به نام خداوند مهربان

وزارت علوم، تحقیقات و فناوری

پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)

طرح پژوهشی

ساخت پیکرهٔ متنی از مقاله‌های پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

مجری

الهام علایی ابوذر

همکاران

نصرالله پاک‌نیت

علی‌اصغر حجت‌پناه

مجتبی زالی

محمدهادی آقالوی

فروردین ۱۴۰۰

وزارت علوم تحقیقات و فناوری
پژوهشگاه علوم و فناوری اطلاعات ایران
ایرانداک
پژوهش . مدیریت دانش . آموزش

شماره:
۳۳/۴۰۷
تاریخ:
۱۴۰۰/۰۱/۲۲
پیوست:
ندارد

بنام خدا

ایرانداک در نیم‌قرن دوم کار خود، همچنان جوان و کوشا، همراه هر پژوهش، پژوهشگر، و سیاست‌گذار
۱۳۹۹ - ۱۳۴۷

گواهی

انجام طرح پژوهشی:

- **عنوان:** ساخت پیکرهٔ متنی از مقاله‌های پژوهش‌نامهٔ پردازش و مدیریت اطلاعات
- **مجری:** دکتر الهام علایی ابوذر
- **همکاران:** دکتر نصرالله پاک‌نیت، علی‌اصغر حجت‌پناه، مجتبی زالی، محمدهادی آقالوئی
- **شمارهٔ قرارداد:** ۳۳/۱۲۱۱۶
- **تاریخ قرارداد:** ۱۳۹۸/۱۰/۲۸
- **تاریخ شروع:** ۱۳۹۸/۱۰/۲۸
- **تاریخ پایان:** ۱۳۹۹/۱۲/۱۱
- **ناظران:** دکتر محمود بی‌جن‌خان، فاطمه جعفری زیارانی

در پژوهشگاه علوم و فناوری اطلاعات ایران گواهی می‌شود. گزارش پایانی این طرح، بر پایهٔ داوری در نشست ۴۰۶ (تاریخ ۱۴۰۰/۱/۱۸) شورای پژوهشی پژوهشگاه به تصویب رسیده است. لازم می‌دانم مراتب قدردانی خود را به مجری، همکاران، و ناظران این طرح تقدیم دارم.

با آرزوی بهروزی
پرویز شهریاری
معاون پژوهش و آموزش

تهران . خیابان انقلاب اسلامی . چهارراه فلسطین . شمارهٔ ۱۰۹۰ صندوق پستی ۱۳۷۱-۱۳۱۸۵
تلفن: ۶۶۴۹۴۹۸۰ - ۶۶۹۵۱۴۳۰ دورنگار: ۶۶۴۶۲۲۵۴ تلفن گویا: ۶۶۴۹۴۹۵۴
irandoc.ac.ir
info@irandoc.ac.ir

ساخت پیکرهٔ متنی از مقاله‌های پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

مدیر طرح پژوهشی: الهام علایی ابوذر، عضو هیات علمی پژوهشکده علوم اطلاعات پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)

نشانی: تهران، خیابان انقلاب، تقاطع فلسطین، شماره ۱۰۹۰

تلفن: ۶۶۹۵۱۴۳۰ (۰۲۱) **دورنگار**: ۶۶۴۶۲۲۵۴ (۰۲۱)

رایانامه: alayi@irandoc.a.ir

چکیده

انجام بسیاری از پژوهش‌های زبان‌شناسی و تصمیم‌گیری‌ها در برنامه‌ریزهای زبانی، تنها با استفاده از یک پیکرهٔ زبانی امکان‌پذیر است. در پژوهش حاضر پیکره‌ای ساخته شده است که شامل متون مقاله-های موجود در پژوهش‌نامهٔ پردازش و مدیریت اطلاعات است. ۶۷۷ مقاله و ۴۷۷۴۷۹۰ واژه. موضوع این مقاله‌ها شامل کتابداری و اطلاع‌رسانی، علم اطلاعات و دانش‌شناسی، فناوری اطلاعات، مدیریت اطلاعات، مدیریت دانش، زبان‌شناسی، زبان‌شناسی رایانشی، اصطلاح‌شناسی، هستان‌شناسی و سایر حوزه-های مرتبط با پردازش اطلاعات است. بنابراین، محتوای این پیکره عمومی نیست و حاوی متون بسیار تخصصی و میان‌رشته‌ای است و از این نظر برای پردازش‌هایی که مستلزم بهره‌گرفتن از متون تخصصی است، بسیار ارزشمند است. پس از نمونه‌گیری و وارد کردن متون مقاله‌ها در پیکره، فرادادهٔ مربوط به مقاله‌ها (مانند عنوان مقاله، نام نویسنده/نویسندگان و موضوع مقاله) نیز از طریق کاربرگه ثبت مقالات وارد پیکره شد. سپس، ابتدا نرمال‌سازی ماشینی و به دنبال آن، برچسب‌گذاری ماشینی (نوعاً برچسب‌گذاری اجزای واژگانی کلام (POS)) انجام شد و اطلاعات در پایگاه داده ذخیره شده است و در نهایت، برچسب‌ها دستی کنترل شده‌اند. همچنین در این طرح پژوهشی، در کنار ساخت پیکره، سامانه‌ای طراحی شد به نام «سامانهٔ پیکره‌های ایرانداک» که پیکرهٔ متنی ایرانداک در این سامانه قرار گرفت و در آینده می-توان پیکره‌های دیگری را نیز در این سامانه قرار داد.

واژه‌های کلیدی: پیکره، نرمال‌سازی، برچسب‌گذاری اجزای واژگانی کلام، سامانهٔ پیکره‌های ایرانداک

فهرست مطالب

عنوان شماره صفحه

فهرست جدول‌ها

فهرست شکل‌ها

سپاسگزاری

به این وسیله از ناظر محترم این طرح پژوهشی، آقای دکتر محمود بی‌جن‌خان که از تجربه و نظرات ارزشمند ایشان در ساخت پیکره بسیار بهرمند گشتم، و از رئیس محترم پژوهشگاه، آقای دکتر سیروس علیدوستی برای حمایت‌های مادی و معنوی از این طرح پژوهشی، کمال تشکر و قدردانی را می‌نمایم. از دوستان و همکاران گرامی، خانم دکتر لیلا نامداریان، آقای دکتر بهروز رسولی، آقای دکتر محمد ربیعی، خانم ستاره مدرسی و آقای علیرضا کبودان نیز که در تهیه داده‌ها و تسهیل فرآیند ساخت پیکره بنده را همراهی نمودند بسیار سپاسگزارم. همچنین از همراهی صمیمانهٔ اعضای محترم شورای پژوهش کمال تشکر و قدردانی را می‌نمایم.

فصل اول

مقدمه

۱-۱- مقدمه

پیکره را می‌توان اینگونه تعریف کرد: پیکره[1]، مجموعه‌ای نظام‌مند، رایانه‌ای شده، معتبر و موثق از زبان است که برای مطالعات زبان‌شناسی مورد استفاده قرار می‌گیرد (سینکلر:۲۰۰۴) و زبان‌شناسی پیکره‌ای[2]، اصول و شیوه‌های به کارگیری پیکره در پژوهش‌های زبان‌شناسی است. در حقیقت، زبان‌شناسی پیکره‌ای، بررسی زبان مبتنی بر داده‌هایی از کاربرد واقعی زبان است. داده‌های زبانی موجود در پیکره هم می‌تواند به صورت متن باشد و هم گفتار ضبط شده و آوانویسی شده. اصولاً هر مجموعه‌ای که متشکل از بیش از یک متن باشد را می‌توان پیکرهٔ متنی[3] خواند، اما واژهٔ تخصصی «پیکره» یا عبارت «پیکرهٔ متنی» در زبان-شناسی مدرن به کار می‌رود و به مجموعه‌ای از متون اطلاق می‌شود که برای تجزیه و تحلیل زبان‌شناسی یا پردازش زبان طبیعی مورد استفاده قرار می‌گیرد.

بسیاری از پژوهش‌های زبان‌شناسی و تصمیم‌گیری‌ها در برنامه‌ریزی زبانی، تنها با استفاده از یک پیکرهٔ زبانی امکان‌پذیر است. برخی از کاربردهای پیکره شامل: پردازش زبان طبیعی و درک و بازشناسی گفتار، تبدیل متن به گفتار و برعکس، تدوین فرهنگ‌ها، آموزش و پژوهش، ساخت پایگاه‌های دادهٔ زبانی، بررسی واژه‌های هم‌آیند در زبان‌های گوناگون، پایشگری زبان برای پیگیری و ردگیری دگرگونی‌های زبانی، ترجمهٔ ماشینی، توسعهٔ مفاهیم و منابع در پیوند با واژگان، نگارش و گسترش مهارت‌های نوشتاری، آموزش و یادگیری زبان با شناخت گویش‌ها و گوناگونی زبانی، معناشناسی، تحلیل کلام، زبان‌شناسی اجتماعی، زبان‌شناسی حقوقی، واکاوی ژانرهای ادبی و پژوهش‌های دستوری است. در استفاده از پیکره در مطالعات زبان‌شناسی سه رویکرد رایج وجود دارد: رویکرد پیکره‌بنیاد[4]: در این رویکرد نظریه مطرح می‌گردد و برای سنجش نظریه از پیکره استفاده می‌شود. رویکرد برگرفته از پیکره[5]:

---

[1] corpus
[2] corpus linguistics
[3] text corpus
[4] corpus-based approach
[5] corpus-driven approach

در این رویکرد از پیکره برای بازیابی اطلاعات استفاده می‌شود. رویکرد پیکره‌یار[1]: این رویکرد ترکیبی است از به کارگیری روش‌های کمی و کیفی.

پیکره انواع گوناگونی دارد؛ اتکینز و همکاران (۱۹۹۲) انواع پیکره را از چند دیدگاه مختلف بررسی کرده‌اند و بر آن اساس پیکره‌ها را به انواع «متن کامل[2]»، «نمونه‌ای[3]»، «نظارتی[4]»، «بسته[5]»، «باز[6]»، «هم-زمانی[7]»، «درزمانی[8]»، «عمومی[9]»، «اصطلاح‌شناسی[10]»، «یک‌زبانه[11]»، «دوزبانه[12]»، «چندزبانه[13]»، «منفرد[14]»، «موازی[15]»، «مرکزی[16]»، «پوسته‌ای[17]»، «هسته‌ای[18]» و «پیرامونی[19]» طبقه‌بندی کرده‌اند.

۱-۲- ضرورت انجام پژوهش

تهیهٔ یک پیکرهٔ زبانی مبتنی بر داده‌های علمی ایرانداک (مقاله‌های پژوهش‌نامهٔ پردازش و مدیریت اطلاعات) از دو نظر حائز اهمیت است: ۱- ایرانداک را در شمار پژوهشگاه‌ها و دانشگاه‌هایی قرار می‌دهد که پیکرهٔ زبانی مفید (مبتنی بر داده‌های علمی معتبر) تهیه کرده‌اند ۲- می‌توان بسیاری از پژوهش‌های حوزهٔ پردازش متن (مانند انواع طبقه‌بندی داده‌ها، استخراج اطلاعات گوناگون از متن و غیره) با استفاده از این پیکره انجام داد. ساخت پیکره‌ها کاری پیچیده، زمان‌بر، دارای گام‌های گوناگون، و میان‌رشته است. بنابراین برای کارایی و اثربخشی در ساخت آن‌ها، نیاز است که پیش از آغاز کار، امکان‌سنجی[20] انجام شود. امکان‌سنجی؛ گام‌ها، هزینه‌ها، نیروی انسانی، و مانند آن‌ها را برای یک پروژه،

---

[1] corpus-assisted approach
[2] whole text
[3] samples
[4] monitor
[5] open
[6] closed
[7] synchronic
[8] diachronic
[9] general
[10] thesaurus
[11] monolingual
[12] bilingual
[13] multilingual
[14] single
[15] parallel
[16] central
[17] cluster
[18] nuclear
[19] Perimeter
[20] feasibility study

بررسی و به مدیریت پروژه کمک می‌کند تا آن را با آمادگی و آینده‌نگری بیشتری به سرانجام برساند. خبرگانی که به تجزیه و تحلیل پیکرهٔ زبانی می‌پردازند لزوماً در ساخت پیکره/پیکره‌هایی که از آن‌ها استفاده می‌کنند، تبحر ندارند؛ در حقیقت همیشه این خطر متوجه آنان است که ممکن است پیکره‌ای بسازند که حاوی اطلاعاتی باشد که یا از قبل می‌دانند، و بنابراین اطلاعات جدیدی نیست، و یا می‌توانند جزئیات زبان‌شناسی آن را حدس بزنند. حال آنکه شرایط مطلوب این است که پیکره باید طوری طراحی و ساخته شود که الگوهای ارتباطی جوامعی که آن زبان در آن استفاده می‌شود در پیکره منعکس شود. صرفنظر از محتوای اسناد مکتوب و گفتاری پیکره، اسناد باید حاصل جمع‌آوری مکالمات مردم و مطالبی باشد که مردم می‌نویسند یا می‌خوانند. بنابراین، محتوای پیکره باید بدون در نظر گرفتن زبان پیکره و بر اساس کارکرد کاربردی زبان در جامعهٔ مورد استفاده باشد. مهم‌ترین شاخص‌های اصلی ساخت پیکرهٔ زبانی شامل: نمونه‌گیری[۱]، نمایندگی[۲]/ نماینده بودن حجم داده‌های زبانی، توازن[۳]، اندازه[۴]، نوع پیکره: عمومی[۵] یا تخصصی[۶] و یک‌دستی[۷] است (سینکلر: ۲۰۰۴). این شاخص‌ها کلی است و در ساخت هر نوع پیکره‌ای، چه نوشتاری/متنی و چه گفتاری/آوایی لحاظ می‌شوند. در طرح پژوهشی پیشین نگارنده، با عنوان «امکان‌سنجی ساخت پیکره از داده‌های علمی پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)»، ابعاد و مولفه‌های امکان‌سنجی ساخت پیکره (نوعاً، پیکرهٔ متنی) استخراج شد و مدلی کلی برای ساخت پیکره پیشنهاد و اعتباریابی شد. سپس مدل امکان‌سنجی ساخت پیکره در پنلی متشکل از متخصصان موضوعی و اجرایی در ایرانداک به شور گذاشته و اعتباریابی شد و مدل پایانی برای ساخت پیکره از داده‌های علمی ایرانداک به دست آمد و در پایان، برنامهٔ ساخت پیکره از داده‌های علمی ایرانداک مشخص شد که در آن چگونگی پرداختن به ابعاد و مولفه‌های مدل مشخص شده است. در امکان‌سنجی ساخت پیکرهٔ متنی باید به هفت بعد و ۳۳ مؤلفه پرداخت (جدول ۱-۱). این ابعاد؛ فنی، اقتصادی، زمان‌بندی، قانونی‌حقوقی، عملیاتی، تأمین مالی و بازاریابی هستند. هر بعد نیز مؤلفه‌هایی را در بر دارد. در بعد فنی، نیاز به نیروی متخصص در تیم ساخت پیکره مانند زبان‌شناس همگانی/ عمومی، زبان-شناس رایانشی، برنامه‌نویس رایانه‌ای، پژوهشگرانی که تجربهٔ ساخت پیکره دارند، بررسی می‌شود. بررسی

[1] sampling
[2] representativeness
[3] balance
[4] size
[5] general corpus
[6] specialized corpus
[7] homogenity

نیاز به پشتیبانی برون‌سازمانی از پیکره و رابط کاربر (User interface) برای کار کردن کاربران با پیکره از دیگر ویژگی‌های فنی ساخت پیکره است. در بعد فنی گام‌هایی هم برای ساخت پیکره وجود دارد که شامل انتخاب نوع متون، تعیین حجم متون، پیش‌پردازش متون، و حاشیه‌نویسی (برچسب‌گذاری و...) است. در بعد اقتصادی باید برای مشاوره گرفتن و به کارگیری نیروهای متخصص، برآورد هزینه انجام شود. دربارهٔ برنامه‌نویسی رایانه، حاشیه‌نویسی دستی، آموزش کاربران، خرید نرم‌افزار، خرید وسایل و تجهیزات اداری (رایانه، میز، صندلی و....)، تأمین زیرساخت (سرور، شبکه، و...)، نگهداری زیرساخت (سرور، شبکه، و...)، ساخت ابزار گردآوری متن، ساخت ابزار کار با داده‌های پیکره، و به‌روز کردن داده-های پیکره در پیکره‌های پایشگر/ به‌روزشونده نیز باید برآورد هزینه شود. مؤلفهٔ دیگری که در بعد اقتصادی دیده می‌شود، تصمیم‌گیری دربارهٔ چگونگی دسترسی کاربران به پیکره (رایگان یا با پرداخت هزینه) است. در بعد زمان‌بندی، باید زمان‌بندی پروژهٔ ساخت پیکره و گام‌های آن روشن شود. مؤلفه‌هایی که در بعد حقوقی‌قانونی بررسی می‌شوند؛ الزام‌های قانونی و حقوقی و اجازه‌نامهٔ کاربرد متون و حقوق مادی پیکره (پس از ساخت) هستند تا ساخت پیکره از دیدگاه حقوقی‌قانونی با مانعی برخورد نکند. تأمین مالی بعد دیگری از امکان‌سنجی ساخت پیکره است که در آن باید منابع تأمین مالی برای ساخت پیکره پیش‌بینی شود. در بعد بازاریابی نیز شناخت کاربران پیکره و آزمون عملکرد پیکره، پیش از عرضهٔ آن به بازار باید بررسی شوند (علایی و علیدوستی: ۱۳۹۸).

جدول ۱-۱. ابعاد و مؤلفه‌های امکان‌سنجی ساخت پیکرهٔ متنی (برگرفته از علایی و علیدوستی: ۱۳۹۸)

ابعاد	مؤلفه‌ها
۱. فنی	۱-۱. بررسی نیاز به زبان‌شناس همگانی/ عمومی در تیم ساخت پیکره
	۱-۲. بررسی نیاز به زبان‌شناس رایانشی در تیم ساخت پیکره
	۱-۳. بررسی نیاز به برنامه‌نویس رایانه‌ای در تیم ساخت پیکره
	۱-۴. بررسی نیاز به پشتیبانی برون‌سازمانی از پیکره
	۱-۵. طراحی زیربنای پیکره (چارچوب و ساختار)، به عنوان گامی از ساخت پیکره
	۱-۶. انتخاب نوع متون، به عنوان گامی از ساخت پیکره

	۱-۷.	تعیین حجم متون، به عنوان گامی از ساخت پیکره
	۱-۸.	پیش‌پردازش متون، به عنوان گامی از ساخت پیکره
	۱-۹.	حاشیه‌نویسی (برچسب‌گذاری و...)، به عنوان گامی از ساخت پیکره
	۱-۱۰.	بررسی نیاز به رابط کاربر[1] برای کار کردن کاربران با پیکره
	۱-۱۱.	بررسی نیاز به همکاری پژوهشگرانی که تجربهٔ ساخت پیکره دارند
۲. اقتصادی	۲-۱.	برآورد هزینهٔ متخصصان ساخت پیکره
	۲-۲.	برآورد هزینهٔ برنامه‌نویسی رایانه
	۲-۳.	برآورد هزینهٔ حاشیه‌نویسی دستی
	۲-۴.	برآورد هزینهٔ آموزش کاربران
	۲-۵.	برآورد هزینهٔ خرید نرم‌افزار
	۲-۶.	برآورد هزینهٔ وسایل و تجهیزات اداری (رایانه، میز، صندلی و....)
	۲-۷.	برآورد هزینهٔ اولیهٔ تأمین زیرساخت (سرور، شبکه، و...)
	۲-۸.	برآورد هزینهٔ نگهداری زیرساخت (سرور، شبکه، و...)
	۲-۹.	برآورد هزینهٔ ساخت ابزار گردآوری متن
	۲-۱۰.	برآورد هزینهٔ ساخت ابزارِ کار با داده‌های پیکره
	۲-۱۱.	تصمیم‌گیری دربارهٔ چگونگی دسترسی کاربران به پیکره (رایگان یا با پرداخت هزینه)
	۲-۱۲.	برآورد هزینهٔ به‌روز کردن داده‌های پیکره در پیکره‌های پایشگر/ به‌روزشونده
۳. زمان‌بندی	۳-۱.	برآورد زمان‌بندی ساخت پیکره
	۳-۲.	فازبندی زمان ساخت پیکره
۴. قانونی‌حقوقی	۴-۱.	بررسی الزام‌های قانونی و حقوقی در ساخت پیکره
	۴-۲.	بررسی اجازه‌نامهٔ کاربرد متون در ساخت پیکره

[1] User Interface:UI

	۴-۳.	بررسی حقوق مادی پیکره (پس از ساخت)
۵. عملیاتی	۵-۱.	بررسی پذیرش و آمادگی سازمان در کاربرد پیکره
	۵-۲.	بررسی دستاوردهای سازمان با ساخت پیکره
۶. تأمین مالی	۶-۱.	پیش‌بینی منابع تأمین مالی برای ساخت پیکره
۷. بازاریابی	۷-۱.	شناخت کاربران پیکره
	۷-۲.	آزمون عملکرد پیکره پیش از عرضه به بازار

اکنون که ابعاد و مولفه‌های ساخت پیکرهٔ متنی از داده‌های علمی ایرانداک مشخص شده است، نگارنده قصد دارد در این طرح پژوهشی اقدام به ساخت پیکرهٔ متنی نماید و با در نظر گرفتن ملاحظات حقوقی، قصد دارد این پیکره را از متون مقاله‌های موجود در «پژوهش‌نامهٔ پردازش و مدیریت اطلاعات» بسازد. از آنجائیکه این پیکره از مقاله‌های مجلهٔ پردازش و مدیریت اطلاعات ساخته می‌شود، محتوای این پیکره عمومی نیست و حاوی متون بسیار تخصصی و میان‌رشته‌ای است و از این نظر برای پردازش‌هایی که مستلزم بهره گرفتن از متون تخصصی است، بسیار ارزشمند خواهد بود.

۱-۳- سوال پژوهش

این طرح پژوهشی سعی دارد حین ساخت پیکره به سوال‌های زیر نیز توجه داشته باشد:

- چه حجمی از مقاله‌های پژوهش‌نامهٔ پردازش و مدیریت اطلاعات را می‌توان در پیکره وارد کرد؟
- از چه ابزارهایی می‌توان برای پیش‌پردازش متون فارسی در پیکره استفاده کرد؟
- با توجه به نیازهای ایرانداک در زمینه‌های پژوهشی مربوط به پردازش متن، و با توجه به تخصصی بودن پیکره، چه نوع حاشیه‌نویسی‌های دیگری را می‌توان در طرح‌های پژوهشی آینده به پیکره اضافه کرد؟ در حقیقت چه نوع اطلاعات زبانی توضیحی-تفسیری مانند اطلاعات تخصصی نحوی، گفتمانی و.... را می‌توان در طرح‌های پژوهشی مجزا به این پیکره اضافه کرد؟

در این طرح پژوهشی ابتدا در فصل دوم، به پیشینهٔ پژوهش، که در واقع مرور مطالعات انجام شده در حوزهٔ ساخت پیکرهٔ زبانی است، پرداخته می‌شود و در پایان برخی از پیکره‌های زبان فارسی معرفی می‌شوند. در فصل سوم به روش ساخت پیکره و واردکردن داده در پیکره پرداخته می‌شود. فصل چهارم، چگونگی نرمال‌سازی و حاشیه‌نویسی پیکره و ابزارهای مربوطه توضیح داده می‌شود. فصل پنجم استقرار سامانه پیکره ایرانداک است و در نهایت فصل ششم، نتیجه‌گیری و ارائهٔ پیشنهادات برای پژوهش‌های آینده است.

فصل دوم

پیشینهٔ پژوهش

۲-۱- مقدمه

در این فصل پژوهش‌های انجام شده در حوزهٔ ساخت پیکرهٔ زبانی، مورد بررسی قرار خواهند گرفت و در ادامه به معرفی برخی از پیکره‌های زبان فارسی پرداخته خواهد شد.

۲-۲- مرور پژوهش‌های پیشین ساخت پیکرهٔ زبانی

پیکره‌های زبانی و واکاوی پیکره‌بنیاد زبان در زبان‌شناسی پیکره‌ای نقشی کلیدی دارند. ساخت پیکره‌ها کاری پیچیده، زمان‌بر، دارای گام‌های گوناگون، و میان‌رشته است. بنابراین برای کارایی و اثربخشی در ساخت آن‌ها نیاز است که پیش از آغاز کار، امکان‌سنجی انجام شود. امکان‌سنجی؛ گام‌ها، هزینه‌ها، نیروی انسانی، حقوق مادی و معنوی، و مانند آن‌ها را برای یک پروژه، بررسی و به مدیریت پروژه کمک می‌کند تا آن را با آمادگی و آینده‌نگری بیشتری به سرانجام برساند. در این راستا و به منظور ارائهٔ مدلی برای امکان‌سنجی ساخت پیکره، علایی و علیدوستی (۱۳۹۸) پس از بررسی پژوهش‌های پیشین، شاخص‌های کلی ساخت پیکره‌های زبانی را استخراج کرده‌اند. سپس، فرآیند ساخت پیکره را به تفصیل مورد بررسی قرار داده‌اند. پس از بررسی پژوهش‌های پیشین در زمینهٔ امکان‌سنجی، بر پایهٔ اطلاعات به دست آمده از مطالعات امکان‌سنجی، بررسی شاخص‌های ساخت پیکره و فرآیند ساخت پیکره، ابعاد و مولفه‌های امکان‌سنجی ساخت پیکره استخراج شده است و مدلی کلی برای ساخت پیکره پیشنهاد شده است. سپس با روش دلفی و در دو دور اعتباریابی، مدل پایانی به دست آمده است. این مدل دارای هفت بعد فنی، اقتصادی، زمان‌بندی، قانونی‌حقوقی، عملیاتی، تأمین مالی، و بازاریابی و روی‌هم‌رفته ۳۳ مؤلفه است. این مدل و کاربرد آن در پروژه‌های ساخت پیکرهٔ متنی می‌تواند کمک بزرگی برای

پژوهشگران و سازمان‌ها باشد. واینه (۲۰۰۵) نیز به مرور نظرات نویسندگان و پژوهشگران نامی حوزهٔ ساخت پیکره می‌پردازد. این پژوهشگران پیشنهادات کاربردی برای ساخت پیکره (از نوع داده‌های مورد استفاده از پیکره تا انواع و چگونگی حاشیه‌نویسی) ارائه می‌کنند. در حقیقت، در این کتاب، به زبان ساده مراحل ساخت پیکره توضیح داده شده است. بی‌جن‌خان و همکاران (۲۰۱۱) نیز به بررسی موضوعاتی می‌پردازند که هنگام ساخت پیکره باید مدنظر قرار گیرند. آن‌ها پیکره‌ای را معرفی می‌کنند که حاوی ۳۵/۰۵۸ فایل متنی است. اندازهٔ هر متن حداقل ۱۰۰۰ واژه است. در این پیکره، برای حاشیه‌نویسی از دستورالعمل EAGLES پیروی شده است که نتیجهٔ به کار گیری این دستورالعمل، تهیهٔ سلسله‌مراتبی از برچسب‌ها بوده است. برای حاشیه‌نویسی از روش نیمه‌خودکار استفاده شده است. همچنین در ساخت این پیکره، به هم‌نگاره‌ها و ساخت اضافه نیز توجه داشته‌اند.

انواع پیکره‌ها می‌تواند شامل پیکرهٔ آوایی، تخصصی، موازی، متنی و غیره باشد. به عنوان نمونه، در فصلی از کتاب «آواشناسی پیکره‌ای» دوراند و همکاران (۲۰۱۴) فرآیند ساخت پیکره‌های آوایی را توضیح می‌دهند. آن‌ها پس از تعریف پیکرهٔ آوایی، به بررسی اجزای مهم ساخت چنین پیکره‌هایی می‌پردازند؛ این بخش‌ها شامل: ذخیرهٔ پیکره، چگونگی به اشتراک‌گذاشتن و استفاده مجدد از پیکره، نمایندگی و اندازهٔ پیکره، انتخاب دادهٔ خام و حاشیه‌نویسی پیکره است. کلود توریدا (۲۰۱۶) به توضیح گام‌به‌گام ساخت پیکرهٔ تخصصی[1] و تهیهٔ فهرست واژگان حاشیه‌نویسی شده و مبتنی بر فراوانی واژگان در پیکره می‌پردازد. وی از پروژهٔ «آموزش زبان انگلیسی برای اهداف دانشگاهی» که در دانشگاهی در خاورمیانه توسعه یافته است، مثال‌هایی ارائه می‌کند. مراحل ساخت پیکرهٔ تخصصی از نظر کلود توریدا شامل انتخاب متون آموزشی، حذف واژگانی که قاموسی[2] نیستند (مانند حروف اضافه، حروف ربط و....)، تجزیه و تحلیل متن با استفاده از نرم‌افزار AntConc، ایجاد فهرست فراوانی واژه‌ها، توسعهٔ فهرست واژگان حاشیه‌نویسی شده که خود شامل: تعیین مقولهٔ نحوی (POS) واژگان موجود در فهرست، اضافه کردن تعریف واژگان، باهم‌آیی واژگان و نمونه‌ای از جمله‌ای که واژه در آن به کار رفته، است. تابع بردبار (۱۳۹۳) نحوهٔ ساخت پیکرهٔ زبانی موازی[3] به وسیلهٔ پیکره‌های قیاس‌پذیر را توضیح می‌دهد. برای ساخت این پیکرهٔ زبانی، از پایگاه دانش ویکی‌پدیا و شیوه‌ای مبتنی بر بازیابی اطلاعات بر اساس برچسب گروه و لینک، به منظور دسته‌بندی

[1] specialized corpus
[2] Content words
[3] پیکره‌های موازی از متونی تشکیل شده‌اند که با ترجمه‌هایی از یک یا بیش از یک زبان دیگر در ارتباط قرار گرفته‌اند.

مقالات مشابه استفاده شده است. سپس، از برخی ویژگی‌ها که دربرگیرندهٔ شباهت ترجمه و میزان شباهت از نظر ترازبندی است، برای امتیازدهی جمله‌ها استفاده شده است و به منظور دادن وزن بهینه به هر یک از این ویژگی‌ها، از یک مدل خطی استفاده شده است. محمدی (۱۳۹۱) چگونگی ساخت پیکرهٔ تطبیقی فارسی- انگلیسی را توضیح می‌دهد. منابع استفاده شده در این پیکره شامل اسناد خبری روزنامه‌های همشهری و بی‌بی‌سی است و از اسناد به دست آمده، معیارهایی مانند تعداد کلمات کلیدی مشترک، اسم-های خاص یکسان، عنوان‌های مشابه و فاصلهٔ تاریخ انتشار دو خبر استخراج شده است. سپس معیارهای به دست آمده بر اساس میزان اهمیت آن‌ها در ترازبندی متون، با وزن‌های مختلف با یکدیگر ترکیب شده‌اند و در نهایت، به استخراج جمله‌های موازی از پیکرهٔ تطبیقی ساخته شده پرداخته شده است. آیت (۱۳۸۹) نیز مراحل تهیه یک دادگان دایفون ویژه زبان فارسی را این گونه بیان می‌کند: ابتدا پایگاه واژگانی که دایفون‌های زبان را شامل شوند، تهیه شده است. سپس نرم‌افزاری طراحی و پیاده‌سازی شده است که با گرفتن صورت‌های واجی واژه‌ها، دایفون‌هایی را که قرار است از آن استخراج شوند، مشخص می‌کند. در مرحله بعد سیگنال‌های گفتاری واژه‌ها ضبط و بررسی شده است. در پایان نیز جداسازی دایفون‌ها و تهیه دادگان مورد نظر صورت پذیرفته است. برای افزایش دقت دادگان تهیه شده، مراحل جداسازی دایفون‌ها از سیگنال‌های گفتاری ضبط شده با استفاده از سه روش شنوایی، بررسی سیگنال زمانی و مطالعه طیف‌نگاشت، ارزیابی و از ترکیب هر سه روش برای افزایش دقت دادگان استفاده شده است. از جمله تلاش‌هایی که در زمینهٔ ساخت پیکره‌های دوزبانه فارسی- انگلیسی انجام گرفته است، می‌توان به پیکرهٔ ساخته شده توسط دشتبانی (۱۳۹۱) اشاره کرد. وی اقدام به طراحی و ساخت پیکرهٔ دو زبانهٔ فارسی- انگلیسی حوزهٔ فاوا (حوزهٔ فناوری ارتباطات و اطلاعات) کرده است. این پیکره به صورت خودکار ساخته شده است و منابع آن، اسناد تخصصی حوزهٔ فاوا است. در این پژوهش، نرم‌افزاری برای ساخت پیکره طراحی شده است که هزینه و مدت زمان ساخت پیکره را کاهش می‌دهد. علاوه بر این، نرم‌افزار ارائه شده قابلیت مدیریت پیکره را برای کاربران فراهم می‌کند. از ویژگی‌های پیکرهٔ ساخته شده، فراهم کردن مجموعه‌ای متمرکز از اسناد تخصصی است که می‌تواند در پروژه‌های مختلف حوزهٔ فاوا استفاده شود.

مهم‌ترین مرحلهٔ ساخت پیکره‌های چندزبانی، ترازبندی داده‌های پیکره است. در این پروژه روشی برای ترازبندی جمله‌های پیکرهٔ فارسی تخصصی حوزه فاوا و جملات انگلیسی پیکرهٔ تخصصی حوزه فاوا ارائه شده است. هدف این پژوهش، طراحی یک سیستم ترازبندی برای استخراج جمله‌های متناظر دو زبان است. در این روش با استفاده از یک لغت‌نامه دوزبانی و تکنیک‌های هوش مصنوعی، امتیاز نشان دهندهٔ

شباهت دو جمله، محاسبه می‌شود. در نهایت اطلاعات مربوط به نگاشت جمله‌های دو مجموعهٔ انگلیسی و فارسی در پایگاه دادهٔ پیکره، ذخیره می‌گردد.

تهیهٔ پیکرهٔ متنی برای زبان فارسی با مشکلاتی همراه است؛ برخی از این مشکلات توسط قیومی و همکاران (۲۰۱۳) مورد بررسی قرار گرفته است. آن‌ها معتقدند منبع مشکلات مربوط به عوامل متعددی می‌شود، از جمله: نظام نوشتاری فارسی، تلفیق نظام نوشتاری فارسی با خط عربی، شیوهٔ تایپ‌کردن تایپیست‌ها و خلاقیت زبانی. این مشکلات به صورت خودکار و از طریق برنامه‌نویسی یا به صورت دستی حل شده است. به عنوان مثال، دربارهٔ کدهای مربوط به حروف، از همان کدهای عربی استفاده شده است و برخی از کدها که در عربی موجود نیست (مانند کد مربوط به حرف «گ» یا «ژ») به کدهای قبلی اضافه شده است. فاصله‌گذاری درست بین ریشه و وندها که معمولاً تایپیست‌ها رعایت نمی‌کنند، خود باعث مشکل در پردازش متن می‌شود ؛ به عنوان مثال کلمات «کتاب‌ها«، «کتاب ها» و «کتابها» یک کلمه هستند، اما اگر بین «کتاب» و «ها» یک فاصله (و نه نیم‌فاصله) قرار گیرد، در پردازش، دو کلمه در نظر گرفته می‌شوند. این مشکل به این صورت حل شده است که شکل‌های مختلف یک کلمه که در نتیجهٔ اتصال وند به ریشه/ستاک ایجاد شده‌اند، یک شکل در نظر گرفته شوند و همگی با نیم‌فاصله فرض شوند؛ این عمل نیز به صورت خودکار و از طریق برنامه‌نویسی انجام شده است. اما مشکل مربوط به اجزای کلمات مرکب که معمولاً با ساخت اضافه به هم متصل می‌شوند، را نمی‌توان از طریق مذکور حل کرد. زیرا وندها محدود هستند و در برنامه‌نویسی می‌توان وندهای تصریفی و اشتقاقی را لحاظ کرد تا هر جا به ستاک/ ریشه متصل شده باشند، نیم‌فاصله در نظر گرفته شود، اما در مورد کلمات مرکب چون نمی‌توان تعداد ثابتی برای اجزای آن‌ها در نظر گرفت، نمی‌توان چنین برنامه‌نویسی‌ای انجام داد. در این حالت باید شکل‌های مختلف کلمه توسط تحلیلگر ساخت‌واژی تشخیص داده شوند.

۲-۳- برخی از پیکره‌های زبانی فارسی

۲-۳-۱- پیکرهٔ متنی زبان فارسی: این پیکره، توسعه‌یافتهٔ «پیکرهٔ بی‌جن‌خان» است. پیکرهٔ بی‌جن‌خان، مجموعه‌ای است از متون فارسی شامل بیش از دو میلیون واژه که با ۵۵۰ نوع برچسب اجزای واژگانی کلام[1]، برچسب‌گذاری شده‌اند. این پیکره شامل بیش از ۴۳۰۰ برچسب موضوعی چون سیاسی، تاریخی و ... برای متون است و در پژوهشکدهٔ پردازش هوشمند علائم تهیه شده است (بی‌جن‌خان: ۱۳۸۳). پیکرهٔ متنی استاندارد زبان فارسی مجموعه‌ای است از متون نوشتاری و گفتاری زبان فارسی رسمی است و از منابعی مانند روزنامه‌ها، سایت‌ها و مستندات از قبل تایپ شده، جمع‌آوری شده، تصحیح گردیده و برچسب خورده است. حجم این دادگان حدوداً ۱۰۰ میلیون کلمه است و از منابع مختلف تهیه گردیده و دارای تنوعات بسیار زیادی است. پیکرهٔ متنی زبان فارسی دارای قابلیت‌ها و ویژگی‌های زیر است (بی‌جن-خان و همکاران :۲۰۱۱):

- جمع‌آوری و سازمان‌دهی متون نوشتاری و گفتاری رسمی زبان فارسی با حجم ۱۰۰ میلیون کلمه
- ویرایش نیمه‌خودکار اولیه متون
- برچسب‌دهی نحوی-معنایی کلمات برای ۱۰ میلیون کلمه با استفاده از ۸۸۲ برچسب به صورت دستی توسط دانشجویان رشته زبان‌شناسی بر اساس دستورالعمل
- تهیه نویسه‌های UniCode و XML برای پرونده‌های متنی دادگان
- امکان برچسب‌دهی گروه‌های نحوی
- طبقه‌بندی هر پرونده بر حسب موضوع و منبع آن
- پوشش موضوعات مختلف سیاسی، اجتماعی، اقتصادی، فرهنگی، ...
- مجهز به یک نرم‌افزار آماری برای محاسبه و استخراج ویژگی‌های زبانی از قبیل: توزیع احتمالی مشروط، واژگان بسامدی، شناسایی هم‌نگاره‌ها، هم‌آیندها، مطابقه‌ها و ترتیب قاموسی با امکان گزارش‌گیری
- طراحی یک زبان جستجوی هوشمند

[1] Part Of Speech (POS)

۲-۳-۲-پایگاه دادگان زبان فارسی: این پیکره، مجموعه‌ای است برای ذخیره، پردازش و ارائه داده‌های زبانی فارسی. این پایگاه دربرگیرنده پیکره‌های گوناگونی زبان فارسی است؛ پیکره‌های این پایگاه می‌توانند همواره روزآیند شوند. پایگاه دادگان زبان فارسی فراگیر و متنوع است. در واقع فراتر از یک یا چند پیکرهٔ خاص است و کاربران بر پایهٔ نیاز و هدف پژوهشی خود می‌توانند پیکرهٔ مناسب را از آن برگزینند. حتی پژوهندگان می‌توانند پیکره‌های اختصاصی خود را وارد پایگاه کنند و تحلیل‌ها و فهرست‌گیری‌های مورد نظر خود را انجام دهند. پایگاه دادگان زبان فارسی دارای متن‌های نشانه‌گذاری شده از جمله شناسنامهٔ متن، برچسب‌های دستوری، آوایی، ریشه‌ای و معنایی است. این دادگان مجهز به نرم‌افزارهای اختصاصی جستجو، تقطیع و تحلیل متن است که می‌تواند انواع فهرست‌های واژگانی، بسامدی و آماری را ارائه کند. منابع مورد استفاده در این پایگاه از گونه‌های نوشتاری با استفاده از متن‌های معتبر و با رعایت معیارهای مختلف نمونه‌گیری شده است و البته هیچ‌گونه محدودیت و امساکی در مورد آثار مهم ادبی و نویسندگان سرشناس و بویژه صاحب سبک و تاثیرگذار اعمال نمی‌شود. فهرست‌های مفصلی از همهٔ منابع مهم نظم و نثر فارسی معاصر فراهم شده است. این فهرست‌ها به طور جداگانه برای آثار شعری، داستانی، غیرداستانی، نمایشنامه و فیلمنامه، ادبیات کودکان، نشریه‌های ادورای و مجلات علمی، تخصصی و ادبی فراهم گردیده است. شمار آثاری که دراین فهرست‌ها قرار گرفته‌اند، بیش از یک هزار و پانصد عنوان شده است که پس از بررسی و کنار گذاشتن موارد مشابه، بیش از پانصد عنوان برای درونداد پایگاه داده‌ها برگزیده شده است. حدود ۴۵۰ اثر داستانی و غیر داستانی نثر، ۲۵۰ اثر شعری از شاعران معاصر، بیش از ۸۰ عنوان مجله و نشریهٔ علمی، ادبی و تخصصی، نزدیک به ۳۰۰ عنوان نمایش‌نامه و فیلم‌نامه، و ۲۰۰ عنوان ادبیات کودک، چندین عنوان روزنامه و نشریهٔ خبری، برخی از کتاب‌های درسی دانشگاهی و دبیرستانی، برخی از کتاب‌های دبستانی، نامه‌های اداری و بخشنامه‌ها، مجموعهٔ کامل قوانین و مقررات، نشریه‌ها و جزوه‌های پراکنده، پوسترها، دیوارنوشته‌ها و مانند این‌ها ازجمله این متون هستند (عاصی: ۱۳۸۴).

۲-۳-۳-دادگان دایفونی فارسی: واحد پایهٔ مورد استفاده در دادگان صوتی برای بازسازی رایانه‌ای گفتار باید به گونه‌ای انتخاب شود که اولاً حجم حافظهٔ معقولی را اشغال کند، یعنی تعداد واحدهای آوایی مطلوب باشد و ثانیاً بتوان گذرهای آوایی را در دادگان پوشش داد. از جمله واحدهایی که

با هدف تأمین این شرایط، تعریف و مورد استفاده قرار می‌گیرد، دایفون است. دایفون عبارت است از نیمهٔ پایدار یک آوا تا نیمهٔ پایدار آوای بعدی. دایفون از دو نیم‌آوای به‌هم‌چسبیده تشکیل می‌شود، البته باید درنظر داشت که این ترکیب شامل ترکیب سکوت و نیم‌آوا نیز می‌باشد. با این تعریف انواع دایفون‌های زبان فارسی شامل توالی‌هایcc ، cv، vc، c-، -c، -vو v- است، که «-» به معنای سکوت می‌باشد. در دادگان دایفونی فارسی که در پژوهشکده پردازش هوشمند علائم تهیه شده است، تعداد ۹۶۶ دایفون وجود دارد (اسلامی و همکاران:۱۳۸۸).

۲-۳-۴-فارس‌دات تلفنی: دادگان فارس‌دات تلفنی[1]، مجموعه‌ای از عبارات و جملات است که توسط گویندگان فارسی‌زبان از مناطق مختلف کشور از طریق خط تلفن بیان شده است. این دادگان در سطح واج (آوا) با دقت میلی‌ثانیه تقطیع و برچسب‌دهی شده و بصورت فایل‌های مجزا ذخیره گردیده است. تهیهٔ این دادگان برای کانال ارتباطی تلفن، از اهمیت ویژه‌ای برخوردار است (بی-جن‌خان و همکاران: ۲۰۰۳).

۲-۳-۵-پیکرهٔ همشهری: مجموعه همشهری پیکره‌ای است حاوی ۳۱۸ هزار سند مربوط به اخبار سال‌های ۱۳۷۵ تا ۱۳۸۶ که با خزش (Crawl) وب‌سایت همشهری و چندین مرحله پیش‌پردازش و برچسب‌گذاری حاصل آمده است. همهٔ اسناد مجموعه همشهری دارای برچسب «Cat» هستند که نشان می‌دهد هر سند در چه رده‌ای است (اقتصادی، سیاسی و...) (آل احمد و همکاران: ۲۰۰۹).

۲-۳-۶-فارس‌نت: فارس نت (وردنت عمومی زبان فارسی) پایگاه دانشی است که شامل اطلاعاتی درمورد واژه‌ها و ترکیبات زبان (مفاهیم)، اطلاعات نحوی آن‌ها و روابط معنایی میان آن‌ها است. نسخهٔ اول فارس‌نت شامل بیش از ۱۷هزار مدخل واژگانی از مقوله‌های اسم، فعل و صفت است. روابط تحت پوشش در این نسخه روابط درون‌مقوله‌ای مطرح در وردنت انگلیسی (نسخه ۲ و۱) می‌باشد و قابلیت اتصال به وردنت‌های دیگر از طریق نگاشت به وردنت پرینستون نسخهٔ ۳،۰ را نیز داراست. نسخه دوم فارس نت شامل بیش از ۳۰ هزار مدخل واژگانی از مقوله‌های اسم، فعل، صفت و قید است. علاوه بر روابط درون-مقوله‌ای مطرح در وردنت انگلیسی (نسخه ۲ و۱) ، پنج رابطه میان-مقوله‌ای نیز مفاهیم را بهم پیوند می‌دهد و علاوه بر ویژگی‌های در نظر گرفته شده برای واژه‌ها، ویژگی‌های نحوی، ساخت‌واژی و آوایی به واژه‌ها و قاب و ساختار آرگومانی به

[1] TFarsDat

افعال افزوده شده است. این وردنت نیز قابلیت اتصال به وردنت‌های دیگر را از طریق نگاشت به وردنت پرینستون نسخه ۳٫۰ داراست. مجموعه فارس‌نت توسط آزمایشگاه پردازش زبان طبیعی دانشگاه شهید بهشتی و با حمایت پژوهشگاه ارتباطات و فناوری اطلاعات (مرکز تحقیقات مخابرات ایران) تهیه شده است (شمس‌فرد و همکاران: ۲۰۱۰).

۲-۳-۷-پیکرهٔ وابستگی نحوی زبان فارسی: این پیکره، مجموعه‌ای است شامل ۳۰۰۰۰ جملهٔ برچسب-خورده با اطلاعات نحوی و ساخت‌واژی. این پیکره که نخستین پیکرهٔ وابستگی زبان فارسی است، می‌تواند به عنوان زیرساختی اساسی در پردازش رایانه‌ای زبان فارسی به کار رود. همچنین اطلاعات موجود در این پیکره می‌تواند در پژوهش‌های زبان‌شناختی و آموزش و یادگیری زبان فارسی مورد استفاده قرار گیرد. مهم‌ترین دلایل استفاده از دستور وابستگی در این پیکرهٔ نحوی عبارتند از: نتایج رضایت‌بخش در یادگیری خودکار و سازگاری مناسب با طبیعت زبان‌های بی-ترتیب، همچون زبان فارسی (منظور از بی‌ترتیب بودن زبان فارسی این است که برای اکثر جملات، ترتیب خاصی برای جملات نمی‌توان در نظر گرفت. مانند جملهٔ «من در مدرسه کتاب را به علی دادم» که به صورت‌های دیگر نیز می‌توان این مفهوم را بیان کرد: «من در مدرسه به علی کتاب را دادم» ، «من به علی در مدرسه کتاب را دادم» و «من کتاب را به علی در مدرسه دادم». ویژگی‌های پیکرهٔ وابستگی نحوی زبان فارسی شامل: جملات پیکره برگرفته از منابع مختلفی از متون فارسی معاصر هستند. تمامی جملات دارای برچسب روابط نحوی (بر مبنای دستور وابستگی) از قبیل فاعل، مفعول، مسند، مضافٌ‌الیه، بدل و....هستند. تمامی جملات دارای برچسب اطلاعات ساخت‌واژی (برچسب اجزای واژگانی کلام (POS)) از قبیل فعل، اسم، صفت، قید، ضمیر و....هستند. جملات توسط تیمی از زبان‌شناسان مجرب برچسب خورده‌اند و در چند مرحله بازبینی شده‌اند. داده‌های پیکره به صورت تصادفی، به داده‌های یادگیری (۸۰٪)، آزمون (۱۰٪) و ارزیابی (۱۰٪) تقسیم شده است. تعداد کل جملات ۲۹۹۸۲، تعداد کل واژه‌ها ۴۹۸۰۸۱، تعداد واژه‌های منحصر به فرد ۳۷۶۱۸، میانگین طول هر جمله ۶۱/۱۶، تعداد افعال منحصر به فرد ۴۷۸۲ و میانگین حضور هر فعل ۶۷/۱۲ است (رسولی و همکاران: ۲۰۱۳).

۲-۳-۸-پیکرهٔ واژگان زایای زبان فارسی: واژگان زایای زبان فارسی واژگانی است شامل حدود ۵۵ هزار مدخل که هر مدخل دارای اطلاعات مربوط به صورت نوشتاری واژه در خط فارسی، ساخت‌واجی، مقولهٔ واژگانی، الگوی تکیه و بسامد واژه می‌باشد. برای تهیهٔ واژگان زایا، یک

پیکرهٔ متنی ۱۰ میلیون کلمه‌ای ملاک استخراج واژه‌ها قرار گرفته است. واژگان زایا از حدود ۱۰۰ هزار کلمه با بسامدهای متفاوت تشکیل شده است. بعد از حذف صورت‌های تصریفی از فهرست فوق، حدود ۴۴ هزار واژه به مفهوم علمی آن به دست آمد. در بررسی فهرست واژه‌های حاصل از پیکرهٔ متنی معلوم شد که برخی واژه‌های عامیانه و برخی واژه‌های کاملاً علمی در فهرست ۴۴ هزار مدخلی غایب هستند. برای رفع این کاستی فهرست فوق با فرهنگ فارسی امروز (صدری‌افشار، ۱۳۸۱) مقایسه شد و حدود ۱۱ هزار مدخل جدید به فهرست واژه‌ها اضافه شد و واژگان ۵۵ هزار مدخلی به دست آمد (اسلامی و همکاران: ۱۳۸۳).

۲-۳-۹- پیکرهٔ تشخیص خودکار جنسیت: این پیکره شامل دو بخش اصلی است که عبارتند از: ۱- بخش متون رسمی که با مشخص کردن جنسیت نویسندگان متون داستانی موجود در پیکرهٔ بی‌جن‌خان و داستان‌های دیگر برگرفته از اینترنت به دست آمده است. ۲ - بخش متون غیررسمی. از رویکرد وب برای پیکره استفاده شده است. از نظرات کاربران در سایت «هلو کیش» نیز استفاده شده است. برای استخراج نظرات مرتبط با نظردهندگان زن و مرد، فهرستی از اسامی فارسی زن و مرد استفاده شده است که نام نویسندهٔ مدنظر با این فهرست تطبیق داده شده است و نظرات برحسب این فهرست تفکیک شده و در دو دسته زن و مرد قرار گرفته است (مرادی و بحرانی: ۱۳۹۴).

۲-۳-۱۰- پیکره چندزبانه رایانامه‌ها: این پیکره برای تشخیص ریسمان‌های گفتگوی چندزبانه در آزمایشگاه سیستم‌های هوشمند اطلاعات دانشگاه تهران تهیه شده است. نام این پیکره «Multilingual-BC3» است که در حقیقت یک پیکرهٔ ساختگی چندزبانه است که حاصل ترجمهٔ بخشی از پیکره تک‌زبانهBC3 ، توسط عامل انسانی است. پیکرهٔ اولیه BC3 به صورت تک‌زبانه و در زبان انگلیسی توسط آزمایشگاه هوش محاسباتی در دانشگاهBritish Columbia ساخته شده است. این پیکره، یک زیرمجموعه از پیکرهٔ W3C است که دارای برچسب‌های معنایی، نظیر حالت گفتار در سطح جمله و برچسب خلاصه‌سازی گفتگوها است. پیکره ConThread-BC3 یک نسخه از BC3 است که در آن، برچسب‌های نشان‌دهنده ساختار ریسمان‌های گفتگو و همچنین اطلاعات مربوط به برچسب متن اصلی و متن نقل‌قول رایانامه‌ها اضافه شده است. پیکرهٔ Multilingual-BC3، در دو نسخه تهیه شده که نسخه اول، گونه‌ای چندزبانه از نسخه اولیه BC3 و نسخه دوم گونه‌ای چندزبانه از ConThread-BC3 است. شایان

ذکر است که اطلاعات برچسب‌های موجود در نسخه‌های تک‌زبانه مستقل از زبان بوده و قابل گسترش به Multilingual-BC3 خواهند بود (دهقانی و همکاران: ۲۰۱۳).

۲-۳-۱۱- پیکرهٔ نقش‌های معنایی زبان فارسی: این پیکره که حدود ۳۰,۰۰۰ جمله از زبان فارسی معاصر را شامل می شود، به صورت دستی برچسب‌گذاری شده است. این پیکره بر اساس مفهوم نقش‌های معنایی فیلمور، لایه‌ای از اطلاعات مربوط به رابطهٔ محمول موضوع را به ساخت نحوی پیکرهٔ وابستگی اضافه می‌کند. در این مجموعه، افعال، اسم‌ها و صفت‌های گزاره‌ای به عنوان محمول‌های جمله در نظر گرفته شده و بنا بر نوع رویدادشان، در جمله تعیین ظرفیت شده‌اند. خروجی این پیکره بر اساس الگوی همایش زبان‌شناسی رایانه‌ای و پردازش زبان طبیعی (CoNLL)آماده شده است. نقش‌های معنایی مورد استفاده در برچسب‌زنی معنایی شامل دو گروه برچسب‌های موضوعی و برچسب‌های نقشی هستند. برچسب‌های موضوعی، ساخت ظرفیتی محمول‌های جمله را به دست می‌دهند و برچسب‌های نقشی، افزوده‌های توصیف‌گر فعل یا کل جمله را شامل می‌شوند. تعداد نقش‌های معنایی۲۷ مورد و تعداد موضوع‌های نقشی ۱۵ مورد است. دو برچسب وجه‌نما و نفی هم به عنوان برچسب نقشی مورد استفاده قرار گرفته است (میرزایی و مولودی: ۱۳۹۳).

۲-۳-۱۲- پیکرهٔ واژگان نحوی و معنایی افعال مرکب فارسی(نسخه ۱,۰)[1] : مجموعه‌ای است چندزبانه شامل اطلاعات نحوی و معنایی افعال مرکب زبان فارسی، ترجمهٔ انگلیسی و فرانسوی افعال و حداقل یک جمله مثال برای هر فعل. اطلاعات نحوی بر اساس دیدگاه گروس و اطلاعات معنایی بر اساس دیدگاه لوین تهیه شده‌اند. نسخهٔ اول این مجموعه که توسط پژوهشگران دانشگاه سوربن جدید فرانسه تهیه و عرضه شده است شامل اطلاعات مربوط به بیش از ۶۰۰ فعل مرکب شامل همکرد «زدن» است (سامولین و فقیری: ۲۰۱۳).

[1] PersPred

۲-٤-خلاصهٔ فصل

در این فصل به بررسی پژوهش‌هایی پرداخته شد که در حوزهٔ ساخت پیکره انجام شده است. در این پژوهش‌ها، چکیده‌ای از تجربیات پژوهشگران این حوزه دربارهٔ ساخت پیکره‌ها و چالش‌هایی که هنگام ساخت پیکره‌های متنی یا نوشتاری با آن‌ها مواجه شده‌اند، ارائه شده است. برخی دیگر از این پژوهش‌ها شامل توصیف انوع روش‌های ساخت پیکره، نوع پیکرهٔ ساخته شده و حاشیه‌نویسی‌های استفاده شده در پیکره است. در بخش انتهایی این فصل به معرفی برخی از پیکره‌های زبان فارسی پرداخته شد.

فصل سوم

ساخت پیکره و وارد کردن داده‌ها در پیکره

۳-۱- مقدمه

پیکرهٔ متنی ایرانداک در سامانه‌ای قرار داده شده است به نام «سامانهٔ پیکره‌های ایرانداک» تا بتوان در آینده، پیکره‌های دیگری را نیز در این سامانه قرار داد. در این فصل، پس از معرفی سامانهٔ تحت وب پیکره‌های ایرانداک و پرداختن به مشخصات فنی و تکنولوژی‌ها و زبان‌های رایانه‌ای استفاده شده برای سامانه و پیکره، به چگونگی وارد کردن داده‌ها در پیکره پرداخته می‌شود.

۳-۲- معرفی سامانهٔ تحت وب پیکره‌های ایرانداک

به منظور مدیریت و انتشار پیکرهٔ متنی، سامانه‌ای برخط با به‌کارگیری فناوری‌های نوین توسعه داده شد. با استفاده از این سامانه، امکاناتی فراهم گردیده که از طریق آن بتوان پیکره متنی موجود را با به‌کارگیری از ابزارهای ساده مدیریت کرد. امکاناتی که برای بخش مدیریت داده‌های موجود در پیکره درنظر گرفته شده است شامل موارد زیر است:

الف. امکان افزودن، حذف و ویرایش یک واژه

ب. امکان افزودن، حذف و ویرایش برچسب‌های اختصاص داده شده به یک واژه

برای دسترسی به این سامانه سه سطح دسترسی پیش‌بینی شده است. سطح اول برای بازدیدکنندگان پایگاه وب درنظر گرفته شده است. این کاربران می‌توانند بدون عضویت از امکانات جستجو و مشاهده

پیکرهٔ متنی استفاده نمایند. سطح دوم برای کاربرانی طراحی شده است که درخواست دانلود پیکرهٔ متنی را دارند. این سطح دسترسی برای دو هدف درنظر گرفته شده است. هدف اول، جمع‌آوری اطلاعات آماری از مخاطبان پایگاه وب است. از این اطلاعات می‌توان برای فعالیت‌های پژوهشی و امور مشابه بهره برد. هدف دوم از طراحی این سطح دسترسی، مدیریت دانلود دادگان پیکرهٔ متنی است. باتوجه به حجم قابل-ملاحظه پیکرهٔ متنی، این نیاز وجود دارد که داده‌ها با اعمال محدودیت و کنترل بر روی پهنای باند در دسترس قرار گرفته و از دانلود ماشینی جلوگیری شود. سطح سوم برای راهبر پایگاه وب توسعه داده شده که امکانات اشاره شده در بالا برای این سطح دسترسی، پیاده سازی شده است.

برای بازدیدکنندگان و کاربران پایگاه وب نیز امکاناتی برای جستجو و فیلتر کردن برچسب‌های موجود در پیکرهٔ متنی طراحی شده است. این ویژگی‌ها می‌تواند کمک‌کنندهٔ کاربران پایگاه وب و مخاطبان پیکرهٔ متنی برای شناخت بهتر جزئیات پیکره باشد. با توجه به این نکته که معمولاً پیش از بهره‌گیری از پیکره‌های متنی در پروژه‌های پژوهشی در حوزه پردازش زبان طبیعی این نیاز اساسی وجود دارد که دادگان موجود در پیکره و برچسب‌های آن‌ها مورد بررسی قرار گیرد، طراحی این بخش می‌تواند مخاطبان پیکرهٔ متنی را در این گام یاری کند. همچنین بخش آماری برای کاربران توسعه داده شده که به صورت برخط جزئیات شمار برچسب‌ها را نمایش می‌دهد. در این بخش صفحه‌ای توسعه داده شده که شمار برچسب‌های به کار رفته در پیکره متنی را نمایش می‌دهد .

با وارد کردن لینک زیر می‌توان به سامانهٔ پیکره‌های ایرانداک دسترسی داشت:

http://peykare.irandoc.ac.ir/

شکل ۳-۱، صفحه اول سامانهٔ پیکره‌های ایرانداک را نمایش می‌دهد.

شکل ۳-۱: صفحه اول سامانه پیکره‌های ایرانداک

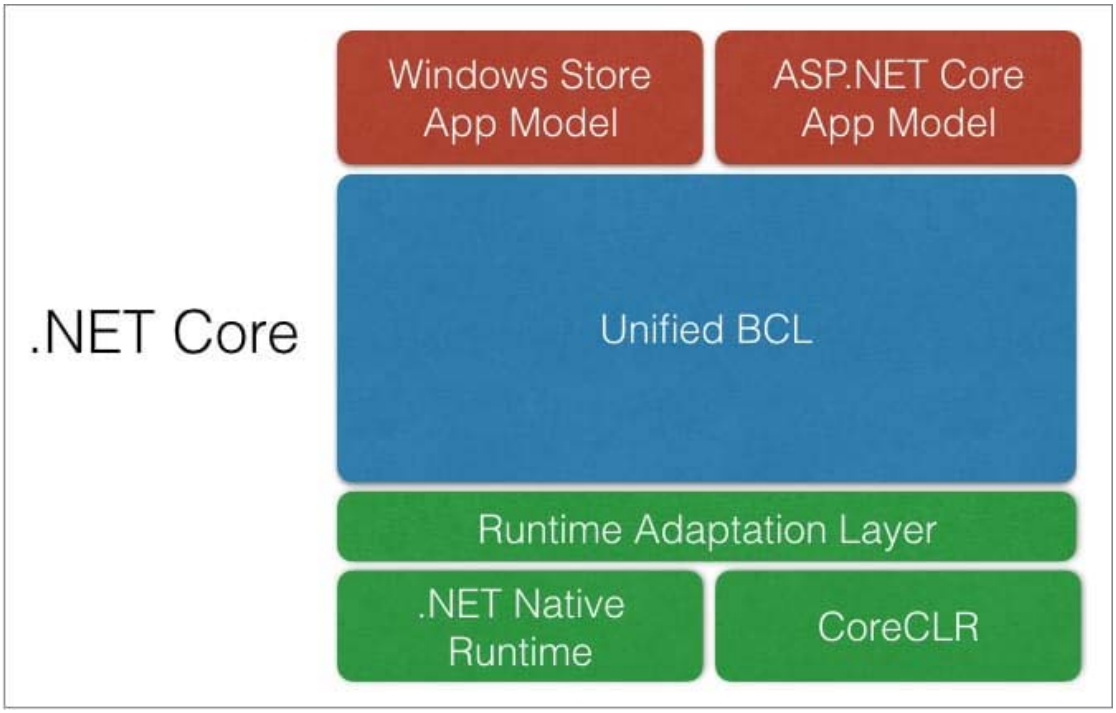
۳-۳- مشخصات فنی سامانهٔ پیکره‌های ایرانداک

۳-۳-۱- چارچوب اصلی سامانه

چارچوب اصلی سامانه، Net Core. (دات نت کور) می‌باشد. Net Core. پلت‌فرمی است چندمنظوره برای توسعه[1] و برنامه‌نویسی که توسط مایکروسافت و مجموعه Net. درGitHub تهیه شده است.

[1] development

چارچوب Net Core. کاملاً cross-platform طراحی شده است، به شکلی که در ویندوز، لینوکس و مک قابل استفاده بوده است. همچنین برای دستگاه‌های مختلف، فضاهای ابری و سناریوهای مرتبط با اینترنت اشیاء کاربرد دارد. شکل ۳-۲، چارچوب اصلی Net Core. را نشان می‌دهد.



شکل ۳-۲: چارچوب اصلی Net Core.

۳-۳-۲- زبان‌های برنامه‌نویسی

در طراحی و پیاده‌سازی سامانه پیکره‌های ایرانداک از زبان‌های برنامه‌نویسی زیر استفاده شده است:

- C# (سی‌شارپ[1]) در این سامانه برای توسعهٔ سمت سرور(Back End) از این زبان برنامه‌نویسی استفاده شده است. این زبان در بطن چارچوب Net Core. قرار دارد.
- HTML[2] (اچ‌تی‌ام‌ال)
- CSS (سی‌اس‌اس)[1]

[1] C sharp
[2] Hyper Text Markup Language (HTML)

- Javascript: در سرتاسر سامانهٔ پیکره‌های ایرانداک به صورت گسترده از کدهای JavaScript استفاده شده است. این زبان به همراه چارچوب[۲] Jquery و Angular بخش اصلی کدهای Client Side (سمت کاربر) را تشکیل می‌دهد.

۳-٤- تکنولوژی‌های استفاده شده در سامانه

در سامانه پیکره‌های ایرانداک از تکنولوژی‌های زیر برای توسعه و پیاده‌سازی استفاده شده است:

- Jquery (جی کوئری)
- AngularJs
- Bootstrap
- Linq[۳]
- Entity Framwork
- Sql Server

سامانه پیکره‌های ایرانداک از طریق تکنولوژی Entity Framework در Net Core. از امکانات Sql Server استفاده می‌کند. در واقع این تکنولوژی استفاده از دستورات Sql را از طریق زبان Linq در خود کد سامانه امکان‌پذیر می‌کند.

۳-۵- معماری سامانه

۳-۵-۱- الگوی توسعه MVC

الگوی MVC مخفف سه کلمه Model (مدل) و View (نمایشگر) و controller (کنترلگر) است. برخی از برنامه‌نویسان، همچنان از ASP.NET که بر مبنای فرم‌های وب و Postback است، استفاده می‌کنند. برخی از ویژگی‌های MVC سود می‌برند و برخی دیگر هم دو پلت‌فرم را ترکیب می‌کنند؛ این موضوع بیانگر این است که هیچکدام از پلت‌فرم‌ها ناقض یکدیگر نیستند. در واقع MVC بر روی

---

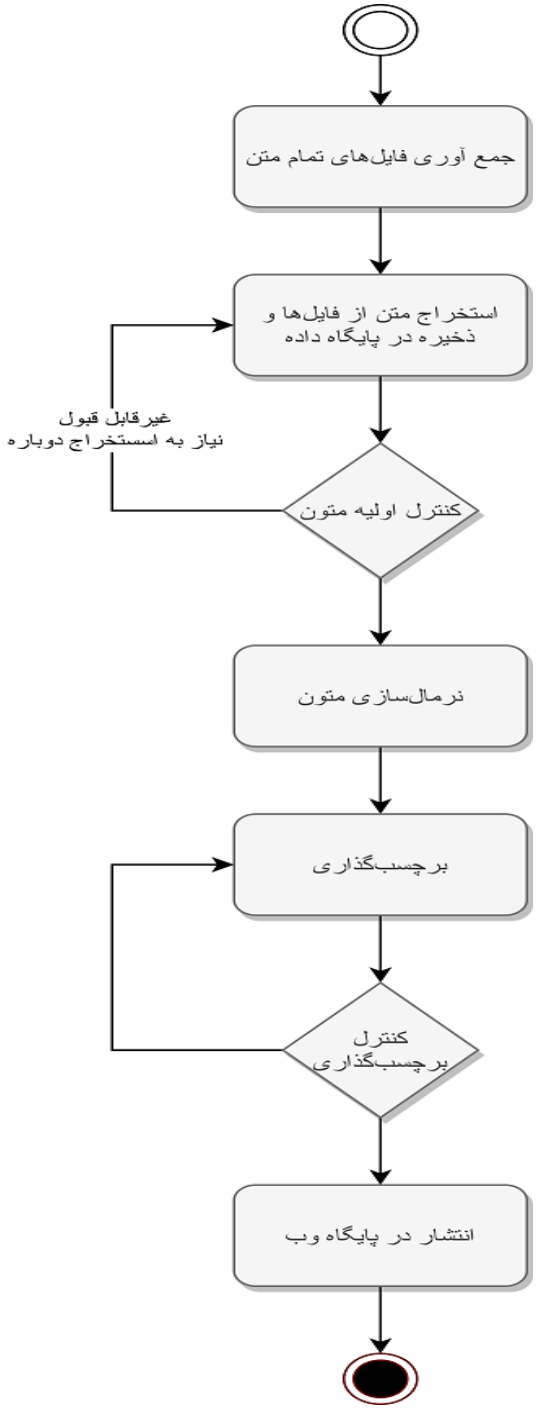
[1] Cascading Style Sheets (CSS)
[2] framework
[3] Language-Integrated Query

معماری‌های چندلایه‌ای برای تفکیک بخش‌های مختلف برنامه (بخش‌های منطقی برنامه مانند داده‌ها، مجوزها، کنترل صحت داده‌ها و لایه‌های مرتبط با کاربر نهایی) قرار می‌گیرد.

۳-۵-۲- اجزای تشکیل‌دهندهٔ MVC

- Model (مدل): قسمتی از برنامه کاربردی است که مسئول بازیابی داده از بانک اطلاعاتی، ذخیره آن، تبدیل آن به شی یا objectها و پیاده‌سازی منطق برنامه برای داده‌های دامنهٔ مسئله است. در حقیقت، بار اصلی معماری MVC بر عهدهٔ این بخش است. به عنوان مثال، یک object Product ممکن است اطلاعات را از بانک اطلاعاتی بازیابی کرده، بر روی آن‌ها عملیاتی را انجام دهد و سرانجام نتیجه را در بانک اطلاعاتی یا در جدول Products ذخیره کند.
- View (نمایشگر): جزئی از برنامه است که واسط کاربری برنامه (UI) را می‌سازد. معمولاً این UI از داده‌های مدل ساخته می‌شود. در حقیقت، نقطه پایان برنامه کاربردی است. به کاربر نتایج عملیات و بازیابی و نمایش داده از طریق برقراری ارتباط با دو بخش دیگر، یعنی مدل و کنترلگر، را نشان می‌دهد. برای مثال، هنگامی که کاربر در فرم ورود سیستم، رمز عبور خود را وارد می‌کند، اکثر برنامه‌نویسان در همان فرم اقدام به چک کردن رمز عبور می‌کنند که این عمل مغایر با قوانین MVC است. در MVC ، هنگامی که کاربر رمز عبور را وارد کرد، رمز عبور بدون هیچگونه اعمالی به بخش‌های دیگر فرستاده می‌شود و فقط یک نتیجهٔ ساده یا خبر از بخش‌های دیگر دریافت می‌کند که از طریق آن اجازهٔ ورود به برنامه داده می‌شود.
- Controllers (کنترلگرها): قسمت‌هایی از برنامه هستند که مدیریت تعامل با کاربر را بر عهده دارند. می‌توان گفت که واسط بین مدل و نمایشگر هستند؛ یعنی مدل کار می‌کند و در انتها نمایشگری را برای نشان دادن واسط کاربری انتخاب می‌کند. ورودی کاربر را مدیریت کرده و به آن‌ها پاسخ می‌دهد و با کاربر تعامل می‌کند. برای مثال، کنترلگر، عبارت‌های پرس‌وجوی بانک اطلاعاتی را مدیریت کرده و آن‌ها را به مدل ارسال می‌کند، وظیفه اجرای پرس‌وجوها با مدل است.

شکل ۳-۳، نمودار جریانی(block diagram) از ماجول‌های ساخت پیکره را نشان می‌دهد.



شکل ۳-۳: نمودار جریانی(block diagram) از ماجول‌های ساخت پیکره

۳-٦- جمع‌آوری و وارد کردن داده‌ها در پیکره

در طراحی کلی پیکره، ملاحظاتی از قبیل نوع متون مورد استفاده، تعداد متون، انتخاب متون خاص، طول نمونه‌های متون و....وجود دارد که هر کدام مستلزم تصمیم‌گیری دربارهٔ نمونه‌گیری است. تصمیماتی که در این زمینه گرفته می‌شود بر مناسب بودن پیکره برای انواع تجزیه و تحلیل زبانی تاثیر می‌گذارد. نمونه‌گیری در واقع عمل انتخاب متون مربوط به هر ژانر با توجه به هدف تهیهٔ پیکره است. برخی از معیارهایی که بر اساس آن‌ها نمونه‌گیری صورت می‌پذیرد شامل: شکل[1]متن (گفتاری/ نوشتاری/ الکترونیکی)، نوع متن (کتاب/ مجله/ نامه)، حوزهٔ متن (آکادمیک/ عمومی)، زبان (زبان‌ها یا گونه‌های زبانی پیکره) و مکان متن (به عنوان مثال، انگلیسی بریتانیا باشد یا استرالیا) است. داده‌های پیکرهٔ متنی ایرانداک از مقاله‌های پژوهش‌نامهٔ پردازش و مدیریت اطلاعات تهیه شده است؛ بنابراین، گونهٔ زبانی متون موجود در مجله، گونهٔ نوشتاری و رسمی است و حوزهٔ متون، آکادمیک است. توضیح مختصری دربارهٔ تاریخچه و حوزهٔ این مجله در پاراگراف بعدی ارائه می‌گردد:

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات دارای پروانه انتشار از وزارت فرهنگ و ارشاد اسلامی و یک ادواری علمی و دارای فرآیند داوری[2] است که اولین شمارهٔ آن با نام «نشریه فنی مرکز مدارک علمی» در مهر ماه سال ۱۳۵۱ منتشر گردید و انتشار آن با تغییر نام به «فصل‌نامه‌ی علوم اطلاع‌رسانی»، سپس به «فصل‌نامه علوم و فناوری اطلاعات» و اکنون به «پژوهش‌نامهٔ پردازش و مدیریت اطلاعات» ادامه یافته است. مالکیت مادی و معنوی این ادواری از آن پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) است. پژوهش‌نامهٔ پردازش و مدیریت اطلاعات، گزارش پژوهش‌های نوین را در قالب مقاله‌های پژوهشی، مروری، و مطالعه موردی در زمینهٔ نظریه‌پردازی، رهیافت‌های فلسفی، روش‌شناسی، روندپژوهی، آینده‌نگاری، معماری، مدل، الگوریتم، استاندارد، رهنمود، زبان، جنسیت، آموزش، و یادگیری؛ در قلمرو «اطلاعات و دانش علمی و فناورانه» و در پیوند با محورهای زیر، برای بررسی و انتشار می‌پذیرد:

الف. پردازش اطلاعات

- تولید، فراهم‌آوری، و انتقال
- سازمان‌دهی

[1] mode
[2] peer review

- ذخیره و بازیابی
- حفظ و آرشیو
- ارزیابی و تحلیل
- مدیریت کیفیت
- بازنمایی و دیداری‌سازی
- ابزارهای زبانی و زبان‌شناختی

ب.جامعه و اطلاعات

- سیاست‌گذاری
- اخلاق و حقوق
- اقتصاد، صنعت، و بازار
- نهاد و سازمان
- نوآوری و کارآفرینی
- رسانه
- مشارکت بخش عمومی و خصوصی
- دسترسی آزاد، علم آزاد، و داده آزاد
- پایش و ارزیابی تأثیر
- اشتراک منابع

پ.انسان و اطلاعات

- تعامل انسان و ماشین
- رفتار اطلاع‌یابی
- تجربه و رابط کاربری
- سواد (اطلاعات، دیجیتال، محتوا، و رسانه)

ت.فناوری و اطلاعات

- سیستم‌های مدیریت اطلاعات و دانش
- سیستم‌های هوشمند و پیشنهاددهنده
- سیستم‌های جست‌وجو
- معماری‌های پلتفرمی
- امنیت
- اینترنت اشیا
- کسب‌وکار الکترونیک
- اکتشاف دانش
- مدیریت محتوا
- تحلیل داده‌های کلان

ث.دیگر زمینه‌ها، موضوع‌ها، و فناوری‌های نوین در پردازش و مدیریت اطلاعات و دانش علمی و فناورانه

نمونه‌گیری به این صورت انجام شد: جامعهٔ آماری که شامل مجموعه مقاله‌های موجود در پژوهش‌نامهٔ پردازش و مدیریت اطلاعات است، شامل ۱۰۸۹ مقاله است؛ این مقاله‌ها در بازه زمانی سال ۱۳۵۱ تا ۱۳۹۹ در این مجله چاپ شده‌اند. ابتدا تصمیم بر این بود که همهٔ این مقاله‌ها در پیکره وارد شوند، اما از آنجائیکه فایل ورد (word) همهٔ این مقاله‌ها موجود نبود و در حال حاضر به دلیل محدودیت‌های خاصی که با آن مواجه هستیم، امکان تبدیل فایل‌های پی‌دی‌اف (pdf) به ورد، وجود ندارد، این تصمیم گرفته شد که مقاله‌هایی که فایل ورد آن‌ها موجود است، در پیکره وارد شود؛ در نهایت، متن ۶۷۷ مقاله که فایل ورد آن‌ها موجود بود، وارد پیکره شد. بنابراین، روش نمونه‌گیری خاصی برای انتخاب متون پیکره، به کار برده نشده است و هر آنچه موجود بوده است، وارد پیکره شده است. موضوع این مقاله‌ها شامل کتابداری و اطلاع‌رسانی، علم اطلاعات و دانش‌شناسی، فناوری اطلاعات، مدیریت اطلاعات، مدیریت دانش، زبان-شناسی، زبان‌شناسی رایانشی، اصطلاح‌شناسی، هستان‌شناسی و سایر حوزه‌های مرتبط با پردازش اطلاعات

است. یکی از راه‌های مرتب کردن انواع فایل، استفاده از اطلاعات توصیفی است. این داده‌های توصیفی، فراداده[1] نامیده می‌شود و بسته به نوع فایل، متفاوت است. اطلاعات به این صورت وارد پیکره شده است که فراداده از طریق کاربرگه ثبت مقالات وارد شده است. سپس از طریق خروجی گرفتن از پایگاه داده SQLSERVER به برنامهٔ نرمال‌سازی و برچسب‌گذاری منتقل شده و مجدد اطلاعات در پایگاه داده ذخیره شده است.

برای پیشگیری از خراب شدن داده‌ها هنگام انتقال دادن داده‌ها به پایگاه داده، بخش بزرگی از داده‌ها به صورت دستی وارد پایگاه داده شد. اطلاعات موجود در همهٔ مقاله‌های پیکره شامل عنوان مقاله، موضوع مقاله، نام نویسنده و متن مقاله در قسمت مربوطه وارد شد. نمونه‌ای از این فضا در شکل ۳-۴ قابل مشاهده است.

شکل ۳-۴: بخش افزودن داده‌ها به پیکره

[1] Metadata

۳-۷- خلاصه فصل

در این فصل، پس از معرفی سامانهٔ تحت وب پیکره‌های ایرانداک و پرداختن به مشخصات فنی و تکنولوژی‌ها و زبان‌های رایانه‌ای استفاده شده برای آن، به چگونگی وارد کردن داده‌ها در پیکره پرداخته شد.

فصل چهارم

نرمال‌سازی و حاشیه‌نویسی داده‌های پیکره

٤-١- مقدمه

در این فصل ابتدا به تعریف «نرمال‌سازی» و چگونگی نرمال‌سازی داده‌های واردشده در پیکره پرداخته می‌شود. سپس مفهوم «حاشیه‌نویسی» توضیح داده خواهد شد و در نهایت، نوع حاشیه‌نویسی/برچسب-گذاری استفاده شده در پیکرهٔ متنی ایرانداک و ابزار استفاده شده برای این منظور، معرفی می‌گردد و در نهایت نحوهٔ کنترل دستی برچسب‌ها توضیح داده می‌شود.

٤-٢- نرمال‌سازی داده‌ها

در ساخت پیکرهٔ متنی، پس از انتخاب متونی که در پیکرهٔ متنی مورد استفاده قرار خواهند گرفت، باید پیش‌پردازش‌هایی روی متن انجام شود که به اصطلاح «نرمال‌سازی متن[1]» نامیده می‌شود. این پیش-پردازش‌ها در متون فارسی می‌تواند شامل، یک‌دست کردن فاصله‌ها و نشانه‌گذاری‌های درون متن، یکسان کردن کاراکترهای استفاده شده در متون (مانند انواع «ی»، «ک»، همزه و...)، یکسان کردن روش اتصال وندهای گوناگون به ستاک، اصلاح غلط‌های املایی، ارتباط دادن کلمات چندواژه‌ای و یکسان در نظر گرفتن آن‌ها و....باشد. ابتدا باید همهٔ نویسه‌های (کاراکترهای) متن با جایگزینی با معادل استاندارد آن، یکسان‌سازی گردند. متون مختلف ممکن است بسیار به هم شبیه باشند اما به دلیل تفاوت‌های ساده ظاهری از نظر ماشین متفاوت باشند؛ به همین دلیل ابتدا باید این تفاوت‌های سادهٔ ظاهری برطرف گردد. همچنین در این مرحله، اصلاحات دیگری نیز به منظور پردازش دقیق‌تر متون انجام می‌پذیرد. برای رسیدن به این هدف، قبل از مقایسهٔ متون، پیش‌پردازش‌هایی روی آن‌ها انجام می‌شود. طبیعتاً هر چه این پیش‌پردازش‌ها

[1] normalization

قوی‌تر باشد، نتایج حاصل ازمقایسه متون قابل‌اطمینان‌تر خواهد بود. در پردازش خط فارسی، با توجه به نزدیکی که با خط عربی دارد، همواره در پردازش تعدادی از نگاره‌ها مشکلاتی وجود دارد که در اولین گام باید مشکلات مربوط به این نگاره‌ها را برطرف ساخت. علاوه بر این، اصلاح و یکسان‌سازی نویسهٔ نیم‌فاصله و فاصله در کاربردهای مختلف آن و مواردی مشابه برای یکسان‌سازی متون، از اقدامات لازم قبل از شروع فازهای مختلف ساخت پیکره است. در این راستا می‌توان از ابزارهای آماده استفاده کرد. مهم‌ترین ابزارهای پردازش زبان طبیعی در متون عبارتند از:

- ابزار تشخیص جمله[1]: این ابزار باید با توجه به کاراکترهای جداکنندهٔ جمله در خط و زبان، توانایی تشخیص جملات را در متن ورودی داشته باشد. برای ایجاد این ابزار باید ابتدا تمامی کاراکترها، نمادها و احیاناً قواعد دستوری که باعث شکسته شدن جملات می‌شوند، شناسایی گردند. با توجه به پایه بودن جمله در بسیاری از پردازش‌های زبانی، خروجی دقیق این ابزار از درجهٔ اهمیت بالایی برخوردار است. از نمونه‌های انگلیسی آن می‌توان به OpenNLP ، Stanford NLP، NLTKو Freeling اشاره کرد.
- ابزار جداسازی/واحدسازی[2]: ابزاری برای شکستن یک متن بر اساس واحدهای بامعنی مانند کلمه، پاراگراف و نمادهای معنادار مانند space (فاصله) و tab و . ... است؛ لازمهٔ ایجاد این ابزار در فارسی، جمع‌آوری واحدهایی است که در زبان فارسی به عنوان واحدهای مستقل معنایی شناخته می‌شوند؛ سپس بر اساس انتخاب هر کدام از این واحدها، متن بر اساس آن شکسته خواهد شد. از نمونه‌های انگلیسی آن می‌توان به Flex ، JLex، JFLex، ANTLR، Ragelو Quexاشاره کرد.
- ابزار شناسایی ماهیت اسم‌ها[3]: ابزاری برای تشخیص اسم‌ها و نوع آن‌ها اعم از اسامی افراد، اماکن، مقادیر عددی و ... .است. برای تشخیص این‌که یک کلمه اسم است، راه‌های مختلفی وجود دارد که از جملهٔ آن‌ها مراجعه به لغت‌نامه، مراجعه به word-net ، در نظر گرفتن ریشهٔ کلمه، استفاده از قواعد نحوی؛ ساخت‌واژهٔ کلمه و ... می‌باشد. در این ابزار پس از تشخیص اسم‌ها با استفاده یک لغت‌نامه حاوی اسم افراد، مکان ها، مقادیر عددی و ... نوع اسم تشخیص

[1] Sentence recognition
[2] Tokenizer
[3] Named entity recognition

داده می‌شود. از جمله نمونه‌های انگلیسی این ابزار می‌توان به Stanford NER و Illinois NER اشاره کرد.

- ابزار ریشه‌یابی[1]: ابزاری برای ریشه‌یابی لغات و تشخیص نوع کلمه ساخته شده از آن ریشه (اسم مکان، اسم زمان، صفت فاعلی، مفعولی و....) است. معمولاً ریشه‌یابی لغات بر اساس قواعد ساخت‌واژی و سپس حذف پسوندها صورت می‌پذیرد. معروف‌ترین الگوریتم ریشه‌یابی در انگلیسی porter می باشد.
- ابزار تشخیص میزان شباهت[2]: ابزاری برای تشخیص میزان شباهت میان دو عبارت بر اساس پارامترهای مختلف مانند نوع اسم‌های مشابه به کار رفته، استفاده از word-net و... .است. در این ابزار پس از تشخیص نوع کلمات به کار رفته در یک جمله، بر اساس جایگاه آن کلمات در جمله، کلماتی که در جایگاه‌های یکسان قرار دارند، مورد مقایسه قرار می‌گیرند. از نمونه‌های انگلیسی آن می توان به Illinois NESim و Illinois WNSim اشاره نمود.
- ابزار تشخیص گروه‌های نحوی[3]: ابزاری برای تشخیص گروه‌های اسمی، فعلی و .... در یک جمله است. از جمله نمونه‌های انگلیسی آن می‌توان به Illinois Chunker اشاره کرد.
- ابزار برچسب‌دهندهٔ نقش نحوی[4]: ابزاری برای تشخیص نقش نحوی کلمه در جمله است. این ابزار یکی از مهم‌ترین نقش‌ها را در پردازش‌های زبانی بر عهده دارد. دقت در این ابزار بسیار حائز اهمیت است. این ابزار باید نقش‌های نحوی کلمات در جمله‌ها مانند فاعل، مفعول مستقیم، مفعول غیر مستقیم و .... را تشخیص دهد. از جمله نمونه‌های انگلیسی آن می‌توان به OpenNlP، Illinois SRL، Swirl و LTHSRL اشاره کرد. این ابزارها از الگوریتم پارسینگ charniak استفاده می‌کنند.
- ابزار نشانه‌گذاری (در پیکره)[5]: ابزاری برای ایجاد یک نمونه از یک آنتولوژی در یک سند ورودی است. از ابزارهای موجود در انگلیسی می توان به Illinois Curator و Stanford Annotator اشاره کرد.

---

[1] Stemming
[2] Similarity recognition
[3] Chunker
[4] syntactic role labeler
[5] Annotator

- ابزار تعیین هم‌مرجع‌ها[1]: ابزاری برای تعیین مرجع اسمی یک اسم یا یک ضمیر در جملات است. استفاده از ضمایر به جای اسامی در زبان انگلیسی بسیار رایج است، اما در زبان فارسی این امر چندان رایج نیست. از نمونه‌های انگلیسی این ابزار می‌توان به Illinois Coreference package اشاره کرد.
- ابزار برچسب‌گذاری اجزای واژگانی کلام: ابزاری برای مشخص کردن نوع کلمات از قبیل اسم، صفت، قید، فعل و ... است. به عمل انتساب برچسب‌های واژگانی (اسم، صفت، قید، فعل و ...) به کلمات و نشانه‌های تشکیل‌دهندهٔ یک متن برچسب‌گذاری کلمات (POS Tagging) گفته می‌شود؛ به صورتی که این برچسب‌ها نشان‌دهندهٔ مقولهٔ کلمات و نشانه‌ها در جمله باشد.

پس از وارد کردن داده‌ها در پیکرهٔ متنی ایرانداک، نرمال‌سازی به صورت ماشینی انجام شد؛ در این راستا از کد ابزار «هضم» ، استفاده شد. «هضم» ابزاری است برای پردازش زبان فارسی در پایتون. ویژگی-های این ابزار شامل موارد زیر است (sobhe/hazm) :

- تمیز و مرتب کردن متن
- تقطیع جمله‌ها و واژه‌ها
- ریشه‌یابی واژه‌ها
- تحلیل صرفی جمله
- تجزیه نحوی جمله
- واسط استفاده از داده‌های زبان فارسی
- سازگاری با بسته NLTK
- پشتیبانی از پایتون نسخه ۲ و ۳
- تست مداوم کدها

از کدهای ابزار «هضم» که به رایگان در اختیار کاربران قرار گرفته است، کد واحدسازی/جداسازی (tokenizer) در پیکره استفاده شده است. ابزار واحدسازی/جداسازی، مرز واژه‌ها را در متون تشخیص می‌دهد و متن را به دنباله‌ای از واژه‌ها تبدیل می‌کند و آن را برای تحلیل‌های بعدی آماده می‌کند. در

[1] Coreference resolution

حقیقت، واحدسازی، تکه‌تکه متن به قسمت‌های کوچکی به نام واحد است. واحدسازی در سطح واژه رخ می‌دهد و واحدهای استخراج شده به عنوان ورودی ماجول‌های دیگر مانند ریشه‌یاب، برچسب‌گذاری و....استفاده می‌شوند. بنابراین، بخشی از پیش‌پردازش متن، واحدسازی/جداسازی است. همچنین برای یکدست کردن انواع «ی» و «ک» و «کسره اضافه روی «ه» (مانند «هٔ») از دستورات TSQL در برنامه پایگاه داده SQKServer نیز استفاده شده است. مقاله‌ها بر اساس دستورالعمل فرهنگستان زبان و ادب فارسی، ویرایش شده‌اند. اما به هر حال ممکن است فاصله‌ها و نیم‌فاصله‌ها و دیگر تنوعات نگارشی اصلاح نشده باشند. بنابراین، فهرستی از همه واژگان موجود در پیکره در اکسل تهیه شد و در آنجا نرمال‌سازی انجام شده توسط ماشین، تا حد امکان به صورت دستی نیز کنترل شد تا تنوعات نگارشی و فاصله‌ها و نیم-فاصله‌ها و.... بررسی شود و هر جا خطایی صورت گرفته است، صورت درست واژه وارد پیکره شود.

٤-۳- حاشیه‌نویسی داده‌ها

حاشیه‌نویسی، اضافه کردن اطلاعات زبانی توضیحی-تفسیری به پیکره است. یکی از رایج‌ترین نوع حاشیه‌نویسی پیکره، اضافه کردن برچسب‌ها به کلمات پیکره است؛ این برچسب‌ها می‌تواند نشان‌دهندهٔ مقولهٔ کلمات (اسم، فعل، صفت...) باشد که «برچسب‌گذاری اجزای واژگانی کلام» نامیده می‌شود. یکی از کاربردهای این نوع برچسب‌گذاری، تمایز کلماتی است که یکسان نوشته می‌شوند اما معنی و گاهی تلفظ متفاوت دارند و در اصطلاح «هم‌نگاره[1]» خوانده می‌شوند. به عنوان مثال، صورت نوشتاری <present> در انگلیسی ممکن است از مقولهٔ «اسم» باشد (به معنی «هدیه») ، «فعل» باشد (به معنی «هدیه دادن») یا «صفت» باشد (به معنی «حاضر»). این صورت نوشتاری می‌تواند در پیکره به صورت زیر برچسب‌گذاری شده باشد:

*present*_NN1 (singular common noun)
*present*_VVB (base form of a lexical verb)
*present*_JJ (general adjective)

[1] homograph

برخی از پژوهشگران مانند سینکلر (۲۰۰۴، نقل از لیچ: ۲۰۰۴)، ترجیح می‌دهند وارد مقولهٔ حاشیه-نویسی پیکره نشوند؛ زیرا معتقد هستند پیکرهٔ حاشیه‌نویسی نشده، پیکره‌ای خالص است و برای مطالعات زبان‌شناسی چنین پیکره‌هایی را ترجیح می‌دهند. این پژوهشگران به حاشیه‌نویسی اعتماد ندارند و با شک به آن می‌نگرند و معتقد هستند حاشیه‌نویسی نمی‌تواند بدون خطا باشد. اما اکثر پژوهشگران این حوزه بر این باورند که حاشیه‌نویسی، پیکره را بسیار مفیدتر و غنی‌تر از پیکرهٔ خام می‌کند؛ از نظر آن‌ها حاشیه-نویسی، ارزشی مضاعف به پیکره می‌دهد. به عنوان مثال، عمل برچسب‌گذاری اجزای واژگانی کلام که برای پیکرهٔ براون انجام شده است، در سطح گسترده‌ای توزیع شده و مورد استفادهٔ افرادی قرار گرفته است که در راستای پژوهش‌های خود نیازمند برچسب‌گذاری اجزای واژگانی کلام یا پیکرهٔ برچسب-گذاری شده بودند. غیر از برچسب‌گذاری اجزای واژگانی کلام، حاشیه‌نویسی‌های دیگری نیز وجود دارد که به سطوح مختلف تجزیه و تحلیل‌های زبانی یک پیکره مربوط می‌شود. به برخی از رایج‌ترین آن‌ها اشاره می‌کنیم (لیچ: ۲۰۰۴):

- حاشیه‌نویسی آوایی[1]: اضافه کردن اطلاعاتی دربارهٔ نحوهٔ تلفظ یک کلمه در پیکرهٔ گفتاری یا اضافه کردن مشخصه‌های زبرزنجیری[2] (مانند تکیه[3]، آهنگ کلام[4]، وقفه[5] و...) به کلمات تشکیل-دهندهٔ یک پیکره است.
- حاشیه‌نویسی نحوی[6]: اضافه کردن اطلاعاتی دربارهٔ نحوهٔ تجزیهٔ یک جمله به عبارت‌ها و اجزای تشکیل‌دهندهٔ آن است.
- حاشیه‌نویسی معنایی[7]: اضافه کردن اطلاعاتی دربارهٔ مقوله/ طبقه‌بندی معنایی کلمات پیکره است. به عنوان مثال، صورت نوشتاری <cricket> در انگلیسی، فارغ از صورت نوشتاری یا تلفظی یکسان، هم در طبقه‌بندی انواع ورزش (نوعی ورزش) قرار می‌گیرد و هم در طبقه‌بندی دیگر و متفاوتی تحت عنوان حشرات (نوعی حشره).

[1] phonetic annotation
[2] prosodic
[3] stress
[4] intonation
[5] pause
[6] syntactic annotation
[7] semantic annotation

- حاشیه‌نویسی کاربردی[۱]: اضافه کردن اطلاعاتی دربارهٔ کنش گفتار [۲]که در مکالمه اتفاق می‌افتد. به عنوان مثال، پاره گفتار <okay> در انگلیسی، در موقعیت‌های مختلف می‌تواند به منزلهٔ اقرار و تصدیق، درخواست بازخورد، پذیرفتن یا نشانهٔ شروع مرحلهٔ جدیدی از بحث باشد.
- حاشیه‌نویسی گفتمانی[۳]: اضافه کردن اطلاعاتی دربارهٔ ارتباطات ارجاعی[۴] در متن است. به عنوان مثال، ارتباط دادن ضمیر <them> و مرجع آن <the horses> در جملهٔ زیر :

I`ll saddle the horses and bring them round.

- حاشیه‌نویسی سبکی[۵]: اضافه کردن اطلاعاتی دربارهٔ نمود گفتار (گفتار مستقیم[۶]، گفتار غیرمستقیم[۷]...)
- حاشیه‌نویسی واژگانی[۸]: اضافه کردن اطلاعاتی مانند ریشهٔ کلمه.

موارد نام‌برده شده، معروف‌ترین و متداول‌ترین نوع حاشیه‌نویسی هستند؛ می‌توان انواع دیگر حاشیه‌نویسی را از پژوهش‌های انجام شده استخراج کرد یا حتی حاشیه‌نویسی‌های جدیدی را به فهرست حاشیه‌نویسی‌ها اضافه کرد. به عنوان مثال، می‌توان در پیکره‌های آموزشی زبان برچسبی تحت عنوان «برچسب خطا» تعریف کرد که مربوط به خطاهایی است که زبان‌آموزان در بافت‌های گوناگون مرتکب می‌شوند (گرنگر و همکاران: ۲۰۰۲، نقل از لیچ: ۲۰۰۴).

دو راه برای حاشیه‌نویسی وجود دارد: ۱- پیروی کردن از استانداردهای خاص که می‌توان آن‌ها را اساس حاشیه‌نویسی پیکره در نظر گرفت؛ ۲- راه دیگر این است که از ابزارهای خاص استفاده کنیم و حاشیه‌نویسی را به کمک نرم‌افزارهای خاصی که برای این منظور طراحی شده است، انجام دهیم (به عنوان مثال، ابزارهایی برای برچسب‌گذاری اجزای کلام یا ویرایش متن وجود دارد که می‌توان از آن‌ها استفاده کرد) و در نهایت، خروجی این حاشیه‌نویسی‌ها را دستی کنترل کنیم.

همانگونه که ذکر شد، یکی از پرکاربردترین نوع برچسب‌گذاری، برچسب‌گذاری اجزای واژگانی کلام است. این نوع برچسب‌دهی، عملی کاربردی در بسیاری از حوزه‌های پیشرفته‌تر پردازش زبان طبیعی[۱] از

---

[1] pragmatic annotation
[2] speech act
[3] discourse annotation
[4] anaphoric
[5] stylistic annotation
[6] direct speech
[7] indirect speech
[8] lexical annotaion

جمله ترجمهٔ ماشینی، خطایاب، تبدیل متن به گفتار، بازیابی اطلاعات، موتورهای جستجو و کمک به مدل‌های آماری است (مگردومیان: ۲۰۰۴). پس از ساخت پیکرهٔ متنی ایرانداک، با توجه به کاربرد برچسب اجزای واژگانی کلام در پردازش متن، تصمیم گرفته شد این نوع برچسب هم به پیکره اضافه شود. در این راستا تصمیم گرفته شد ابتدا از یکی از ابزارهای آماده برای برچسب‌گذاری ماشینی اجزای واژگانی کلام استفاده شود و سپس، برچسب‌ها دستی کنترل شوند. ابزار استفاده شده برای برچسب-گذاری ماشینی، ابزار «هضم» است که برای نرمال‌سازی متون پیکره نیز از این ابزار استفاده شد. فهرست برچسب‌ها، شامل برچسب‌هایی است که در مقالهٔ بی‌جن‌خان و همکاران (۲۰۱۱) معرفی شده است. این فهرست با مثال در جدول ۴-۱ آورده شده است:

جدول ۴-۱: فهرست برچسب‌ها

POS tag	**Tag Name**	**مثال**
N	Noun	کشاورز، خانه، کتاب
PREP	Preposition	از، در، برای
PUNC	Punction	نقطه، ویرگول، علامت سوال
AJ	Adjective	زیبا، اجتماعی، بزرگ
V	Verb	می‌تواند، خورد، شمرد
CON	Conjunction	که
NUM	Number	۱۲، ۹۰، ۵۰
PRO	Pronoun	من، تو، ایشان
DET	Determiner	این، آن، هر
ADV	Adverb	یقیناً، خوب، بسیار

[1] Natural language processing

POSTP	Postposition	را
RES	Residual	?
CL	Classifier	نوع، دست، چنین
INT	Interjection	ای، یا

نکته: در فهرستی که در جدول ۴-۱ آورده شده است، کسره اضافه نمایش داده نشده است؛ «هضم» به منظور نمایش کسره اضافه، هر گاه پس از واژه‌ای، کسرهٔ اضافه وجود داشته، کنار برچسب، e گذاشته است که نمایانگر کسرهٔ اضافه است. به عنوان نمونه، برچسب عبارت «شناسایی ساختار محتوایی مطالعات» به صورت: شناسایی (Ne) ساختار (Ne) محتوایی (ADJe) مطالعات (N) است.

شایان ذکر است، دقت برچسب‌گذاری ابزار «هضم» ۹۷٫۱٪ گزارش شده است (sobhe/hazm).

بسیاری از فرآیندهای ماشینی برچسب‌گذاری، بن‌واژه‌سازی و غیره مستلزم دسترسی به فهرست واژگان، وندها و واژه‌بست‌ها است. فهرست واژگان را می‌توان از طریق فرهنگ‌های لغت به دست آورد؛ فهرست وندها و واژه‌بست‌های فارسی در جدول ۴-۲ آورده شده است . این فهرست برگرفته از مجموعه ۱۲ مقالهٔ دکتر علی‌اشرف صادقی تحت عنوان « *شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر ۱ تا ۱۲ : ۱۳۷۰-۱۳۷۲*» و خسرو کشانی (در بحث اشتقاق) و لازار: ۱۳۸۹ و قطره: ۱۳۸۶ (در بحث تصریف) است.

جدول ۴-۲: فهرست وندها و واژه‌بست‌های فارسی

وندها	نوع وند	مثال
ها	پسوند تصریفی (شمار)	کتاب‌ها
ان	پسوند تصریفی (شمار)	درختان
ین	پسوند تصریفی (شمار)	معلمین
ات	پسوند تصریفی (شمار)	درجات
ی	پسوند تصریفی (یای نکره)	کتابی (در عبارت «کتابی خوب»)
ه	پسوند تصریفی (معرفه)	دختره
ش	پسوند تصریفی (معرفه)	آشش (در جمله «آشش خوبه»)

تر/ترین	پسوند تصریفی (نشانهٔ صفت تفصیلی و عالی)	خوب‌ترین/بهتر
ت/اد/د/ید/ست	پسوند تصریفی (نشانهٔ ماضی)	کشت/ایستاد/خورد/رسید/آراست
ان	پسوند تصریفی (متعدی ساز)	رساند (رس+ان+د)
می/ب/ن/م	پیشوند تصریفی	می‌گفت/بروم/نرو/مشنو
با	پیشوند اشتقاقی (صفت‌ساز)	باادب
بی	پیشوند اشتقاقی (صفت‌ساز)	بیکار
نا	پیشوند اشتقاقی (صفت‌ساز)	ناکام
هم	پیشوند اشتقاقی (صفت‌ساز)	همکار
ن	پیشوند اشتقاقی (صفت‌ساز)	نشکن
ام/ای/ه/س/ست//ایم/اید/اند (صورت‌های وابسته فعل بودن)	واژه‌بست	زنده‌ام/زنده‌ای/زنده‌ست/زنده‌ایم/زنده‌اید/زنده‌اند
و (صورت پی‌بستی «را»)	واژه‌بست	«و» در جمله «منو می‌بری؟» (من را می‌بری؟)
و (حرف هم‌پایگی یا عطف)	واژه‌بست	«و »در «من و تو»
م (صورت پی‌بستی «هم»)	واژه‌بست	«م» در «منم میام» (من هم می‌آیم)
ها و صورت محاوره آن (آ) (برای تاکید بعد از فعل می‌آید)	واژه‌بست	اومدم آ
م/ت/ش/مان/تان/شان (ضمایر پی‌بستی/پیوسته)	واژه‌بست	مدادم/مدادت/مدادمان/مدادتان/مدادشان
گاه	پسوند اشتقاقی (اسم مکان)	دانشگاه
ستان	پسوند اشتقاقی (اسم مکان)	کردستان
کده	پسوند اشتقاقی (اسم مکان)	دانشکده
ئیه	پسوند اشتقاقی (اسم مکان)	جمشیدیه
ک	پسوند اشتقاقی (اسم مکان)	قلهک
دان/دانی	پسوند اشتقاقی (اسم مکان)	نمکدان/ سگ‌دانی
زار	پسوند اشتقاقی (اسم مکان)	کشتزار
آباد	پسوند اشتقاقی (اسم مکان)	حسن‌آباد
ی	پسوند اشتقاقی (اسم مکان)	پنجری

سار	پسوند اشتقاقی (اسم مکان)	کوهسار
بار	پسوند اشتقاقی (اسم مکان)	جویبار
لاخ	پسوند اشتقاقی (اسم مکان)	سنگ‌لاخ
ان	پسوند اشتقاقی (اسم مکان)	دیلمان
بان	پسوند اشتقاقی (دال بر شغل، محافظت یا دارندگی)	نگهبان
چی	پسوند اشتقاقی (دال بر شغل، محافظت یا دارندگی)	کافه‌چی
دار	پسوند اشتقاقی (دال بر شغل، محافظت یا دارندگی)	صندوق‌دار
ی	پسوند اشتقاقی (دال بر شغل، محافظت یا دارندگی)	بازاری
یار	پسوند اشتقاقی (دال بر شغل، محافظت یا دارندگی)	استادیار
گر	پسوند اشتقاقی (دال بر شغل، محافظت یا دارندگی)	آهنگر
ساز	پسوند اشتقاقی (دال بر شغل، محافظت یا دارندگی)/ پسوندواره فعلی	آهنگ‌ساز
دوز	پسوند اشتقاقی (دال بر شغل، محافظت یا دارندگی) / پسوندواره فعلی	کفش‌دوز
باف	پسوند اشتقاقی (دال بر شغل، محافظت یا دارندگی) / پسوندواره فعلی	فرش‌باف
پز	پسوند اشتقاقی (دال بر شغل، محافظت یا دارندگی) / پسوندواره فعلی	شیرینی‌پز
ه	پسوند اشتقاقی (اسم اشیاء و اسم معنی)	پایه

ک	پسوند اشتقاقی (اسم اشیاء و اسم معنی)	عروسک
ئیه	پسوند اشتقاقی (اسم اشیاء و اسم معنی)	فطریه
ی	پسوند اشتقاقی (اسم اشیاء و اسم معنی)	گوشی
گان	پسوند اشتقاقی (اسم اشیاء و اسم معنی)	ناوگان
واره	پسوند اشتقاقی (اسم اشیاء و اسم معنی)	جشنواره
ئینه	پسوند اشتقاقی (اسم اشیاء و اسم معنی)	گنجینه
ک	پسوند اشتقاقی (تصغیر و تحبیب)	طفلک
چه	پسوند اشتقاقی (تصغیر و تحبیب)	تاریخچه
ی	پسوند اشتقاقی (تصغیر و تحبیب)	مامانی
ه	پسوند اشتقاقی (تأنیث)	همشیره
ا	پسوند اشتقاقی (تأنیث)	آتوسا
یه	پسوند اشتقاقی (تأنیث)	خیریه
ی	پسوند اشتقاقی (مصدر/حاصل مصدر)	آزادی
ن	پسوند اشتقاقی (مصدر/حاصل مصدر)	رفتن
ا	پسوند اشتقاقی (مصدر/حاصل مصدر)	ژرفا
کار	پسوند اشتقاقی (اسم‌ساز)	فراموش‌کار
بندی	پسوند اشتقاقی (اسم‌ساز)	طبقه‌بندی
گری	پسوند اشتقاقی (اسم‌ساز)	وحشی‌گری
کاری	پسوند اشتقاقی (اسم‌ساز)	کتک‌کاری
بازی	پسوند اشتقاقی (اسم‌ساز)	کاغذبازی
ش	پسوند اشتقاقی (اسم‌ساز)	آرایش

ار	پسوند اشتقاقی (اسم‌ساز)	خریدار
ه	پسوند اشتقاقی (اسم‌ساز)	خنده
مان	پسوند اشتقاقی (اسم‌ساز)	زایمان
ان	پسوند اشتقاقی (اسم‌ساز)	آشتی‌کنان
ئیت	پسوند اشتقاقی (اسم‌ساز)	حساسیت
مند	پسوند اشتقاقی (دال بر دارندگی و اتصاف)	ثروتمند
ور/اور/وار	پسوند اشتقاقی (دال بر دارندگی و اتصاف)	بارور/سزاوار
ناک	پسوند اشتقاقی (دال بر دارندگی و اتصاف)	اندوهناک
ه	پسوند اشتقاقی (دال بر دارندگی و اتصاف)	همه‌کاره
نده	پسوند اشتقاقی (فاعلی و مفعولی)	فرساینده
ان	پسوند اشتقاقی (فاعلی و مفعولی)	لرزان
ا	پسوند اشتقاقی (فاعلی و مفعولی)	توانا
ار	پسوند اشتقاقی (فاعلی و مفعولی)	پرستار
ی	پسوند اشتقاقی (فاعلی و مفعولی)	شکاری
گار	پسوند اشتقاقی (فاعلی و مفعولی)	خواستگار
ند	پسوند اشتقاقی (فاعلی و مفعولی)	خوشایند
ه	پسوند اشتقاقی (فاعلی و مفعولی)	پخته
ی	پسوند اشتقاقی (دال بر نسبت)	دنده‌ای
ین	پسوند اشتقاقی (دال بر نسبت)	آهنین
ئینه	پسوند اشتقاقی (دال بر نسبت)	دیرینه
گان	پسوند اشتقاقی (دال بر نسبت)	خدایگان
گانه	پسوند اشتقاقی (دال بر نسبت)	پنج‌گانه
وند	پسوند اشتقاقی (دال بر نسبت)	شهروند
ئیه	پسوند اشتقاقی (دال بر نسبت)	تحریریه
و	پسوند اشتقاقی (دال بر نسبت)	اخمو
گین	پسوند اشتقاقی (دال بر نسبت)	غمگین

گون	پسوند اشتقاقی (دال بر شباهت)	گندم‌گون
فام	پسوند اشتقاقی (دال بر شباهت)	گل‌فام
آسا	پسوند اشتقاقی (دال بر شباهت)/ پسوندواره فعلی	رعدآسا
سان	پسوند اشتقاقی (دال بر شباهت)	گربه‌سانان
وار	پسوند اشتقاقی (دال بر شباهت)	دیوانه‌وار
وش	پسوند اشتقاقی (دال بر شباهت)	مهوش
انه	پسوند اشتقاقی (دال بر شباهت)	دلیرانه
سا	پسوند اشتقاقی (دال بر شباهت)	پریسا
وی	پسوند اشتقاقی (صفت‌ساز)	دنیوی
باره	پسوند اشتقاقی (دال بر کیفیت جسمی یا روحی)	شکمباره
آگین	پسوند اشتقاقی (دال بر کیفیت)	زهرآگین
آنی	پسوند اشتقاقی (صفت‌ساز)	عصبانی
گرا	پسوند اشتقاقی (صفت‌ساز)/ پسوندواره فعلی	سنت‌گرا
نما	پسوند اشتقاقی (اسم‌ساز) / پسوندواره فعلی	سال‌نما
طلب	پسوند اشتقاقی (صفت‌ساز)/ پسوندواره فعلی	استقلال‌طلب
پرست	پسوند اشتقاقی (صفت‌ساز)/ پسوندواره فعلی	خداپرست
سنج	پسوند اشتقاقی (اسم ابزار)/ پسوندواره فعلی	فشارسنج
بر	پسوند اشتقاقی (اسم دارو)/ پسوندواره فعلی	تب‌بر
پذیر	پسوند اشتقاقی (صفت‌ساز)/ پسوندواره فعلی	امکان‌پذیر
خیز	پسوند اشتقاقی (صفت‌ساز)/ پسوندواره فعلی	نفت‌خیز

شناس	پسوند اشتقاقی (صفت‌ساز)/ پسوندواره فعلی	هواشناس
دان	پسوند اشتقاقی (صفت‌ساز)/ پسوندواره فعلی	ریاضیدان
نشین	پسوند اشتقاقی (اسم‌ساز، معرف نژادو و قوم و طبقه اجتماعی) / پسوندواره فعلی	کارگرنشین
خواه	پسوند اشتقاقی (صفت‌ساز/ اسم‌ساز، معرف یک رژیم یا نظام اجتماعی) / پسوندواره فعلی	مشروطه‌خواه
دار	پسوندواره فعلی	امانتدار
گیر	پسوندواره فعلی	دندانگیر
زن	پسوندواره فعلی	چشمک‌زن
کش	پسوندواره فعلی	زجرکش
بند	پسوندواره فعلی	ماست‌بند
خوار	پسوندواره فعلی	گوشتخوار
فروش	پسوندواره فعلی	دستفروش
گو	پسوندواره فعلی	حقیقتگو
پوش	پسوندواره فعلی	ژنده‌پوش
خوان	پسوندواره فعلی	آوازه‌خوان
آور	پسوندواره فعلی	سودآور
جو	پسوندواره فعلی	حقیقت‌جو
انداز	پسوندواره فعلی	چشم‌انداز
رو	پسوندواره فعلی	تندرو
کن	پسوندواره فعلی	خردکن
آمیز	پسوندواره فعلی	تمسخرآمیز
کوب	پسوندواره فعلی	دیوارکوب
انگیز	پسوندواره فعلی	رقت‌انگیز
تاب	پسوندواره فعلی	بازتاب
چین	پسوندواره فعلی	دورچین

نویس	پسوندواره فعلی	خوشنویس
پرور	پسوندواره فعلی	شهیدپرور
شکن	پسوندواره فعلی	دندان‌شکن
بخش	پسوندواره فعلی	فرح‌بخش
پیچ	پسوندواره فعلی	طناب‌پیچ
افکن	پسوندواره فعلی	مردافکن
ریز	پسوندواره فعلی	سرریز
آرا	پسوندواره فعلی	زینت‌آرا
بر	پسوندواره فعلی	بیرون‌بر
گداز	پسوندواره فعلی	جانگداز
بار	پسوندواره فعلی	سبک‌بار
نواز	پسوندواره فعلی	گوش‌نواز
پرواز	پسوندواره فعلی	بلندپرواز
نشین	پسوندواره فعلی	خوش‌نشین
افشان	پسوندواره فعلی	زرافشان
کش	پسوندواره فعلی	دخترکش
گرد	پسوندواره فعلی	بیابان‌گرد
گذار	پسوندواره فعلی	خدمتگذار
تراش	پسوندواره فعلی	پیکرتراش
جوش	پسوندواره فعلی	خودجوش
کن	پسوندواره فعلی	پوست‌کن
مال	پسوندواره فعلی	لگدمال
یاب	پسوندواره فعلی	طلایاب
اندیش	پسوندواره فعلی	دوراندیش
زا	پسوندواره فعلی	بیماری‌زا
شمار	پسوندواره فعلی	گاه‌شمار
افروز	پسوندواره فعلی	جنگ‌افروز
پر	پسوندواره فعلی	؟
گشا	پسوندواره فعلی	مشکل‌گشا
رس	پسوندواره فعلی	فریادرس

گستر	پسوندواره فعلی	مهرگستر
خواب	پسوندواره فعلی	خوش‌خواب
نورد	پسوندواره فعلی	دریانورد
پسند	پسوندواره فعلی	مشکل‌پسند
دزد	پسوندواره فعلی	بچه‌دزد
ران	پسوندواره فعلی	لوکوموتیوران
فریب	پسوندواره فعلی	دلفریب
نام	پسوندواره فعلی	خوش‌نام
نگار	پسوندواره فعلی	چهره‌نگار
افزا	پسوندواره فعلی	روح‌افزا
آموز	پسوندواره فعلی	دست‌آموز
پاش	پسوندواره فعلی	نمک‌پاش
بو	پسوندواره فعلی	خوش‌بو
دوش	پسوندواره فعلی	ماردوش
شو	پسوندواره فعلی	تاشو
آفرین	پسوندواره فعلی	جان‌آفرین
تراز	پسوندواره فعلی	هم‌تراز
خند	پسوندواره فعلی	پوزخند
سرا	پسوندواره فعلی	ترانه‌سرا
آویز	پسوندواره فعلی	رخت‌آویز
دو	پسوندواره فعلی	سگ‌دو
چران	پسوندواره فعلی	گاوچران
نوش	پسوندواره فعلی	دردنوش
ربا	پسوندواره فعلی	آهن‌ربا
پیوند	پسوندواره فعلی	؟
تاز	پسوندواره فعلی	یکه‌تاز
رسان	پسوندواره فعلی	پیام‌رسان
روب	پسوندواره فعلی	خاک‌روب
انجام	پسوندواره فعلی	سرانجام
پیرا	پسوندواره فعلی	چمن‌پیرا

خا	پسوندواره فعلی	شکرخا
دم	پسوندواره فعلی	آتشین‌دم
گزین	پسوندواره فعلی	خلوت‌گزین
آزار	پسوندواره فعلی	مردم‌آزار
بیز	پسوندواره فعلی	آردبیز
خر	پسوندواره فعلی	بزخر
فرسا	پسوندواره فعلی	طاقت‌فرسا
ورز	پسوندواره فعلی	دادورز
آگند	پسوندواره فعلی	کج‌آگند
باش	پسوندواره فعلی	شادباش
ترس	پسوندواره فعلی	خداترس
چسب	پسوندواره فعلی	دلچسب
اندوز	پسوندواره فعلی	مال‌اندوز
چرخ	پسوندواره فعلی	تک‌چرخ
زدا	پسوندواره فعلی	گندزدا
گسار	پسوندواره فعلی	غم‌گسار
توان	پسوندواره فعلی	کم‌توان
چر	پسوندواره فعلی	علف‌چر
آزما	پسوندواره فعلی	بخت‌آزما
آشوب	پسوندواره فعلی	دل‌آشوب
پران	پسوندواره فعلی	مگس‌پران
توز	پسوندواره فعلی	کینه‌توز
خراش	پسوندواره فعلی	گوش‌خراش
شتاب	پسوندواره فعلی	؟
گوار	پسوندواره فعلی	خوش‌گوار
آشام	پسوندواره فعلی	خون‌آشام
ی	قیدساز	آخر عمری/آمدنی/برگشتنی
اً	قیدساز	اخیراً/دائماً
ک	قیدساز	کم‌کمک/نم‌نمک
کی	قیدساز	یواشکی/هول‌هولکی

ه	قیدساز	امروزه
ان	قیدساز	سحرگاهان
انه	قیدساز	پیروزمندانه
وار	قیدساز	دیوانه‌وار
بر	پیشوندی که در معنی فعل تغییر ایجاد می‌کند	برداشت
در	پیشوندی که در معنی فعل تغییر ایجاد می‌کند	درافتاد
باز	پیشوندی که در معنی فعل تغییر ایجاد می‌کند	بازداشت
فرا	پیشوندی که در معنی فعل تغییر ایجاد می‌کند	فراگرفت
فرو	پیشوندی که در معنی فعل تغییر ایجاد می‌کند	فرورفت

٤-٤- جزئیات چگونگی نرمال‌سازی و برچسب‌گذاری ماشینی داده‌های پیکره

برای اجرای کد برچسب‌گذاری ابزار «هضم»، ابتدا لازم است جدولی در پایگاه داده‌ای mysql ایجاد شود. به منظور ایجاد این جدول اسکریپ زیر باید اجرا شود. با اجرای این کد در پایگاه داده، جدولی با ساختار جدول ۴-۳ ایجاد خواهد شد. از این جدول برای نگهداری داده‌های برچسب‌گذاری‌شده استفاده می‌شود.

```
SET SQL_MODE = "NO_AUTO_VALUE_ON_ZERO";
SET time_zone = "+00:00";


/*!40101 SET
@OLD_CHARACTER_SET_CLIENT=@@CHARACTER_SET_CLIENT */;
/*!40101 SET
@OLD_CHARACTER_SET_RESULTS=@@CHARACTER_SET_RESULTS */;
/*!40101 SET
@OLD_COLLATION_CONNECTION=@@COLLATION_CONNECTION */;
```

```
/*!40101 SET NAMES utf8mb4 */;
--
-- Database: `corpus`
--
CREATE TABLE IF NOT EXISTS `tags` (
  `id` bigint(20) NOT NULL AUTO_INCREMENT,
  `document_id` int(11) NOT NULL,
  `word` varchar(500) COLLATE utf8_unicode_ci NOT NULL,
  `tag` varchar(20) COLLATE utf8_unicode_ci NOT NULL,
  PRIMARY KEY (`id`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8 COLLATE=utf8_unicode_ci;

/*!40101 SET CHARACTER_SET_CLIENT=@OLD_CHARACTER_SET_CLIENT
*/;
/*!40101 SET
CHARACTER_SET_RESULTS=@OLD_CHARACTER_SET_RESULTS */;
/*!40101 SET COLLATION_CONNECTION=@OLD_COLLATION_CONNECTION
*/;
```

جدول ۴-۳: ساختار جدول برای نگهداری کلمات و برچسب‌ها

شماره	نام فیلد	توضیحات
۱	id	کلید اصلی این جدل است و دارای اعداد ترتیبی
۲	document_id	دارای کلید خارجی به شماره سندی است که برای برچسب‌گذاری استفاده می‌شود.
۳	word	برای نگهداری کلمات از این فیلد استفاده می‌شود.
۴	tag	برای برچسب هر کلمه از این فیلد استفاده می‌شود.

پس از ایجاد جدول، از کد زیر که به زبان پایتون و سازگار با نسخه python 3.x است، برای برچسب-گذاری اسناد می‌توان استفاده کرد.

```
from __future__ import unicode_literals
from hazm import *
import mysql.connector
```

این فایل شامل مدل آموزش‌دیده برای برچسب‌گذاری است که توسط «هضم» توسعه داده شده است.

```
tagger = POSTagger(model='./resources-0.5/postagger.model')

db = mysql.connector.connect(user='user', password='password',
                             host='host_name_or_ip',
                             database='name_of_database')

cursor = db.cursor()
cursor.execute("SELECT * FROM hazm")

for document in cursor.fetchall():
    document_id = document[0]
    document_text = document[1]
    print(document_id)
    tags = tagger.tag(word_tokenize(document_text))
    c = 0
    for tag in tags:
        values = (document_id, tag[0], tag[1])
        cursor.execute("INSERT INTO `tags` (`id`, `document_id`, `word`, `tag`)
VALUES (NULL, %s, %s, %s)", values)
        c += 1
        if c > 1000:
            c = 0
            db.commit()
```

```
db.close()
```

قطع کد زیر توسعه داده شده توسط «هضم» برای برچسب‌گذاری یک رشته (string) استفاده می‌شود.

```
# coding: utf-8

from __future__ import unicode_literals
from nltk.tag import stanford
from .SequenceTagger import SequenceTagger

class POSTagger(SequenceTagger):
	"""
	>>> tagger = POSTagger(model='resources/postagger.model')
	>>> tagger.tag(['من', 'به', 'مدرسه', 'رفته_بودم', '.'])
	[('من', 'PRO'), ('به', 'P'), ('مدرسه', 'N'), ('رفته_بودم', 'V'), ('.', 'PUNC')]
	"""

class StanfordPOSTagger(stanford.StanfordPOSTagger):
	"""
	>>> tagger = StanfordPOSTagger(model_filename='resources/persian.tagger',
path_to_jar='resources/stanford-postagger.jar')
	>>> tagger.tag(['من', 'به', 'مدرسه', 'رفته_بودم', '.'])
	[('من', 'PRO'), ('به', 'P'), ('مدرسه', 'N'), ('رفته_بودم', 'V'), ('.', 'PUNC')]
	"""

	def __init__(self, model_filename, path_to_jar, *args, **kwargs):
		self._SEPARATOR = '/'
```

```
            super(stanford.StanfordPOSTagger,
self).__init__(model_filename=model_filename, path_to_jar=path_to_jar, *args,
**kwargs)

    def tag(self, tokens):
        return self.tag_sents([tokens])[0]

    def tag_sents(self, sentences):
        refined = map(lambda s: [w.replace(' ', '_') for w in s], sentences)
        return super(stanford.StanfordPOSTagger, self).tag_sents(refined)
```

قطع کد زیر توسعه داده شده توسط «هضم» برای نرمال‌سازی یک رشته (string) استفاده می‌شود.

```
# coding: utf-8

from __future__ import unicode_literals
import re
from .Lemmatizer import Lemmatizer
from .WordTokenizer import WordTokenizer
from .utils import maketrans

compile_patterns = lambda patterns: [(re.compile(pattern), repl) for pattern, repl in
patterns]

class Normalizer(object):
    def __init__(self, remove_extra_spaces=True, persian_style=True,
persian_numbers=True, remove_diacritics=True, affix_spacing=True,
token_based=False, punctuation_spacing=True):
        self._punctuation_spacing = punctuation_spacing
        self._affix_spacing = affix_spacing
        self._token_based = token_based
```

```
translation_src, translation_dst = ' یکی ','"“ی‌کی"'''
if persian_numbers:
        translation_src += '0123456789%'
        translation_dst += '٪۰۱۲۳۴۵۶۷۸۹'
self.translations = maketrans(translation_src, translation_dst)

if self._token_based:
        lemmatizer = Lemmatizer()
        self.words = lemmatizer.words
        self.verbs = lemmatizer.verbs
        self.tokenizer = WordTokenizer(join_verb_parts=False)
        self.suffixes = {'ی', 'ای', 'ها', 'های', 'تر', 'تری', 'ترین', 'گر', 'گری', 'ام', 'ات', '
اش"}

self.character_refinement_patterns = []

if remove_extra_spaces:
        self.character_refinement_patterns.extend([
                (r' +', ' '),  # remove extra spaces
                (r'\n\n+', '\n\n'),  # remove extra newlines
                (r'[‌\r]', ''),  # remove keshide, carriage returns
        ])

if persian_style:
        self.character_refinement_patterns.extend([
                ('"([^\n"]+)"', r'«\1»'),  # replace quotation with gyoome
                ('([\d+])\.([\d+])', r'\1٫\2'),  # replace dot with momayez
                (r' ?\.\.\.', ' …'),  # replace 3 dots
        ])
```

```
        if remove_diacritics:
            self.character_refinement_patterns.append(
('[\u064B\u064C\u064D\u064E\u064F\u0650\u0651\u0652]', ''),  # remove FATHATAN, DAMMATAN, KASRATAN, FATHA, DAMMA, KASRA, SHADDA, SUKUN
            )

        self.character_refinement_patterns = compile_patterns(self.character_refinement_patterns)

        punc_after, punc_before = r'\.:!؟؛،»\]\)\}', r'«\[\(\{'
        if punctuation_spacing:
            self.punctuation_spacing_patterns = compile_patterns([
                ('" ([^\n"]+) "', r'"\1"'),  # remove space before and after quotation
                (' (['+ punc_after +'])', r'\1'),  # remove space before
                (' (['+ punc_before +']) ', r'\1'),  # remove space after
                ('(['+ punc_after[:3] +'])([^ '+ punc_after +'\d۰۱۲۳۴۵۶۷۸۹])', r'\1 \2'),  # put space after . and :
                ('(['+ punc_after[3:] +'])([^ '+ punc_after +'])', r'\1 \2'),  # put space after
                ('([^ '+ punc_before +'])(['+ punc_before +'])', r'\1 \2'),  # put space before
            ])

        if affix_spacing:
            self.affix_spacing_patterns = compile_patterns([
                (r'([^ ]ه) ی ', r'\1ی '),  # fix ی space
                (r'(^| )(نمی?) ', r'\1\2'),  # put zwnj after نمی, می
                (r'(?<=[^\n\d '+ punc_after + punc_before +']{2}) (های|گری?|ترین?|تر)(?=[ \n'+ punc_after + punc_before +']|$)', r'\1'),  # put zwnj before تر, گری, گر, ترین, تری, های, ها
                (r'([^ ]ه) ((ا|م|ایم|اش|اند|ای|اید|ات))(?=[ \n'+ punc_after +']|$)', r'\1\2'),  # join ام, ایم, اش, اند, ای, اید, ات
```

```
                ])

        def normalize(self, text):
                text = self.character_refinement(text)
                if self._affix_spacing:
                        text = self.affix_spacing(text)

                if self._token_based:
                        tokens = self.tokenizer.tokenize(text.translate(self.translations))
                        text = ' '.join(self.token_spacing(tokens))

                if self._punctuation_spacing:
                        text = self.punctuation_spacing(text)

                return text

        def character_refinement(self, text):
                """
                >>> normalizer = Normalizer()
                >>> normalizer.character_refinement('اصلاح کاف و یای عربی')
                'اصلاح کاف و یای عربی'
                >>> normalizer.character_refinement('4.2 قراردادی به ارزش "2012عراق سال
میلیارد دلار" برای خرید تجهیزات نظامی با روسیه امضا کرد.')
                'عراق سال ۲۰۱۲ قراردادی به ارزش «۴٫۲ میلیارد دلار» برای خرید تجهیزات نظامی با روسیه امضا
کرد.'
                >>> normalizer.character_refinement('رمـــان')
                'رمان'
                >>> normalizer.character_refinement('بُشقابِ مَن را بِگیر')
                'بشقاب من را بگیر'
                """

                text = text.translate(self.translations)
```

```
        for pattern, repl in self.character_refinement_patterns:
            text = pattern.sub(repl, text)
        return text

    def punctuation_spacing(self, text):
        """
        >>> normalizer = Normalizer()
        >>> normalizer.punctuation_spacing('اصلاح ( پرانتزها ) در متن .')
        'اصلاح (پرانتزها) در متن.'
        >>> normalizer.punctuation_spacing('1396 تهران، 22:00 در ساعت 0.5نسخه ')
        '1396 تهران، 22:00 در ساعت 0.5نسخه '
        >>> normalizer.punctuation_spacing('۷ میلیون.اتریش ۹.')
        '۷ میلیون.اتریش ۹.'

        """

        for pattern, repl in self.punctuation_spacing_patterns:
            text = pattern.sub(repl, text)
        return text

    def affix_spacing(self, text):
        """
        >>> normalizer = Normalizer()
        >>> normalizer.affix_spacing('خانه ی پدری')
        'خانه‌ی پدری'
        >>> normalizer.affix_spacing('فاصله میان پیشوند ها و پسوند ها را اصلاح می کند.')
        'فاصله میان پیشوندها و پسوندها را اصلاح می‌کند.'
        >>> normalizer.affix_spacing('می روم')
        'می‌روم'
        >>> normalizer.affix_spacing('حرفه ای')
        'حرفه‌ای'
```

```
            >>> normalizer.affix_spacing('محبوب ترین ها')
            'محبوب‌ترین‌ها'
            """

            for pattern, repl in self.affix_spacing_patterns:
                text = pattern.sub(repl, text)
            return text

    def token_spacing(self, tokens):
            """
            >>> normalizer = Normalizer(token_based=True)
            >>> normalizer.token_spacing(['کتاب', 'ها'])
            ['کتاب‌ها']
            >>> normalizer.token_spacing(['او', 'می', 'رود'])
            ['او', 'می‌رود']
            >>> normalizer.token_spacing(['ماه', 'می', 'سال', 'جدید'])
            ['ماه', 'می', 'سال', 'جدید']
            >>> normalizer.token_spacing(['اخلال', 'گر'])
            ['اخلال‌گر']
            >>> normalizer.token_spacing(['پرداخت', 'شده', 'است'])
            ['پرداخت', 'شده', 'است']
            >>> normalizer.token_spacing(['زمین', 'لرزه', 'ای'])
            ['زمین‌لرزه‌ای']
            """

            result = []
            for t, token in enumerate(tokens):
                joined = False

                if result:
                    token_pair = result[-1]+"+token
```

```
                    if token_pair in self.verbs or token_pair in self.words and
self.words[token_pair][0] > 0:
                        joined = True

                    if t < len(tokens)-1 and token+'_'+tokens[t+1] in
self.verbs:
                        joined = False

                elif token in self.suffixes and result[-1] in self.words:
                    joined = True

            if joined:
                result.pop()
                result.append(token_pair)
            else:
                result.append(token)

        return result
```

٤-٥- کنترل دستی برچسب‌ها

کنترل دستی برچسب‌ها به این صورت انجام گرفت که ابتدا متن ۵۰ مقاله به عنوان نمونه انتخاب شد و برچسب‌ها به دقت مورد مطالعه قرار گرفت و به اصلاح برچسب‌های غلط پرداخته شد. به منظور اصلاح برچسب‌های غلط در پیکره، امکان ویرایش در صفحهٔ هر مقاله فراهم شد، به این صورت که یکی‌یکی روی هر واژه‌ای که برچسب غلط داشت، کلیک می‌شد و پنجره‌ای باز می‌شد که در آن امکان اصلاح املایی واژه (اگر نیاز داشت) و برچسب آن فراهم می‌شد(شکل ۴-۱).

شکل ۴-۱: پنجره اصلاح برچسب

ضمن اصلاح برچسب‌های غلط، به بررسی خطای موجود در برچسب‌گذاری نیز پرداخته شد تا اگر الگویی برای خطاها مشاهده می‌شود، بتوان به صورت ماشینی اصلاح کرد. این خطاها را می‌توان در دسته-بندی زیر قرار داد:

- در بسیاری موارد واژهٔ «می‌توان» و «نمی‌توان» به دلیل ختم شدن به «ن» که نشانهٔ مصدر/ اسم است، یا به دلیل اینکه نویسنده بین پیشوند تصریفی «می-/نمی-» و «توان»، فاصله گذاشته است، به اشتباه برچسب اسم(N) دارد.
- در بسیاری موارد پسوند جمع «-ها»، پسوند جمع «-ها» همراه با کسرهٔ اضافه («-های») و پسوند جمع «-ها» همراه با کسرهٔ اضافه و «-ی» نکره («هایی»)، با فاصله به تکواژ قبل از خود وصل شده است (مانند «کتاب ها، کتاب های، کتاب هایی) و این مسئله باعث شده است این واژه‌ها دو برچسب داشته باشند (به عنوان نمونه واژهٔ «کتاب» برچسب اسم N/ و پسوند تصریفی «ها» برچسب اسم N/ یا صفت/ADJ).
- در برخی موارد «-ای» از تکواژ قبل از خود جدا افتاده است (مانند «کتابخانه ای»).
- در بسیاری موارد واژه‌هایی که به پسوندهای «-ساز/سازی»(مانند «نمایه‌سازی»)، «-رسان/رسانی»(مانند «اطلاع‌رسانی»)، «-گیر/گیری»(مانند «نتیجه‌گیری»)، «-دار/داری»(مانند «معنی-

داری»)، «–شناس/شناسی»(مانند «زبان‌شناسی»)، «–سنج/سنجی»(مانند «علم‌سنجی»)، «–گذار/گذاری»(مانند «برچسب‌گذاری»)، «–پذیر/پذیری»(مانند «ریسک‌پذیری»)، «–بند/بندی»(مانند «طبقه‌بندی»)، «–افزار/افزاری» (مانند «نرم‌افزار»، «–نویس/نویسی»(مانند «برنامه‌نویسی»)، «–یاب/یابی»(مانند «دستیابی»)، «–آور/آوری»(مانند «جمع‌آوری»)، «–جو/جویی(مانند «جست‌وجو» و....» ختم می‌شوند، به دلیل فاصله‌ای که نویسنده بین تکواژها گذاشته است، به اشتباه دو برچسب (یا بیشتر) گرفته‌اند و پسوندها برچسب اسم(N) یا فعل(V) ممکن است گرفته باشند.

- واژهٔ «داده» به دلیل حوزهٔ موضوعی مجلهٔ پردازش و مدیریت اطلاعات، بسیار در متون گوناگون تکرار شده است و در برخی موارد، به جای اینکه برچسب «اسم (N)» بگیرد، برچسب «فعل»(V) گرفته است.

در ادامه، به منظور تسریع فرآیند کنترل دستی برچسب‌ها، برخی از داده‌ها را که به دلیل عدم رعایت نیم‌فاصله، به اشتباه دو برچسب داشتند، به صورت ماشینی اصلاح شدند؛ به این صورت:

- هر جا تکواژ «–ها»، «–های» و «–هایی» جدا از تکواژ قبل است و یک واژهٔ مجزا در نظر گرفته شده است، با نیم‌فاصله به واژه/تکواژ قبل از آن متصل گردد.
- هر جا تکواژهای «می–» و «نمی–» جدا از تکواژ بعد است و یک واژهٔ مجزا در نظر گرفته شده است، با نیم‌فاصله به واژه/تکواژ بعد از آن متصل گردد.
- هر جا تکواژ «–تر»، «–ترین» جدا از تکواژ قبل است و یک واژهٔ مجزا در نظر گرفته شده است، با نیم‌فاصله به واژه/تکواژ قبل از آن متصل گردد.
- هر جا تکواژ «–ای» جدا از تکواژ قبل است و یک واژهٔ مجزا در نظر گرفته شده است، با نیم‌فاصله به واژه/تکواژ قبل از آن متصل گردد.
- هر جا تکواژ «–ی» جدا از تکواژ قبل است و یک واژهٔ مجزا در نظر گرفته شده است، با نیم‌فاصله به واژه/تکواژ قبل از آن متصل گردد.
- هر جا تکواژهای «–ساز/سازی»، «–گیر/گیری»، «–شناس/شناسی»، «–بند/بندی»، «–رسان/رسانی»، «–سنج/سنجی»، «–گذار/گذاری»، «–پذیر/پذیری»، «–نویس/نویسی»، «–افزار/افزاری»، «–یاب/یابی»، «–دار/داری»، «–آور/آوری»، «–جو/جویی»، جدا از تکواژ قبل است و یک واژهٔ مجزا در نظر گرفته شده است، با نیم‌فاصله به واژه/تکواژ قبل از آن متصل گردد.

پس از این مرحله، متون مقاله‌های ویرایش شده، مجدداً به ابزار «هضم» داده شد تا برچسب‌گذاری مجدد انجام شود و در نهایت، برچسب‌ها مجدداً دستی کنترل شدند. به این ترتیب تا حد امکان برچسب-های نادرست، اصلاح شدند. البته کماکان تقریباً دو درصد خطا وجود دارد. به عنوان نمونه، متن برچسب-گذاری‌شدهٔ زیر نشان می‌دهد، در هر ۳۰۰ واژه، ممکن است، چهار یا پنج خطای برچسب‌گذاری وجود داشته باشد.

٤-٦- خلاصه فصل

در این فصل ابتدا به تعریف «نرمال‌سازی» و چگونگی نرمال‌سازی داده‌های واردشده در پیکره پرداخته شد. سپس مفهوم «حاشیه‌نویسی» توضیح داده شد و در نهایت، نوع حاشیه‌نویسی/برچسب‌گذاری استفاده شده در پیکرهٔ متنی ایرانداک و چگونگی ابزارهای استفاده شده برای این منظور، معرفی شد و در نهایت نحوهٔ کنترل دستی برچسب‌ها نیز توضیح داده شد.

فصل پنجم

استقرار سامانهٔ پیکره‌های ایرانداک

۵-۱- استقرار سامانهٔ پیکره‌های ایرانداک

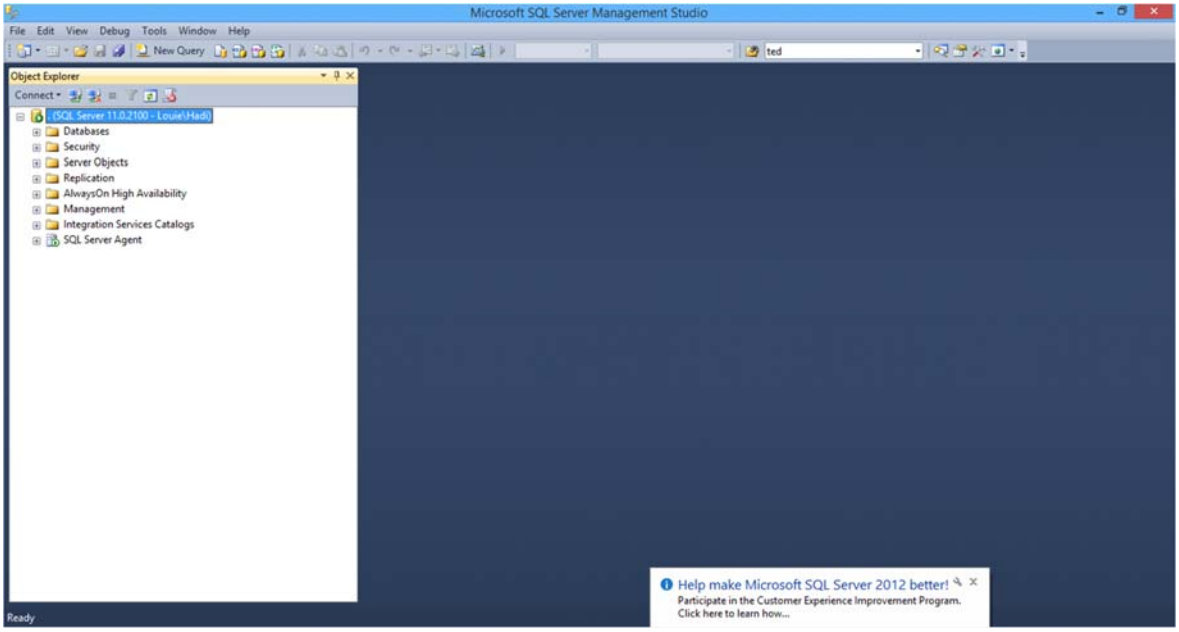
تمام تنظیمات سامانه در فایل App.Settings قرار دارد. این تنظیمات شامل کلیدهای مورد استفاده در سامانه، تنظیمات پایگاه داده، و ... است.

مسیر این فایل در ریشه فولدر اصلی سامانه با نام App.Settings قرار دارد.

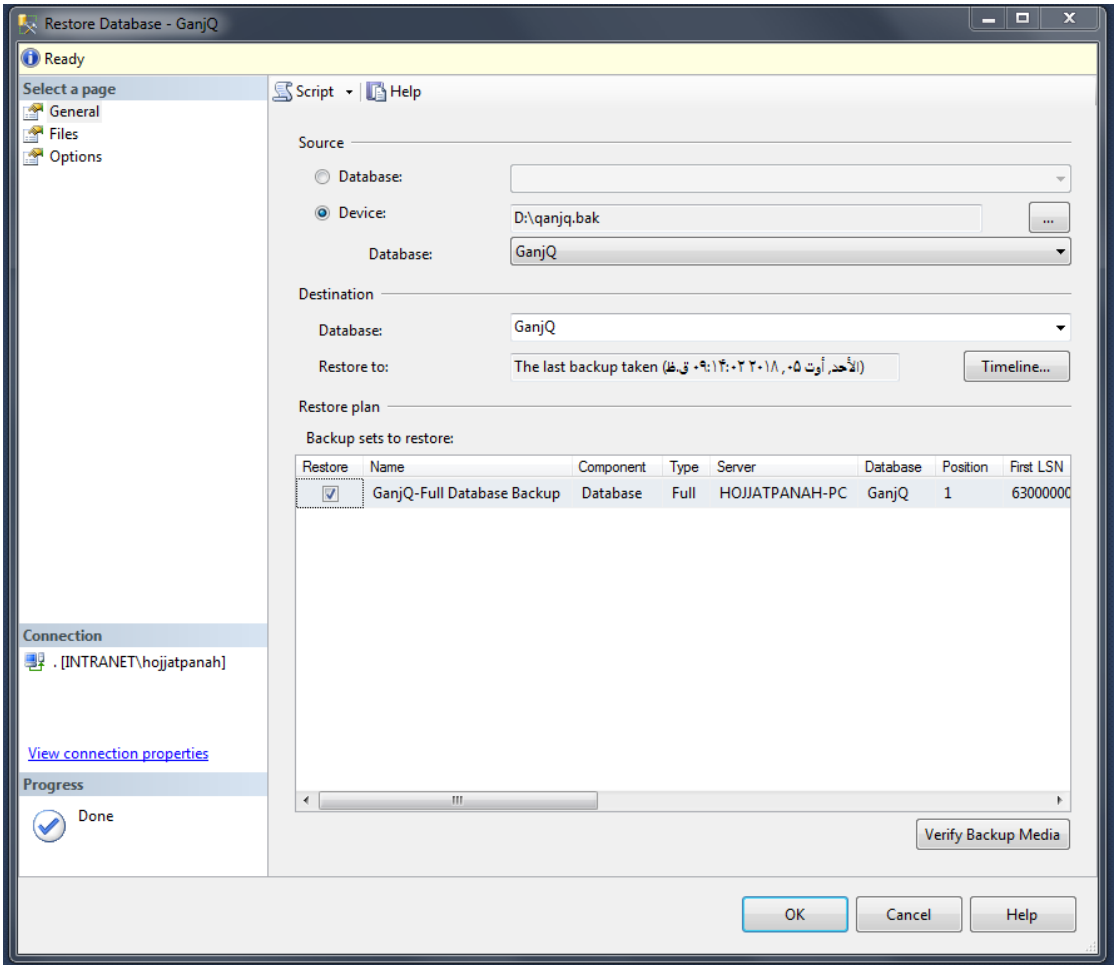
۵-۲-پایگاه داده سامانه

این سامانه از Sql Server 2019 به عنوان موتور پایگاه داده استفاده می‌کند. برای راه‌اندازی پایگاه داده کافیست از فایل پشتیبانی که در پیوست آمده است استفاده و مراحل زیر را طی کرد.

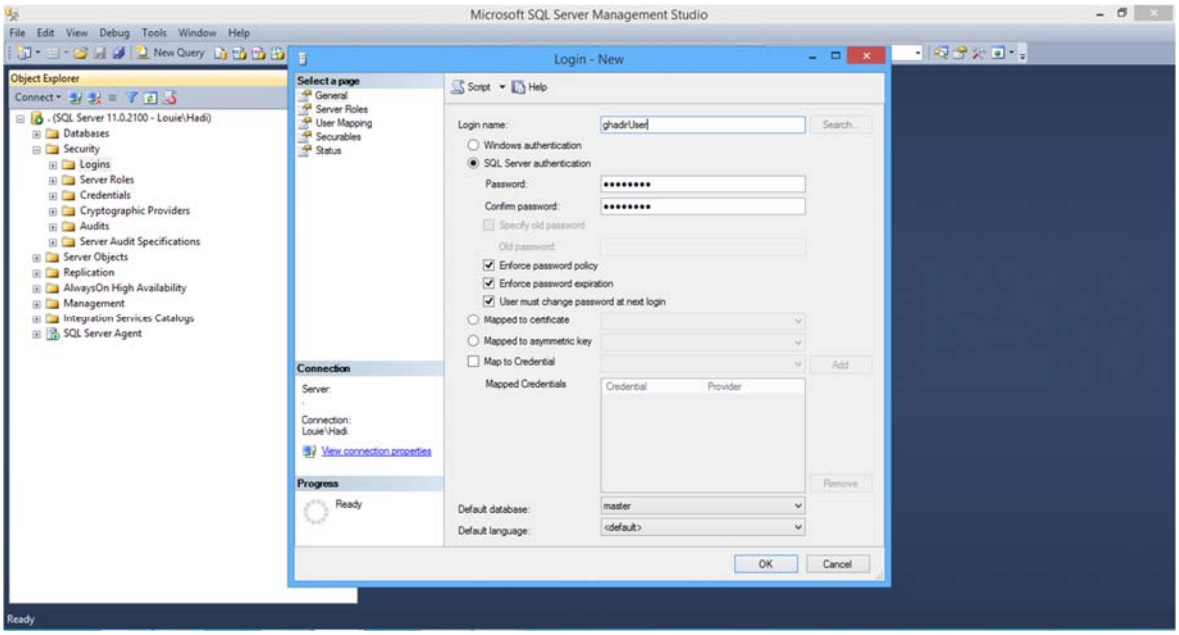
۱. باز کردن برنامه Sql Server Management Studio



۲. کلیک سمت راست بروی Databases و زدن گزینهٔ Restore Databases


Restore Database - GanjQ
Ready
Select a page
General
Files
Options
Script
Help
Source
Database:
Device:
D:\qanjq.bak
Database:
GanjQ
Destination
Database:
GanjQ
Restore to:
The last backup taken (الأحد, أوت ٠٥, ٢٠١٨ ٠٩:١٢:٠٢ ق.ظ)
Timeline...
Restore plan
Backup sets to restore:
Restore
Name
Component
Type
Server
Database
Position
First LSN
GanjQ-Full Database Backup
Database
Full
HOJJATPANAH-PC
GanjQ
1
Connection
. [INTRANET\hojjatpanah]
View connection properties
Progress
Done
Verify Backup Media
OK
Cancel
Help

۳. ساختن کاربر اتصال به پایگاه داده از طریق زدن گزینه Security و سپس Logins و در نهایت New Login



توجه کنید که در قسمت Server Roles باید دسترسی‌های متناسب با موتور پایگاه داده و سروری که سامانه در آن قرار گیرد، داده شود.

پس از Restore کردن پایگاه داده کافیست در فایل Web.Config سامانه که توضیح آن در قسمت قبل داده شد، تنظیم شود.

در فایل Web.Config بخشی با عنوان ConnectionString وجود دارد که وظیفه تنظیم نحوه اتصال به پایگاه داده را بر عهده دارد. این تنظیمات بر اساس Restore کردن پایگاه داده و کاربری که تعریف شده است، امکان اتصال سامانه به پایگاه داده را فراهم می‌کند. این تنظیم برای نمونه در زیر آمده است:

```
<connectionStrings>
    <add name="ProjectContext" connectionString="data source=.;initial
    catalog=Corpus;  User  ID=QUser;  password=****;
    App=EntityFramework;  MultipleActiveResultSets=true"
    providerName="System.Data.SqlClient" />
</connectionStrings>
```

۵-۳- نصب Net Core Framework. نسخهٔ 3.1.1

به طور کلی Net Core. مجموعه‌ای از فایل‌های مورد نیاز سیستم عامل (شامل فایل های DLL و رجیستری و واسطه‌های استاندارد ارتباط برنامه‌ها با یکدیگر) است که برای اجرای برنامه‌های نوشته شده تحت Net Core. ضروری می‌باشد. یعنی یک پکیج کامل از تمام dll های مورد نیاز برنامه‌هایی که با خود Net Core. نوشته شده‌اند.

برای نصب این Framework کافیست فایل Exe آن را اجرا کرده و مراحل نصب تا انتها انجام شود. پس از نصب تمام فایل‌های مورد نیاز برای استفاده در سامانه ایجاد می‌شود.

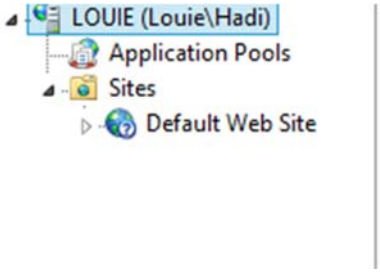
۵-٤- تنظیم وب سرور

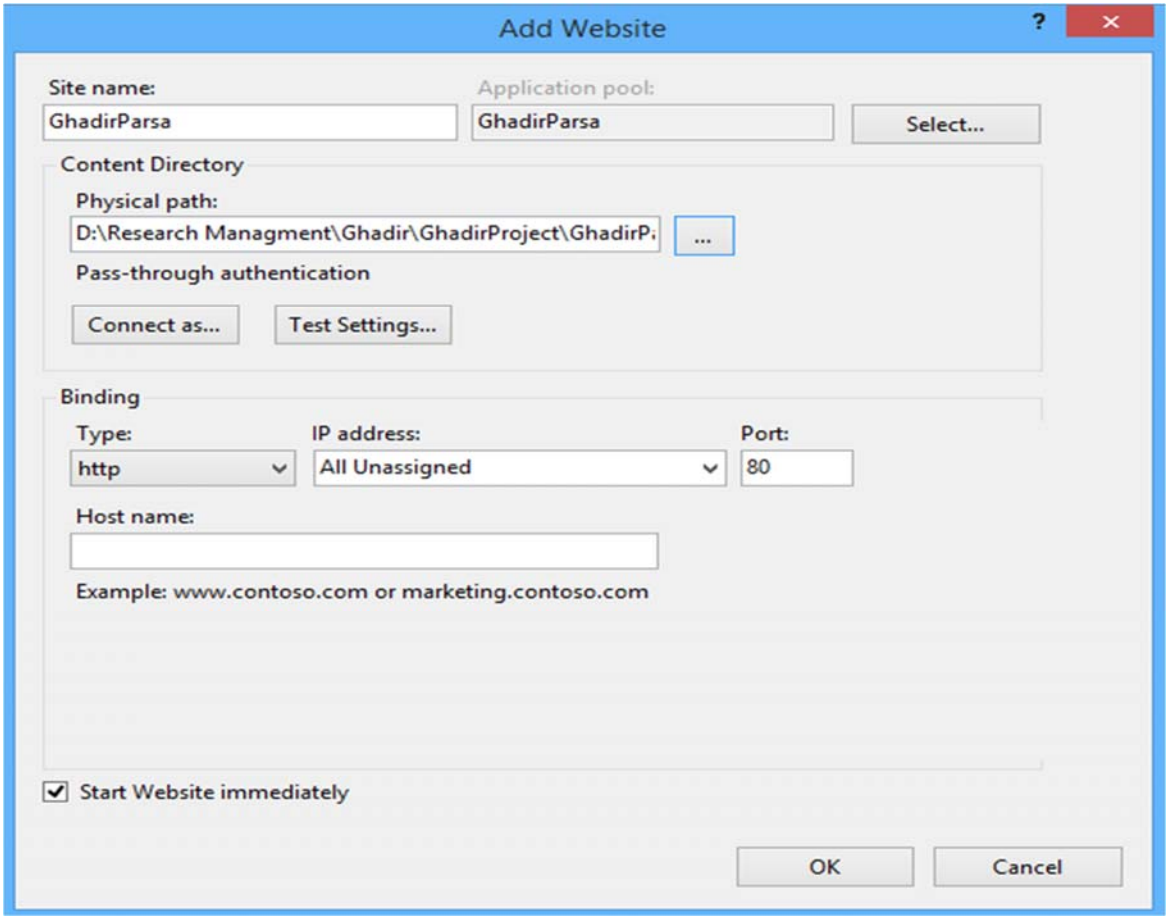
این سامانه بروی وب سرور IIS 8 قرار می‌گیرد. سرویس IIS[1] ، سرویس‌های اطلاعات اینترنتی است که توسط شرکت Microsoft عرضه شده و یک سرور برای کنترل کردن محتویات و دسترسی به سایت‌های وب یا FTP کاربر بر روی هارد ایجاد می‌کند .برای مثال، هنگامی که شما می‌خواهید سایتتان را منتشر کنید قبل از upload کردن آن می‌خواهید آن را آزمایش کنید و اگر با asp طراحی می‌کنید، قبل از نصب Visual Studio.Net ، بهتر است این سرویس را نصب کنید و گرنه مشکلاتی را برای شما به همراه خواهد داشت.

[1] Internet Information Services

برای گذاشتن سورس برنامه روی وب سرور کافیست مراحل زیر را انجام شود:

۱. رفتن به قسمت Control Panel
۲. انتخاب گزینه Administrator Tools
۳. انتخاب گزینه Internet Information Services (IIS) Manager
۴. کلیک راست روی Sites و انتخاب Add New website
۵. در این قسمت نام سایت مورد نظر وارد شده و بعد مسیر فیزیکی سورس برنامه انتخاب شودو در نهایت پورتی که سامانه قرار است بر روی آن بالا بیاید، انتخاب می شود.
۶. با زدن روی گزینه Ok سایت ایجاد می‌شود و می‌توان از طریق کلید Browse سایت را بر روی مرورگر مشاهده کرد.

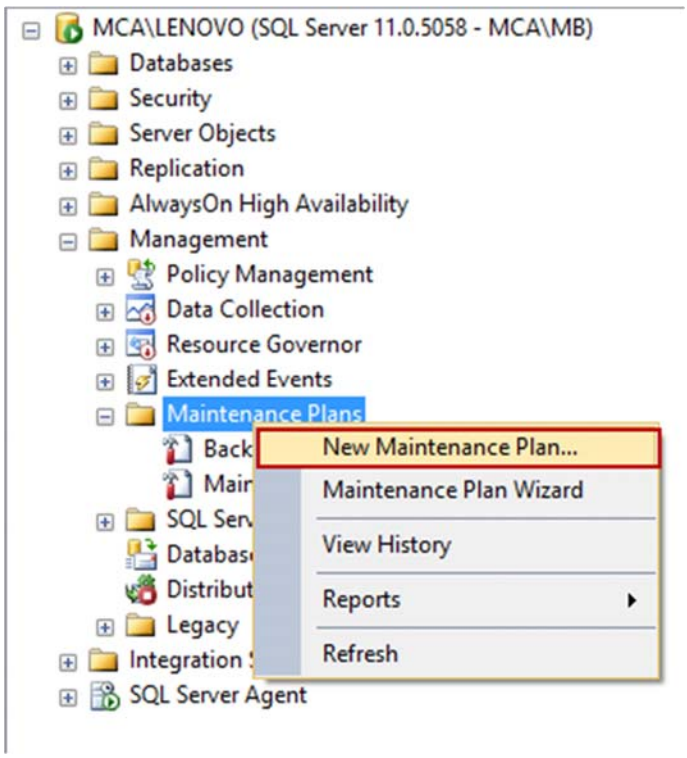




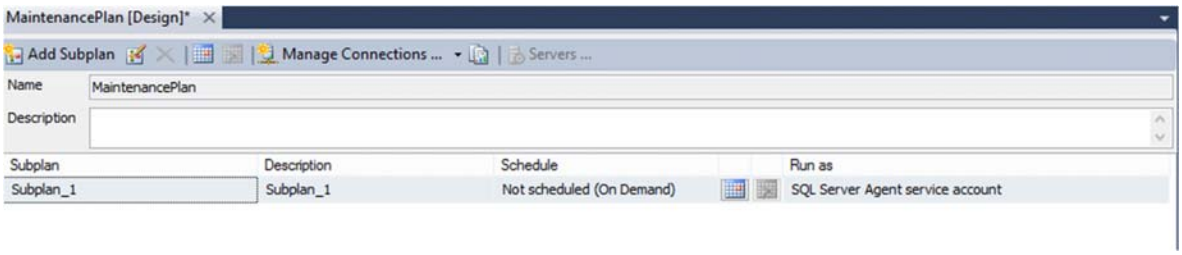
۵-۵- تنظیم پشتیبان‌گیری از سامانه

برای پشتیبان‌گیری خودکار پایگاه داده سامانه باید مراحل زیر را طی کرد.

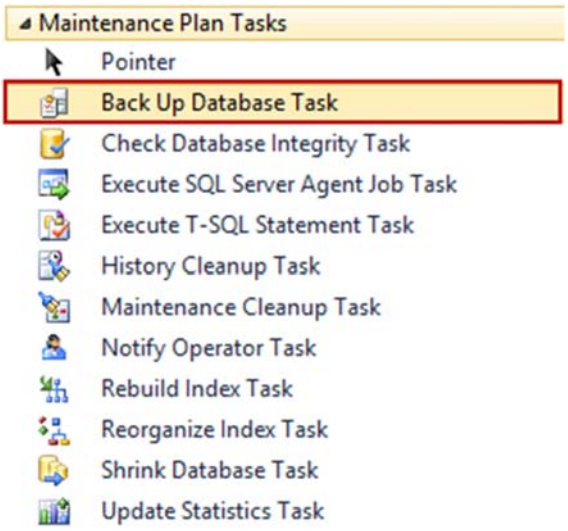
۱. پس از باز کردن Sql Server Management Studio و رفتن به قسمت Object Explore، بر روی گزینه Management کلیک کرده و گزینه Maintenace Plans را انتخاب می‌کنیم. پس از آن روی گزینه New Maintenace Plan کلیک می‌کنیم.



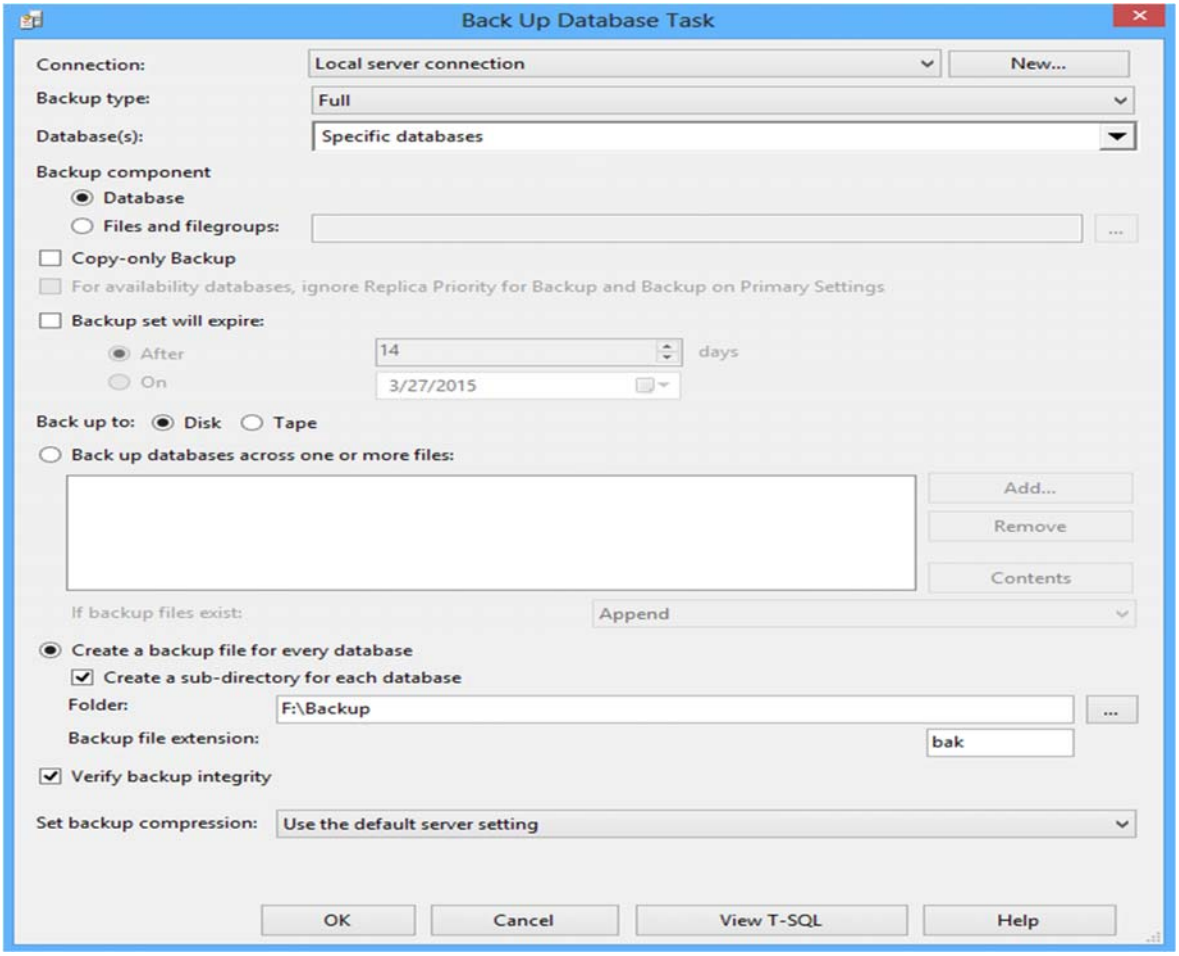
نرم‌افزار Sql Server بخشی را برای تنظیمات پشتیبان‌گیری به همراه جعبه ابزاری برای کنترل و زمان‌بندی در اختیار می‌گذارد.



۲. در بخش Maintenance Plan Tasks گزینه Backup Database Task را انتخاب می‌کنیم.



۳. دوباره بر روی گزینه آمده روی صفحه کلیک می‌کنیم و صفحه‌ای مشابه صفحه زیر ظاهر می‌گردد:



در این قسمت محلی که باید پشتیبان گرفته شود و زمان‌بندی‌های لازم انجام می‌شود. به عنوان مثال می‌توان تنظیم کرد هر هفته در یک ساعت خاص پشتیبان گرفته شود. لازم به ذکر است که امکان پاک کردن پشتیبان نیز قابل زمانبدی است.

۵-۶-نصب فونت گزارش‌های سامانه

گزارش‌های سامانه از فونت B Mitra استفاده می‌کنند. برای نصب فونت بروی سرور باید به بخش Control Panel رفته و پوشه Fonts انتخاب شود. و فونت مورد نظر در آن قسمت کپی و نصب شود. لازم به ذکر است برای اینکه فونت بروی گزارش‌های اعمال شود، نیاز است که سرور Restart شود.

فصل ششم

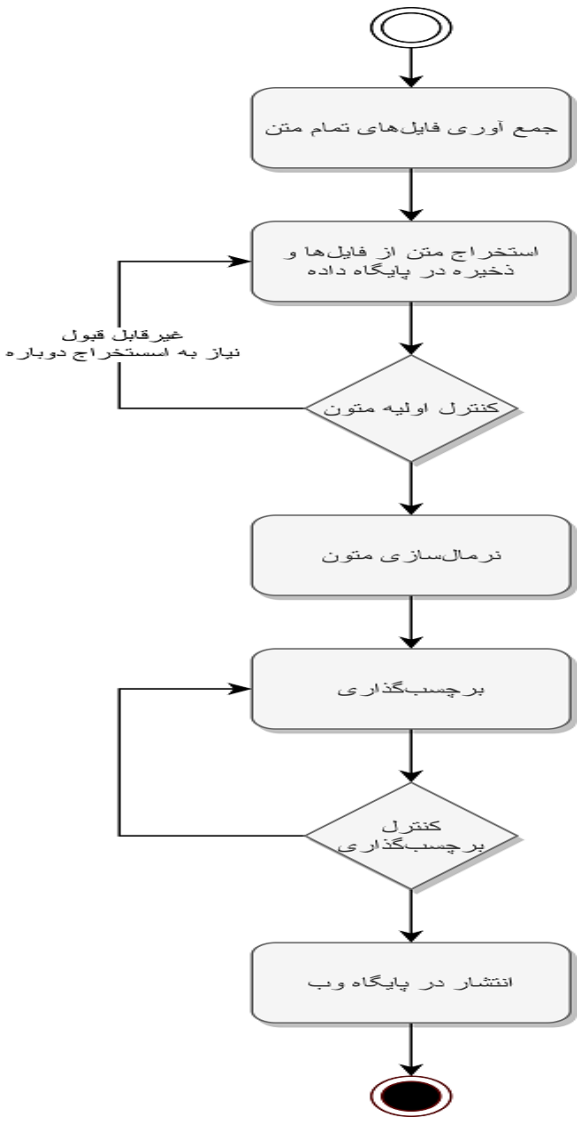
نتیجه‌گیری

٦-١- جمع‌بندی

پیکره، مجموعه‌ای نظام‌مند، رایانه‌ای شده، معتبر و موثق از زبان است که برای مطالعات زبان‌شناسی مورد استفاده قرار می‌گیرد. بسیاری از پژوهش‌های زبان‌شناسی و تصمیم‌گیری‌ها در برنامه‌ریزی زبانی، تنها با استفاده از یک پیکرهٔ زبانی امکان‌پذیر است. برخی از کاربردهای پیکره شامل: پردازش زبان طبیعی و درک و بازشناسی گفتار، تبدیل متن به گفتار و برعکس، تدوین فرهنگ‌ها، آموزش و پژوهش، ساخت پایگاه‌های دادهٔ زبانی، بررسی واژه‌های هم‌آیند در زبان‌های گوناگون، پایشگری زبان برای پیگیری و ردگیری دگرگونی‌های زبانی، ترجمهٔ ماشینی، توسعهٔ مفاهیم و منابع در پیوند با واژگان، نگارش و گسترش مهارت‌های نوشتاری، آموزش و یادگیری زبان با شناخت گویش‌ها و گوناگونی زبانی، معنا-شناسی، تحلیل کلام، زبان‌شناسی اجتماعی، زبان‌شناسی حقوقی، واکاوی ژانرهای ادبی و پژوهش‌های دستوری است. تهیهٔ یک پیکرهٔ زبانی مبتنی بر داده‌های علمی ایرانداک (مقاله‌های پژوهش‌نامهٔ پردازش و مدیریت اطلاعات) از دو نظر حائز اهمیت است: الف. ایرانداک را در شمار پژوهشگاه‌ها و دانشگاه‌هایی قرار می‌دهد که پیکرهٔ زبانی مفید (مبتنی بر داده‌های علمی معتبر) تهیه کرده‌اند . ب. می‌توان بسیاری از پژوهش‌های حوزهٔ پردازش متن (مانند انواع طبقه‌بندی داده‌ها، استخراج اطلاعات گوناگون از متن و غیره) با استفاده از این پیکره انجام داد. در این طرح پژوهشی پیکرهٔ متنی ساخته شد که محتوای آن متون مقاله-های موجود در «پژوهش‌نامهٔ پردازش و مدیریت اطلاعات» است. در مرحلهٔ نمونه‌گیری، متن ۶۷۷ مقاله که فایل ورد(word) آن‌ها موجود بود، وارد پیکره شد. موضوع این مقاله‌ها شامل کتابداری و اطلاع-

رسانی، علم اطلاعات و دانش‌شناسی، فناوری اطلاعات، مدیریت اطلاعات، مدیریت دانش، زبان‌شناسی، زبان‌شناسی رایانشی، اصطلاح‌شناسی، هستان‌شناسی و سایر حوزه‌های مرتبط با پردازش اطلاعات است. علاوه بر وارد کردن متن مقاله‌ها در پیکره، فرادادهٔ مربوط به هر مقاله (شامل عنوان مقاله، نام نویسنده/نویسندگان و موضوع مقاله) نیز از طریق کاربرگه ثبت مقالات وارد پیکره شد. پس از وارد کردن داده‌ها و فراداده‌ها در پیکرهٔ متنی ایرانداک، از طریق خروجی گرفتن از پایگاه داده SQLSERVER داده‌ها به برنامهٔ نرمال‌سازی و سپس، برچسب‌گذاری منتقل شد. نرمال‌سازی به صورت ماشینی انجام شد؛ در این راستا از کد ابزار «هضم» ، استفاده شد. از کدهای ابزار «هضم» که به رایگان در اختیار کاربران قرار گرفته است، کد واحدسازی/جداسازی (tokenizer) در پیکره استفاده شده است. ابزار واحدسازی/جداسازی، مرز واژه‌ها را در متون تشخیص می‌دهد و متن را به دنباله‌ای از واژه‌ها تبدیل می‌کند و آن را برای تحلیل‌های بعدی آماده می‌کند. بنابراین، بخشی از پیش‌پردازش متن، واحدسازی/جداسازی است. همچنین برای یکدست کردن انواع «ی» و «ک» و «کسره اضافه روی «ه» (مانند «هٔ») از دستورات TSQL در برنامه پایگاه داده SQKServer نیز استفاده شده است. یکی از پرکاربردترین نوع برچسب‌گذاری، برچسب‌گذاری اجزای واژگانی کلام(POS) است. این نوع برچسب‌گذاری، عملی کاربردی در بسیاری از حوزه‌های پیشرفته‌تر پردازش زبان طبیعی از جمله ترجمهٔ ماشینی، خطایاب، تبدیل متن به گفتار، بازیابی اطلاعات، موتورهای جستجو و کمک به مدل‌های آماری است. پس از ساخت پیکرهٔ متنی ایرانداک، با توجه به کاربرد برچسب اجزای واژگانی کلام در پردازش متن، این نوع برچسب‌گذاری هم به پیکره اضافه شد. در این راستا تصمیم گرفته شد ابتدا از یکی از ابزارهای آماده برای برچسب‌گذاری ماشینی اجزای واژگانی کلام استفاده شود و سپس، برچسب‌ها دستی کنترل شوند. ابزار استفاده شده برای برچسب‌گذاری ماشینی، ابزار «هضم» است که برای نرمال‌سازی متون پیکره نیز از این ابزار استفاده شد. فهرست برچسب‌ها، شامل برچسب‌هایی است که در مقالهٔ بی‌جن‌خان و همکاران (۲۰۱۱) معرفی شده است. در نهایت، برچسب‌ها چک شدند و تا حد امکان از طریق انجام

اصلاحات روی کدهای ابزارهای نرمال‌سازی و کنترل دستی، برچسب‌های غلط اصلاح شدند. نمودار جریانی(block diagram) از ماجول‌های ساخت پیکره به این صورت است:



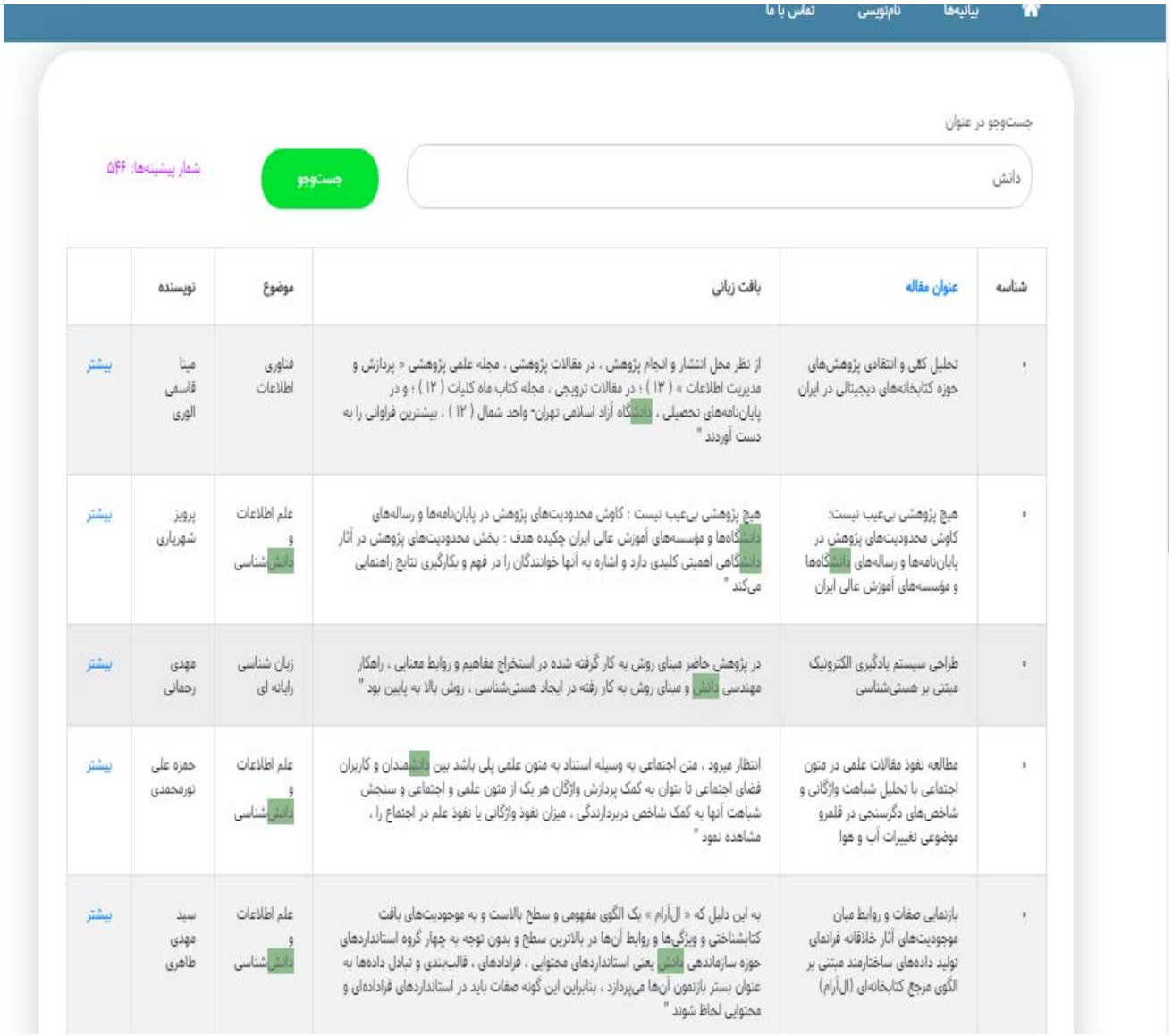
نمونه‌ای از جستجوی یک واژه در پیکرهٔ متنی ایرانداک به شکل زیر است:

ابتدا وارد لینک زیر می‌شویم:

http://peykare.irandoc.ac.ir/

این صفحه نمایش داده می‌شود:

سپس، یک واژه، به عنوان نمونه واژهٔ «دانش» را در قسمت جستجو وارد می‌کنیم و صفحه‌ای مانند زیر نمایش داده می‌شود که نشان می‌دهد این واژه چندبار در پیکره تکرار شده است و در چه مقاله‌هایی این واژه یافت می‌شود:



برای دیدن واژه در متن هر مقاله می‌توان روی گزینهٔ «بیشتر» کلیک کرد و صفحهٔ زیر نمایش داده می‌شود:

نویسنده	مینا قاسمی الوری

تمام متن

تمام متن | برچسب گذاری شده

تحلیل کلی و انتقادی پژوهش‌های حوزه کتابخانه‌های دیجیتالی در ایران چکیده پژوهش حاضر با هدف تحلیل محتوای کلی و انتقادی پژوهش‌های انجام شده حوزه کتابخانه‌های دیجیتال پژوهشگران ایرانی از منظر روش تحقیق ، جامعه پژوهش ، موضوعات ، نویسندگان و وابستگی سازمانی آن‌ها و نتایج حاصل از پژوهش‌های پیشین انجام شد . روش پژوهش تحلیل محتوای کلی است و برای این منظور اطلاعات نویسندگان ، عنوان ، چکیده ، کلیدواژه‌های مقالات و پایان‌نامه‌های مربوط به این حوزه مورد تجزیه‌وتحلیل قرار گرفت . جامعه آماری پژوهش شامل ۲۰۹ بروندادِ پژوهشی ایران (۱۳۸ مقاله و ۷۱ پایان‌نامه) در حوزه کتابخانه‌های دیجیتالی است که مبتنی بر ملاک اعتبار درون‌سنجی تمامی آن برگزیده شدند . از نظر محل انتشار و انجام پژوهش ، در مقالات پژوهشی ، مجله علمی پژوهشی « پردازش و مدیریت اطلاعات » (۱۳) ؛ در مقالات ترویجی ، مجله کتاب ماه کلیات (۱۲) ؛ و در پایان‌نامه‌های تحصیلی ، دانشگاه آزاد اسلامی تهران- واحد شمال (۱۲) ، بیشترین فراوانی را به دست آوردند . از نظر موضوعات به کار رفته ، نتایج نشان داد موضوعات « ارزیابی کتابخانه‌های دیجیتال » ، « مفاهیم کتابخانه‌های دیجیتال ، مجازی و الکترونیک » ، « امکان‌سنجی ایجاد کتابخانه دیجیتال و « محیط رابط کاربر کتابخانه‌های دیجیتال » بیشترین فراوانی را کسب کردند . نتایج تحلیل محتوا گویای آن است که موضوع کتابخانه‌های دیجیتال در پژوهش‌های پیشین در سال‌های اخیر فرایند رو به رشدی داشته و موضوعات جدیدی از جمله وب معنایی ، هستی نگاشت ، یکپارچه‌سازی و سازماندهی معنایی در این حوزه توجه پژوهشگران را به خود جلب کرده‌است . همچنین ، بررسی پژوهش‌ها نشان داد که اولاً بخش عمده تحقیقات درصدد معرفی و ترویج مفهومی و مصداقی کتابخانه دیجیتالی بودند ، یعنی دانشی که عمده آن به زبان انگلیسی موجود است ؛ ثانیاً ، موضوع بخش زیادی از تحقیقات تکراری است ؛ ثالثاً ، پژوهشی که از نظر مفهومی و یا خلق فناوری به توسعه حوزه کتابخانه دیجیتالی بپردازد ، یافت نشد . اضافه بر این ، نتایج تحقیق برخی کژی‌های روش‌شناسی تحقیقات پیشین را نشان داد که نمونه بارز آن ، ذکر روش‌های تحقیقی است که کاملاً قابل تأمل می‌باشند . در مجموع ، بررسی موضوعات مورد مطالعه و سیر زمانی انجام آن‌ها در حوزه نظری و کاربردی کتابخانه‌های دیجیتال در ایران ، منجر به شناسایی خلأهای پژوهشی مربوط به این حوزه شد . لذا به پژوهشگران پیشنهاد می‌شود به سایر جنبه‌های کتابخانه‌های دیجیتال که مورد بررسی قرار نگرفته‌اند و یا مطالعات کمتری به آن اختصاص داده شده ، بپردازند . کلیدواژه‌ها : کتابخانه‌های دیجیتال ، مقالات ، پایان‌نامه‌ها ، تحلیل محتوا مقدمه دسترسی مؤثر به اطلاعات و منابع اطلاعاتی همیشه یکی از دغدغه‌های بشر بوده‌است . به همین دلیل ، نهادهایی مانند کتابخانه‌ها از روزگاران باستان تشکیل شده‌اند تا منابع اطلاعاتی را گردآوری ، سازماندهی ، نگهداری و ارائه کنند تا دسترسی به آن‌ها به بهترین شیوه امکان‌پذیر شود . با نگاهی تاریخی به این نهادها طی قرون گذشته در خواهیم‌یافت که همواره با تغییرهایی همراه بوده‌اند ، گاهی در شیوه گردآوری منابع ، گاهی در نحوه سازماندهی منابع و گاهی در سطح دسترسی به منابع آن‌ها . اما شاید بتوان گفت سرآمد این تغییرها در دهه ۱۹۹۰ رخ داد . پیدایش وب که همه عرصه‌ها را با تحول‌های چشمگیری مواجه کرد ، کتابخانه‌ها را نیز وارد عصر دیجیتالی کرد . در این عصر ، کتابخانه‌ها وارد فضای برخط شدند و محدودیت‌های دسترسی به منابع تا حد زیادی از بین رفت . « کتابخانه‌های دیجیتالی » عنوان جدیدی بود که برای این‌گونه از کتابخانه برگزیده شد (کوشا ۱۳۸۵) . از منظر اجتماعی ، فدراسیون کتابخانه دیجیتالی (۱۹۹۸) "این کتابخانه‌ها را سازمان‌هایی می‌داند که در آن کارکنان متخصص به انتخاب ، سازمان‌دهی ، و کمک برای دسترسی به منابع اطلاعاتی می‌پردازند و در آن فرایند تفسیر ، توزیع ، حفاظت از یکپارچگی اطلاعات دیجیتالی و نیز اطمینان از وجود مجموعه‌ای از آثار دیجیتالی در مدت زمان طولانی مورد توجه قرار می‌گیرد . تا از این طریق بتوان اطلاعات دیجیتال را به سرعت و به طور اقتصادی برای استفاده یک جامعه یا مجموعه‌ای از جوامع در دسترس قرار داد . این تعریف با توجه به هدف نهایی کتابخانه‌های دیجیتالی که تأمین رضایت کاربران است ، تمامی ارکان اساسی دستیابی به این هدف را معرفی می‌کند و در عین حال ، تمامی انواع کتابخانه‌های دیجیتالی با زمینه‌ها و محتوای متنوع را نیز در برمی‌گیرد . از همه مهم‌تر ، قابلیت استفاده آن‌ها را برای انواع کاربران و کاربردهای مختلف آن را مد نظر قرار می‌دهد . از این رو ، لازم است توجه بیشتری نسبت به وضعیت این‌گونه کتابخانه‌ها از ابعاد مختلف صورت گیرد" (نقل در نوروزی ، غلامی و جعفری‌فر ۱۳۹۶ ، ۱۴۸) . همان‌گونه که جعفری فر (۱۳۹۲) اشاره دارد ، با این وجود ، کتابخانه‌های دیجیتالی فعلی در کشور فاصله زیادی تا رسیدن به نقطه استاندارد ، و با نمونه‌های خارجی خود دارند . برخلاف کتابخانه‌های سنتی که سازوکار مشخصی داشته و پشتوانه قرن‌ها تجربه را دارا هستند ، کتابخانه‌های دیجیتالی نوپا بوده و تحقیقات در این حوزه در مقایسه با کتابخانه‌های سنتی بسیار محدودتر است . لازمه ارتقاء این توان و ظرفیت ، بهبود وضعیت تولید اطلاعات علمی در این حوزه است

برای دیدن برچسب واژه‌ها می‌توان روی گزینهٔ «برچسب‌گذاری شده» کلیک نمود که در آنصورت صفحهٔ زیر نمایش داده می‌شود:

٦-۲- پیشنهادات برای پژوهش‌های آینده

- با توجه به نیاز تشخیص افعال مرکب در ترجمهٔ ماشینی مبتنی بر پیکره، می‌توان ابتدا از میان رویکردهای رایج برای تشخیص افعال مرکب در زبان فارسی، یکی را انتخاب نمود و در پیکرهٔ متنی ایرانداک به جست‌وجوی افعال مرکب (جدانشدنی و جداشدنی) پرداخت و برچسب «فعل مرکب/ Compound Verb (CV) را به آن‌ها اختصاص داد.
- با توجه به کاربرد برچسب معنایی در بسیاری از حوزه‌های پردازش زبان طبیعی، مانند استخراج اطلاعات، پرسش و پاسخ، خلاصه‌سازی و ترجمهٔ ماشینی، می‌توان برچسب‌گذاری نقش معنایی را نیز به پیکرهٔ متنی ایرانداک افزود.
- با توجه به موجود بودن چکیده‌های فارسی و انگلیسی پایان‌نامه‌ها و مقاله‌های پژوهش‌نامهٔ پردازش و مدیریت اطلاعات در ایرانداک، می‌توان با بررسی ملاحظات حقوقی، پیکره‌ای موازی حاوی چکیده‌های فارسی و انگلیسی مقاله‌ها یا پایان‌نامه‌ها ساخت و به سامانهٔ پیکره‌های ایرانداک اضافه کرد.

فهرست منابع فارسی

- آیت، سید سعید. ۱۳۸۹. طراحی و پیاده‌سازی دادگان دایفون زبان فارسی برای کاربرد زبانشناسی رایانه‌ای. *پژوهش‌های زبان‌شناسی*. دوره ۲. شماره ۳. صص ۱۱-۱.
- اسلامی، محرم.، مسعود شریفی آتشگاه، صدیقه علیزاده لمجیری، و طاهره زندی. ۱۳۸۳. واژگان زایای زبان فارسی .*مجموعه مقالات اولین کارگاه پژوهشی زبان فارسی و رایانه. تهران*.
- اسلامی، محرم.، جواد شیخ‌زادگان، زهرا احمدی‌نیا، و علی بهرامی راد. ۱۳۸۸. مراحل و نحوهٔ تهیهٔ دادگان‌های صوتی هجایی و دایفونی برای سامانهٔ تبدیل متن به گفتار فارسی. *دوفصل‌نامه علمی-پژوهشی پردازش علائم و داده‌ها*، (۲)، ۳-۱۲.
- بی‌جن‌خان، محمود. ۱۳۸۳. نقش پیکره‌های زبانی در نوشتن دستور زبان: معرفی یک نرم‌افزار رایانه‌ای. *مجلهٔ زبان‌شناسی* ۱۹(۲): ۶۷-۴۸.
- تابع بردبار، علیرضا. ۱۳۹۳. *ایجاد پیکره زبانی موازی به وسیله پیکره‌های قیاس‌پذیر*. کارشناسی ارشد. شیراز: دانشگاه شیراز.
- دشتبانی، شکوفه. ۱۳۹۱. *ساخت پیکرهٔ متنی فارسی حوزهٔ فاوا*. همدان: دانشگاه بوعلی سینا. دانشکدهٔ فنی و مهندسی.
- صادقی، علی‌اشرف. ۱۳۷۰. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۱)». تهران: نشر دانش. شمارهٔ ۶۴.
- صادقی، علی‌اشرف. ۱۳۷۰. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۲)». تهران: نشر دانش. شمارهٔ ۶۵.
- صادقی، علی‌اشرف. ۱۳۷۰. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۳)». تهران: نشر دانش. شمارهٔ ۶۷.
- صادقی، علی‌اشرف. ۱۳۷۱. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۴)». تهران: نشر دانش. شمارهٔ ۶۹.

- صادقی، علی‌اشرف. ۱۳۷۱. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۵)». تهران: نشر دانش. شمارهٔ ۷۰.
- صادقی، علی‌اشرف. ۱۳۷۱. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۶)». تهران: نشر دانش. شمارهٔ ۷۱.
- صادقی، علی‌اشرف. ۱۳۷۱. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۷)». تهران: نشر دانش. شمارهٔ ۷۲.
- صادقی، علی‌اشرف. ۱۳۷۱. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۸) ».تهران: نشر دانش. شمارهٔ ۷۴.
- صادقی، علی‌اشرف. ۱۳۷۲. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۹)». تهران: نشر دانش. شمارهٔ ۷۵.
- صادقی، علی‌اشرف. ۱۳۷۲. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۱۰)».تهران: نشر دانش. شمارهٔ ۷۶.
- صادقی، علی‌اشرف. ۱۳۷۲. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۱۱)». تهران: نشر دانش. شمارهٔ ۷۷.
- صادقی، علی‌اشرف. ۱۳۷۲. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۱۲)». تهران: نشر دانش. شمارهٔ ۷۹ و ۸۰.
- عاصی، مصطفی. ۱۳۸۴. گزارش کوتاهی از شکل‌گیری پایگاه داده‌های زبان فارسی در اینترنت. *مجله پژوهشگران شماره ۲. صفحه ۱۳.* https://hawzah.net/fa/Magazine/View/6280/6282/69701
- علایی، الهام.، و سیروس علیدوستی. ۱۳۹۸. ساخت پیکرهٔ متنی: طراحی مدل امکان‌سنجی. *پژوهش‌های زبان‌شناسی تطبیقی*. مقاله‌های آماده انتشار. پاییز و زمستان ۹۹.
- قطره، فریبا. ۱۳۸۶. «مشخصه‌های تصریفی در زبان فارسی امروز». *دستور*. شماره ۳. ۵۲-۸۱.
- کشانی، خسرو. ۱۳۷۱. *اشتقاق پسوندی در زبان فارسی امروز*. تهران: مرکز نشر دانشگاهی.

- لازار، ژیلبر. ۱۳۸۹. *دستور زبان فارسی معاصر*. ترجمهٔ مهستی بحرینی و توضیحات و حواشی هرمز میلانیان. تهران: انتشارات هرمس. چاپ دوم.
- محمدی، رویا. ۱۳۹۱. *ساخت پیکرهٔ تطبیقی فارسی-انگلیسی و استخراج جملات موازی از آن*. پایان‌نامهٔ کارشناسی ارشد. دانشگاه الزهرا (س). دانشکدهٔ فنی و مهندسی.
- مرادی، مهدی.، و محمد بحرانی. ۱۳۹۴. تشخیص خودکار جنسیت نویسنده در متون فارسی. *پردازش علائم و داده‌ها* ۱۲ (۴): ۸۳-۹۴.
- میرزایی، آزاده.، و امیرسعید مولودی. ۱۳۹۳. نخستین پیکرهٔ نقش‌های معنایی در زبان فارسی. *علم زبان* ۲ (۳): ۴۸-۲۹.

فهرست منابع لاتین

- AleAhmad, A., H. Amiri, E. Darrudi, M. Rahgozar, and F. Oroumchian. 2009. Hamshahri: A Standard Persian Text Collection. *Knowledge-Based Systems*. Dubai 22(5): 382–387. Elsevier.

- Atkins, S., J. Clear, and N. Ostler. 1992. Corpus design criteria. *Literary and Linguistic Computing*. 7 (1). 1-16

- Bijankhan, M., J. Sheykhzadegan, M. Bahrani, and M. Ghayoomi. 2011. Lesson from building a Persian written corpus: Peykare. *Language resources and evolution* 45 (2): 143-164. Springer.

- Bijankhan, M., J. Sheykhzadegan, M. R. Roohani, R. Zarrintare, S. Z. Ghasemi, and M. E. Ghasedi. 2003. Tfarsdat - The Telephone Farsi Speech Database. In *Proceeding of EUROSPEECH,* 1525-1528, Geneva, Switzerland.

- Claude Toriida, M. 2016. Steps for creating specialized corpus and developing an annotated frequence-based vocabulary list. *TESL Canada journal/ revue TESL du Canada* 34 (11): 87-105.

- Dehghani, M., A. Shakery, M. Asadpour, and A. Koushkestani. 2013. A learning approach for email conversation thread reconstruction. *Journal of Information Science* (JIS) 39(6): 846-863.

- Durand, J., U. Gut, and K. Gjert. 2014. *The handbook of corpus phonology*. Oxford University Press (OUP).

- Ghayoomi, M., S. Momtazi, and M. Bijankhan. 2013. A study of corpus development for Persian. *International journal on Asian language processing* 20 (1): 17-33.

- https://github.com/sobhe/hazm

- Leech, G. 2004. *Developing Linguistic Corpora: a Guide to Good Practice. adding linguistic annotaion. Edited by Martin Wynne .ahds.literature, languages and linguistics*. The Oxford Text Archive.

- Megerdoomian, K. 2004. Developing a Persian part-of-speech tagger. *Proceedings of the 1st workshop on Persian language and computer*. Pp. 99-105.

- Rasooli, M., M. Kouhestani, and A. Moloodi. 2013. Development of a Persian Syntactic Dependency Treebank. *In The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*: 306-314. Atlanta, USA.

- Samvelian, P., and P. Faghiri. 2013. Introducing PersPred, A Syntactic and Semantic Database for Persian Complex Predicates. In *Proceedings of the 9th Workshop on Multiword Expressions, Atlanta, Georgia, USA. Association for Computational Linguistics*, 11-20.

- Shamsfard, M., A. Hesabi, H. Fadaei, N. Mansoory, A. Famian, S. Bagherbeigi, E. Fekri, et al. 2010. Semi Automatic Development of Farsnet; the Persian Wordnet. *Proceedings of 5th Global WordNet Conference* (GWA). Mumbai, India.

- Sinclair, J. 2004. *Developing Linguistic Corpora: a Guide to Good Practice. Chapter 1: Corpus and Text — Basic Principles.* Edited by Martin Wynne .ahds.literature, languages and linguistics. The Oxford Text Archive.

- Waynne, M. 2005. *Developing linguistic corpora: a guide to good practice.* Oxbow books. Literary and linguistic computing 22 (1).

Abstract

Numerous linguistic studies as well as language planning involves the use of corpus to analyze different aspects of language. In present study a corpus was built from published articles of the journal of information processing and management. The corpus include 677 articles and 4774790 words. The articles cover hot topics including information science, information technology, information management, knowledge management, linguistics, computational linguistics, terminology, onthology and the like; so, actually the corpus is the spcialized one which could be considered as a prerequisite for specific language processing. After sampling, in present study, data was inserted to corpus, then metadata (including the title, topic and author/authors of each article) was also inserted. Then, normalization and POS tagging were done via a software and finally the tags were manually controlled. Besides, a website was also desighned titled “IranDoc corpora” and the present corpus was inserted to it. In future more corpora could be added to the website.

Key words: corpus, normalization, POS tagging, IranDoc corpora

Iranian Research Institute
for Information Science and Technology (Irandoc)

Building a text corpus from the articles published in journal of information processing & management

By:

Elham Alayiaboozar

Nasrollah Pakniyat
Aliasghar Hojjatpanah
Mojtaba Zali
Mohammadhadi Aghalooy

April 2021