0.45	0.63	0.524
لایت ۱۰	0.37	0.57	0.448
NLTK	0.26	0.41	0.256

با این حال، نتایج به دست آمده، نتایج [۲۶] را تاییدکردند. برای ارائه یک دید کلی در مورد عملکرد هر ریشه‌یاب، جدول (۵) نمونه‌ای از خروجی بدست آمده را نشان‌می‌دهد.

(جدول-۵).خروجی ریشه‌یاب‌ها برای برخی از واژگان نمونه ریشه‌یاب-نور

(Table-5).the output of the stemmers for some of Noor-Stem sample words

واژه	تاشفین	لایت ۱۰	NLTK
عنه	عنه	عنه	عنه
مما	ما	مما	مما
يحيىَ	حيى	حيا	يحيى
الشيطان	شيطان	شوط	شيطان
امارات	مارات	مرا	امارات
أصله	صله	اصل	صله
لسوف	وف	سوف	سوف
فمر	مر	مرر	مر
فكأنَ	كان	كان	ان
لندنَ	دن	ندا	ندن

۶- نتیجه‌گیری

ریشه‌یابی فرآیند استخراج استم یا میانوند از یک فعل یا یک اسم‌است. برخی زبان‌ها مانند عربی از نظر ریخت‌شناسی نسبت به هر زبان دیگری پیچیده‌تر هستند. در این مقاله، ما ریشه‌یاب-نور را به‌عنوان یک مجموعه‌داده معیار برای روش‌های ریشه‌یابی عربی معرفی می‌کنیم. ما عملکرد سه ریشه‌یاب عربی تاشفین، لایت ۱۰ و NLTK را روی این مجموعه‌داده با یکدیگر مقایسه می‌کنیم. نتایج نشان دهنده برتری سنجهٔ اف تاشفین در مقایسه با سایر روش‌ها است.

۷- مراجع 7-Refrenca

۱ .سریانی، حبیب؛ هاشمی، سید محسن. استمر «نور»؛ موتور هوشمند تشخیص میانوند واژگان عربی، هوش مصنوعی و علوم اسلامی، دوره ۱، ۵۴ صفحه

[1] N. Y. Habash, “Introduction to Arabic natural language processing,” Synth. Lect. Hum. Lang. Technol., vol. 3, no. 1, pp. 1–187, 2010.

[2] H. Alshalabi, S. Tiun, N. Omar, F. N. AL-Aswadi, and K. A. Alezabi, “Arabic light-based stemmer using new rules,” J. King Saud Univ. Inf. Sci., 2021.

[3] D. H. Abd, W. Khan, K. A. Thamer, and A. J. Hussain, “Arabic Light Stemmer Based on ISRI Stemmer,” in International Conference on Intelligent Computing, 2021, pp. 32–45.

[4] M. N. Al-Kabi, S. A. Kazakzeh, B. M. A. Ata, S. A. Al-Rababah, and I. M. Alsmadi, “A novel root based Arabic stemmer,” J. King Saud Univ. Inf. Sci., vol. 27, no. 2, pp. 94–103, 2015.

[5] S. Levin, “Toward Proto-Nostratic: A new approach to the comparison of Proto-Indo-European and Proto-Afroasiatic. By Allan R. Bomhard,” Diachronica, vol. 2, no. 1, pp. 97–104, 1985.

[6] R. Mohammed, “New Arabic stemming based on Arabic patterns,” Iraqi J. Sci., vol. 57, no. 3C, pp. 2324–2330, 2016.

[7] L. S. Larkey, L. Ballesteros, and M. E. Connell, “Light stemming for Arabic information retrieval,” in Arabic computational morphology, Springer, 2007, pp. 221–243.

[8] Y. Jaafar, D. Namly, K. Bouzoubaa, and A. Yousfi, “Enhancing Arabic stemming process using resources and benchmarking tools,” J. King Saud Univ. Inf. Sci., vol. 29, no. 2, pp. 164–170, 2017.

[9] R. Mamoun and M. Ahmed, “Arabic text stemming: Comparative analysis,” in 2016 Conference of Basic Sciences and Engineering Studies (SGCAC), 2016, pp. 88–93.

[DOI: 10.61186/jsdp.21.1.101]

[10] R. A. Sameer, “Modified light stemming algorithm for Arabic language,” Iraqi J. Sci., vol. 57, no. 1B, pp. 507–513, 2016.
[11] A. Abdelali, K. Darwish, N. Durrani, and H. Mubarak, “Farasa: A fast and furious segmenter for arabic,” in Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Demonstrations, 2016, pp. 11–16.
[12] A. Ayedh, G. Tan, K. Alwesabi, and H. Rajeh, “The effect of preprocessing on arabic document categorization,” Algorithms, vol. 9, no. 2, p. 27, 2016.
[13] M. Mustafa, A. S. Eldeen, S. Bani-Ahmad, and A. O. Elfaki, “A comparative survey on arabic stemming: approaches and challenges,” Intell. Inf. Manag., vol. 9, no. 02, p. 39, 2017.
[14] Y. A. Alhaj, J. Xiang, D. Zhao, M. A. A. Al-Qaness, M. Abd Elaziz, and A. Dahou, “A Study of the Effects of Stemming Strategies on Arabic Document Classification,” IEEE Access, vol. 7, pp. 32664–32671, 2019, doi: 10.1109/ACCESS.2019.2903331.
[15] Y. A. Al[1] Y. A. Alhaj, M. A. A. Al-qaness, A. Dahou, M. Abd Elaziz, D. Zhao, and J. Xiang, “Effects of Light Stemming on Feature Extraction and Selection for Arabic Documents Classification,” in Studies in Computational Intelligence, vol. 874, Springer, 2020, M. A. A. Al-qaness, A. Dahou, M. Abd Elaziz, D. Zhao, and J. Xiang, “Effects of Light Stemming on Feature Extraction and Selection for Arabic Documents Classification,” in Studies in Computational Intelligence, vol. 874, Springer, 2020, pp. 59–79.
[16] F. S. Utomo, N. Suryana, and M. S. Azmi, “Stemming Impact Analysis On Indonesian Quran Translation And Their Tafsir Classification For Ontology Instances,” IIUM Eng. J., vol. 21, no. 1, pp. 33–50, 2020.
[17] A. Alnaied, M. Elbendak, and A. Bulbul, “An intelligent use of stemmer and morphology analysis for Arabic information retrieval,” Egypt. Informatics J., 2020.
[18] M. El-Defrawy, Y. El-Sonbaty, and N. A. Belal, “Cbas: Context based arabic stemmer,” arXiv Prepr. arXiv1611.00027, 2015.
[19] M. Nabil, M. Aly, and A. Atiya, “Labr: A large scale arabic sentiment analysis benchmark,” arXiv Prepr. arXiv1411.6718, 2014.
[20] “Building an International Corpus of Arabic (ICA): progress of compilation stage,” in 7th international conference on language engineering, Cairo, Egypt, 2007, pp. 5–6.
[21] M. Diab, N. Habash, O. Rambow, and R. Roth, “LDC Arabic treebanks and associated corpora: Data divisions manual,” arXiv Prepr. arXiv1309.5652, 2013.
[22] R. M. Sallam, H. M. Mousa, and M. Hussein, “Improving Arabic text categorization using normalization and stemming techniques,” Int. J. Comput. Appl, vol. 135, no. 2, pp. 38–43, 2016.
[23] M. I. Eldesouki, W. Arafa, and K. Darwish, “Stemming techniques of Arabic language: Comparative study from the information retrieval perspective,” Egypt. Comput. J., vol. 36, no. 1, pp. 30–49, 2009.
[24] F. N. Al-Aswadi, H. Y. Chan, and K. H. Gan, “Automatic ontology construction from text: a review from shallow to deep learning trend,” Artif. Intell. Rev., vol. 53, no. 6, pp. 3901–3928, 2020.
[25] T. Kanan, O. Sadaqa, A. Almhirat, and E. Kanan, “Arabic light stemming: A comparative study between p-stemmer, khoja stemmer, and light10 stemmer,” in 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), 2019, pp. 511–515.
[26] M. Naili, A. H. Chaibi, and H. H. Ben Ghezala, “Comparative study of Arabic stemming algorithms for topic identification,” Procedia Comput. Sci., vol. 159, pp. 794–802, 2019.

ازل العصواد مدرک کارشناسی را از دانشگاه ذیقار عراق و مدرک ارشد را از دانشگاه صنعتی امیرکبیر در رشته کامپیوتر دریافت کرد. وی در حال حاضر دانشجوی دکترای رشته هوش مصنوعی دانشگاه علم و صنعت ایران است. حوزه‌های پژوهشی ایشان پردازش زبان طبیعی است.
نشانی رایانامهٔ ایشان عبارت است از:
azal.alamery2@gmail.com

بهروز مینایی بیدگلی دکترای خود را از دانشکده مهندسی و علوم کامپیوتر دانشگاه میشیگان آمریکا دریافت کرد. او در حال حاضر استاد دانشکده مهندسی کامپیوتر دانشگاه علم و صنعت ایران است و هدایت دو گروه پژوهشی داده‌کاوی و بازی‌های رایانه‌ای را بر عهده دارد. حوزه‌های پژوهشی ایشان متن‌کاوی و پردازش زبان طبیعی‌است.
نشانی رایانامهٔ ایشان عبارت است از:
b_minaei@iust.ac.ir

محمدابراهیم شناسا مدرک کارشناسی را از دانشگاه علم و صنعت ایران و کارشناسی ارشد را از واحد علوم و تحقیقات در رشته مهندسی کامپیوتر دریافت کرد. وی در حال حاضر دانشجوی دکترای هوش مصنوعی و عضو هیئت علمی دانشگاه آزاد واحد تهران-شمال است.

[DOI: 10.61186/jsdp.21.1.101]

حوزه‌های پژوهشی مورد علاقه وی پردازش زبان طبیعی و داده‌کاوی است.

نشانی رایانامهٔ ایشان عبارت است از:
me.shenasa@iau-tnb.ac.ir

حبیب سریانی دانش‌آموخته حوزه علمیه قم بوده و مدرک دکترای خود را در رشته فقه و حقوق جزا دریافت کرده‌است. وی در حال حاضر استادیار گروه حقوق دانشگاه رازی کرمانشاه و پژوهش‌گر فعال در حوزه علوم اسلامی با رویکرد بهره‌مندی از فناوری‌های هوشمند نوین است.

نشانی رایانامهٔ ایشان عبارت است از:
Hsoryani@gmail.com

سید علی حسینی در سال ۱۴۰۰، موفق به اخذ مدرک پسادکتری رشته هوش مصنوعی از دانشگاه علم و صنعت ایران شد. وی همچنین مدرک دکترای هوش مصنوعی خود را از دانشگاه جیرونای اسپانیا، مدرک ارشد را از دانشگاه شهید بهشتی و مدرک کارشناسی را از دانشگاه صنعتی شریف در رشته کامپیوتر اخذ کرد. وی در حال حاضر سرپرست گروه پژوهشی آزمایشگاه پردازش دانش دانشگاه علم و صنعت ایران است.

نشانی رایانامهٔ ایشان عبارت است از:

[DOI: 10.61186/jsdp.21.1.101]