

دانیار، مشاور سازمانی هوشمند و باسابقه‌ی مبتنی بر هوش مصنوعی مولد

طبقه بندی: محرمانه با دسترسی محدود



تحلیلگران هوشمند فناوری اطلاعات امید(امیدتک)

اعتبار پیشنهاد:

اعتبار این پیشنهاد (Proposal) از تاریخ تهیه (۲۷ آبان ۱۴۰۳) به مدت 90 روز می‌باشد.

1 بیان مساله

فرض كنيد، هيات دولت يا يك وزارتخانه مي خواهد آئين نامه جديدي تصويب كند. دانستن تاريخچه آئين نامه ها و تصميمات مرتبط با موضوع در دولت هاي قبل مي تواند بسيار مفيد باشد. دست يابي به تمامي جزئيات مرتبط با اين آئين نامه جديد، شامل تحقيقات انجام شده جهت تصويب آئين نامه هاي مرتبط، افراد درگير در اين تصميم گيري ها، ميزان موفقيت تصميمات و آئين نامه هاي مرتبط، دلايل شكست آنها و غيره، يك دارايي گران بها براي هيات دولت يا وزارتخانه محسوب مي شود.

و يا به عنوان نمونه اي ديگر، هنگام تصميم گيري براي تدوين بودجه سالانه كشور، دانستن اطلاعات جزئي درباره احكام، تبصره ها، منابع، سقف درآمدها و تراز عملياتي، سقف واگذاري دارايي هاي سرمايه اي و غيره مرتبط با لوايح بودجه كل كشور در سال هاي گذشته مي تواند به تدوين دقيق و كارآمد بودجه سال جديد كمك كند.

در هيات دولت و يا هر وزارتخانه، اسناد، نامه ها، طرح ها و راهنماهاي مفصلي طي سال ها تدوين شده كه به صورت يکپارچه اي گرد هم نيامده اند. اطلاعات جزئي درباره اين اسناد مي تواند منجر به ايجاد تحول در تصميمات سازمان شود و باعث شود سازمان متحمل هزينه هاي گزاف نشود.

امروزه با افزايش حجم اسناد و داده ها در سازمان ها، مديريت و استخراج داده هاي موجود در اين اسناد به چالشي بزرگ براي متخصصان و مديران تبديل شده است. حجم زياد اسناد، تنوع فرمت ها، وجود اطلاعات در قالب هاي مختلف مانند متن، تصاوير و جداول و نياز به دسترسي سريع و دقيق به اطلاعات اين مسئله را به يك چالش اساسي تبديل کرده است. با افزايش حجم داده ها و اسناد، نياز به پردازش سريع تر و بهينه تر اطلاعات و نياز به تحليل آنها نيز بيشتر از قبل احساس مي شود.

جستجو و نگهداري حجم زياد اسناد و داده ها و مديريت اطلاعات به طور كلي زمان بر است. همچنين، اين نياز مبرم به باز يابي اطلاعات و استفاده صحيح از آنها براي مديريت و پيشبرد تصميمات سازمان در آينده، نياز به هزينه هاي اضافي و وقت بيش تري دارد. در اين پروپوزال و با معرفي «دانايار» به ارائه راه حلي مبتني بر هوش مصنوعي براي حل اين مساله مي پردازيم.

2 تفسیر تکنولوژی‌های استفاده شده در این طرح

این بخش به معرفی تکنولوژی‌های استفاده شده در این طرح اختصاص داده شده است.

2-1 هوش مصنوعی

هوش مصنوعی (AI) بیانگر توسعه سیستم‌های کامپیوتری است که توانایی انجام وظایفی را دارند که به طور معمول نیاز به هوش انسانی دارند. این حوزه گسترده شامل زیرمجموعه‌های مختلفی مانند یادگیری ماشین، پردازش زبان طبیعی، بینایی ماشین، رباتیک و سامانه‌های خبره می‌شود. سیستم‌های هوش مصنوعی توانایی‌هایی مانند یادگیری، استدلال، حل مسئله، استنتاج و تصمیم‌گیری را شبیه به انسان‌ها شبیه‌سازی می‌کنند.

هوش مصنوعی در سال‌های اخیر به دلیل عواملی همچون حجم بزرگ داده‌ها، افزایش قدرت محاسباتی، الگوریتم‌ها و تکنیک‌های نوین، پیشرفت‌های قابل توجهی داشته است. این فناوری‌ها در زمینه‌های گوناگونی از جمله بهداشت، مالی، حمل و نقل، سرگرمی و امنیت سایبری به کار می‌روند. آن‌ها قادرند صنایع را تغییر داده، کارایی را بهبود بخشند، تصمیم‌گیری را تقویت کرده و امکانات جدیدی فراهم کنند.

یادگیری ماشین به تمرکز بر روی توانایی کامپیوترها برای یادگیری از داده‌ها و انجام پیش‌بینی‌ها یا تصمیم‌گیری‌ها براساس داده‌ها می‌پردازد. این حوزه شامل توسعه الگوریتم‌هایی است که به ماشین‌ها امکان می‌دهند به صورت خودکار الگوها را یاد بگیرند، بینش‌ها را استخراج کنند و عملکرد خود را بدون نیاز به برنامه‌ریزی صریح، به مرور زمان بهبود بخشند.

پردازش زبان طبیعی (NLP) نیز به تحقیق و بررسی تعامل بین کامپیوترها و زبان انسانی می‌پردازد. این زمینه به ماشین‌ها امکان می‌دهد زبان انسان را درک، تفسیر و تولید کنند و وظایفی همچون ترجمه زبانی، تحلیل احساسات، چت‌بات‌ها و تشخیص صوتی را ممکن می‌سازد.

2-1-1 پردازش زبان طبیعی و کاربردهای آن

پردازش زبان طبیعی (NLP) دارای کاربردهای گسترده‌ای در صنایع مختلف است. در زیر به بررسی چند کاربرد اساسی NLP پرداخته می‌شود:

1. دستیارهای مجازی: NLP در توسعه دستیارهای مجازی مانند Siri، Alexa و Google Assistant مورد استفاده قرار می‌گیرد. این دستیارها قادرند به سوالات به زبان طبیعی پاسخ دهند و وظایف مختلف را با اطلاعات دقیق و دستورات کاربر انجام دهند.
2. چت‌بات‌ها: NLP به چت‌بات‌ها امکان می‌دهد با کاربران برای تعامل و ارائه پاسخ‌های خودکار در گفتگوها ارتباط برقرار کنند. این کاربرد در اموری همچون پشتیبانی مشتریان و تعاملات وب سایت‌ها به کار می‌رود.
3. تحلیل احساسات: تکنیک‌های NLP برای تحلیل و تعیین احساسات موجود در داده‌های متنی نظیر نظرات مشتریان یا بازخوردها به کار می‌روند. این تحلیل به فهم نظر جمعی، رضایت مشتریان و تاثیر برند کمک می‌کند.
4. ترجمه زبان: NLP به ترجمه ماشینی متن و گفتار از یک زبان به زبان دیگر کمک می‌کند. سرویس‌هایی نظیر ترجمه گوگل و مترجم مایکروسافت از تکنیک‌های NLP برای ارائه ترجمه‌های دقیق استفاده می‌کنند.
5. استخراج اطلاعات: تکنیک‌های NLP برای استخراج اطلاعات ساختاری از داده‌های متنی بدون ساختار استفاده می‌شوند که این امر به شناسایی موجودیت‌های مختلف نظیر نام‌ها، تاریخ‌ها و مکان‌ها کمک می‌کند.
6. خلاصه‌سازی متن: الگوریتم‌های NLP با انجام خلاصه‌سازی متن، متن‌های بزرگ را به شکل خلاصه‌تر و مفیدتر تبدیل می‌کنند. این فرآیند برای کاربردهای تجمیع اخبار و خلاصه‌سازی اسناد اهمیت دارد.
7. پاسخ به سوالات: سیستم‌های NLP قادر به درک و پاسخ به سوالاتی که به زبان طبیعی مطرح می‌شوند هستند.
8. تشخیص گفتار: NLP نقش مهمی در سیستم‌های تشخیص گفتار دارد.
9. دسته‌بندی متن: این کاربرد شامل دسته‌بندی داده‌های متنی به کلاس‌ها یا موضوعات مختلف است. این فرآیند در طبقه‌بندی اسناد یا تشخیص اسپم اثربخش است.

2-1-2 عامل هوشمند

هر چیزی که محیط را از طریق حسگر شناسایی کند و بر محیطی از طریق اهرم‌ها یا عمل‌کننده‌ها تاثیر بگذارد، «عامل»¹ یا «کنشگر» نامیده می‌شود. عامل‌ها وظایفی از جمله تشخیص دادن، فکر کردن و عمل کردن را در یک برنامه انجام می‌دهند. عامل‌ها انواع گوناگونی دارند که در ادامه به معرفی آن‌ها پرداخته شده است:

«عامل‌های انسانی»²: در بدن انسان چشم‌ها، گوش‌ها و سایر اندام‌های حسی به عنوان عامل و دست‌ها، پاها به عنوان اهرم (محرك) در نظر گرفته می‌شوند.

¹ Agent

² Human Agent

«عامل‌های رباتیک»¹: ربات‌ها، دوربین‌ها، «مسافت‌یاب‌های مادون قرمز»، حسگرها و موتورهای گوناگون دیگر، عامل‌های رباتیک به حساب می‌آیند.

«عامل‌های نرم افزاری»²: مجموعه‌ای از برنامه‌های نرم افزاری که برای اجرای وظیفه‌هایی مشخص طراحی شده‌اند. برای مثال می‌توان به برنامه‌هایی اشاره کرد که بررسی محتوای ایمیل‌های دریافتی و گروه‌بندی آن‌ها را انجام می‌دهند.

عامل هوشمند³، نوعی عامل به حساب می‌آید که می‌تواند وظایف خاص، قابل پیش‌بینی و تکراری را به صورت مستقل انجام دهد و در حین انجام وظایف توان یادگیری هم دارد. می‌توان گفت که این عامل‌ها دارای برخی از خصوصیات روحی انسان از جمله علم، عقیده، قصد و سایر موارد هستند. به عنوان مثال‌هایی برای درک بهتر عامل‌های هوشمند می‌توان به سیستم‌های «پردازش زبان طبیعی» مانند Alexa از شرکت آمازون و Siri از شرکت اپل اشاره کرد. عامل هوشمند دارای دو عملکرد اصلی است که در ادامه به آن‌ها پرداخته شده است:

درک⁴: به عملکردهایی می‌گویند که توسط حسگرها انجام می‌شوند.

عمل⁵: به عملکردهایی گفته می‌شود که به وسیله عمل‌کننده‌ها یا همان محرک‌ها انجام می‌گیرند.

عامل هوشمند برنامه‌ای است که می‌تواند بر اساس محیط، ورودی کاربر و تجربه‌های خود تصمیم بگیرد و خدماتی را انجام دهد. این برنامه‌ها می‌توانند برای جمع‌آوری مستقل اطلاعات در یک زمان‌بندی منظم و برنامه‌ریزی شده یا زمانی استفاده شوند که کاربر در «زمان واقعی» درخواستی داشته باشد. همچنین ممکن است عامل‌های هوشمند به عنوان ربات نیز شناخته شوند.

2-1-3 OCR⁶ تکنولوژی

یکی از تکنولوژی‌های مرتبط با هوش مصنوعی که به طور وسیع مورد استفاده قرار می‌گیرد، فناوری OCR یا نویسه‌خوان نوری است. این فناوری می‌تواند به صورت خودکار متون موجود در تصاویر اسناد را به متون قابل جست‌وجو و ویرایش توسط رایانه‌ها تبدیل کند و کاربردهای گوناگونی همچون ثبت سریع داده‌ها در سیستم، تبدیل اسناد تصویری به متون قابل جست‌وجو و حتی کمک به نابینایان برای مطالعه تصویری دارد و متون دیجیتالی را در سریع‌ترین زمان در اختیار کاربران خود قرار می‌دهد. تکنولوژی OCR و نرم‌افزارهایی که از آن استفاده می‌کنند، از نحوه کار مشخصی برخوردارند. نحوه کار به این صورت است که ابتدا تصاویری که دارای متن و نوشته خاصی هستند توسط دستگاه اسکنر و یا دوربین دیجیتال، اسکن شده و وارد رایانه می‌شوند سپس با استفاده از OCR فایل تصویری مورد نظر تبدیل به فایل متنی شده و در یکی از قالب‌های متنی در کامپیوتر ذخیره می‌شود.

¹ Robotic Agent

² Software Agent

³ Intelligent agent

⁴ Perception

⁵ Action

⁶ Optical Character Recognition

فناوری OCR، عضوی از یک خانواده بزرگ‌تر از فناوری‌های مشابه است که انواع مختلفی دارند. در ادامه آنها را به اختصار معرفی

می‌کنیم:

1. تشخیص هوشمند کلمه¹: این تکنولوژی، متن شکسته یا متون دست نویس را تشخیص می‌دهد. الگوریتم این تکنولوژی با تشخیص یک کلمه دست نویس به جای شناسایی کارکترهای فردی کار می‌کند.
2. تشخیص کاراکتر هوشمند²: این تکنولوژی، متن دست نویس یا شکسته را تشخیص می‌دهد. اما برخلاف مورد قبلی، الگوریتم آن موتور با شناسایی تک کارکترها کار می‌کند و با کمک یادگیری ماشینی تکامل پیدا می‌کند.
3. تشخیص نوری کلمه³: این تکنولوژی، تشخیص متن تایپ شده را به صورت کلمه‌ای هدف قرار می‌دهد و گاهی اوقات به عنوان OCR نیز شناخته می‌شود.
4. تشخیص کاراکتر نوری⁴: این تکنولوژی، متن تایپ شده را تشخیص می‌دهد، اما برخلاف مورد قبلی، الگوریتم آن بر اساس شناسایی تک کارکترها کار می‌کند.
5. تشخیص علامت نوری⁵: تکنیکی برای جمع‌آوری داده‌های مربوط به علائم نوشته شده توسط انسان‌ها است که با تشخیص علائم یا الگوها روی اسناد کار می‌کند.

2-2 مدل‌های زبانی

مدل‌های زبانی که بخش مهمی از سیستم‌های پردازش زبان طبیعی (NLP) را تشکیل می‌دهند، بر تولید و درک زبان انسان متمرکزند. این مدل‌ها بر اساس داده‌های متنی بزرگ آموزش می‌بینند و تلاش می‌کنند تا رابطه بین کلمات، عبارات و جملات را درک کرده و درک نمایند.

2-3 مدل‌های زبانی بزرگ

مدل‌های زبان بزرگ (LLM) مانند GPT-3.5 و GPT-4.0 از شرکت OpenAI، BERT، T5 و PaLM از گوگل و همچنین مدل LLaMA از شرکت فیس‌بوک، به عنوان مدل‌های پیشرفته هوش مصنوعی طراحی شده‌اند که با آموزش بر روی حجم زیادی از

¹ Intelligent Word Recognition

² Intelligent Character Recognition

³ Optical Word Recognition

⁴ Optical Character Recognition

⁵ Optical Mark Recognition

داده‌های متنی، قادر به درک و تولید زبان انسانی هستند. این مدل‌ها برای پردازش و تولید متن به انگیزه‌های خاصی طراحی شده‌اند و سطحی از هماهنگی و توالی در تولید متن دارند.

فرآیند آموزش مدل‌های زبان بزرگ شامل قرار دادن مدل در برابر مجموعه داده‌های بسیار بزرگ است که اغلب تشکیل شده از میلیاردها جمله از منابع گوناگون نظیر کتب، مقالات، وبسایت‌ها و سایر منابع متنی است. این مدل‌ها الگوهای آماری و ساختارهای زبانی موجود در داده‌ها را یاد می‌گیرند و قادرند پاسخ‌های هماهنگ و معناداری ایجاد کنند.

یکی از ویژگی‌های کلیدی مدل‌های زبان بزرگ، توانایی در درک و تولید متن در چندین زبان است. آنها قادر به آموزش بر روی مجموعه داده‌های چند زبانه هستند و این امکان را ایجاد می‌کنند تا ترجمه‌ها، پاسخ به سوالات و مشارکت در مکالمات را در زبان‌های گوناگون ارائه دهند.

مدل‌های زبان بزرگ به عنوان یک ابزار چندمنظوره برای وظایف مختلف از جمله درک زبان طبیعی، تولید زبان طبیعی، پاسخ به سؤال، خلاصه‌سازی متن و ترجمه زبان استفاده می‌شوند. اما توجه به نکته این مهم است که این مدل‌ها به شدت وابسته به داده‌هایی هستند که بر روی آنها آموزش دیده‌اند، که این می‌تواند به ناسازگاری منجر شود. بنابراین ارزیابی دقیق به منظور اطمینان از کیفیت و عدالت خروجی‌های تولید شده توسط این مدل‌ها ضروری است.

2-4 هوش مصنوعی مولد

مدل‌های زبانی بزرگ (LLM) بخشی از یک نهاد بزرگتر در هوش مصنوعی به نام هوش مصنوعی مولد (GAI) هستند. هوش مصنوعی مولد، زیرشاخه‌ای از هوش مصنوعی است که بر تولید مدل‌هایی متمرکز است که قادر به تولید محتوای جدید مانند تصاویر، موسیقی، متن و حتی ویدیو هستند. این مدل‌ها بر روی مجموعه داده‌های بزرگی آموزش می‌بینند و الگوها و ساختارهای موجود در داده‌ها را یاد می‌گیرند تا نمونه‌های جدیدی تولید کنند که شبیه به مثال‌های آموزشی باشند.

هوش مصنوعی مولد کاربردهای مختلفی در زمینه‌های گوناگون دارد:

1. تولید تصاویر: مدل‌های مولد قادر به ایجاد تصاویر واقع‌گرایانه مانند چهره‌های انسان، مناظر طبیعی یا اشیاء هستند. این مدل‌ها در برنامه‌های خلاقانه، طراحی و سرگرمی مورد استفاده قرار می‌گیرند.
2. تولید موسیقی و صدا: هوش مصنوعی مولد قادر به تولید آهنگ‌ها و خلق صداهای جدیدی است. از آن در تولید موسیقی، ساخت موسیقی پس‌زمینه و حتی تولید پیشنهاد‌های شخصی‌سازی برای کاربران استفاده می‌شود.
3. تولید متن: مدل‌های زبانی مانند GPT قادر به تولید متن شبیه به انسان بر اساس ورودی یا پرسش هستند. این مدل‌ها می‌توانند در ایجاد محتوا، چت‌بات‌ها، ترجمه زبان و حتی کمک در تولید ایده‌ها کاربرد داشته باشند.
4. تقویت داده: مدل‌های مولد می‌توانند داده‌های مصنوعی را برای تقویت مجموعه داده‌های موجود ایجاد کنند. این می‌تواند منجر به بهبود عملکرد مدل‌های آموزش دیده بر روی داده‌های محدود شود.

5. شبیه‌سازی و بازی‌سازی: هوش مصنوعی مولد می‌تواند محیط‌های مجازی، شخصیت‌ها و سناریوها را برای شبیه‌سازی و بازی‌سازی ایجاد کند و تجربه‌های واقع‌گرایانه و تعاملی را فراهم کند.

2-5 بازیابی اطلاعات

بازیابی اطلاعات¹ فعالیت بدست آوردن منابع سیستم اطلاعاتی که مربوط به اطلاعات لازم از یک مجموعه است را می‌گویند. جستجوها می‌توانند برپایه جستجوی تمام متن یا سایر جستجوهای مبتنی بر محتوا نمایه‌گذاری شوند. بازیابی اطلاعات علم جستجو اطلاعات در یک سند، جستجو برای خود سندها، جستجو برای فراداده‌ها که داده‌ها را توصیف می‌کنند و برای پایگاه داده‌های متنی، عکسی یا آوایی است. سیستم‌های بازیابی اطلاعات خودکار برای کاهش چیزی که آن را سرریز داده‌ها (اضافه‌بار اطلاعات یا در واقع ریزش کاذب) می‌گویند، استفاده می‌شوند. سیستم بازیابی اطلاعات نرم‌افزاری است که دسترسی، ذخیره و مدیریت کتاب‌ها، مقالات و دیگر سندها را فراهم می‌کند. موتورهای جستجوگر وب (موتور جستجوی وب) از قابل مشاهده‌ترین برنامه‌های کاربردی بازیابی اطلاعات هستند. با افزایش روزافزون حجم اطلاعات ذخیره‌شده در منابع قابل دسترس و گوناگون، فرایند بازیابی و استخراج اطلاعات اهمیت ویژه‌ای یافته‌است. اطلاعات مورد نظر ممکن است شامل هر نوع منبعی مانند متن، تصویر، صوت و ویدئو باشد. برخلاف پایگاه داده‌ها، اطلاعات ذخیره‌شده در منابع اطلاعاتی بزرگ مانند وب و زیرمجموعه‌های آن مانند شبکه‌های اجتماعی از ساختار مشخصی پیروی نمی‌کنند و عموماً دارای معانی تعریف شده و مشخصی نیستند. هدف بازیابی اطلاعات در چنین شرایطی، کمک به کاربر برای یافتن اطلاعات موردنظر در انبوهی از اطلاعات ساختار نیافته‌است.

2-6 موتور جستجو²

در تعریف کلی، هر نرم‌افزار یا سخت‌افزاری که برای پیدا کردن داده‌های مشخصی، در میان حجم زیادی از اطلاعات جستجو کند را موتور جستجو می‌گویند. موتورهای جستجوگر اینترنت در حقیقت یک نرم‌افزار تحت وب هستند که وظیفه‌ی پیدا کردن مرتبط‌ترین محتوا با عبارت جستجو شده توسط کاربر را دارند. وظیفه‌ی هر موتور جستجو ارایه‌ی بهترین محتوا به کاربر با توجه به جستجوی انجام شده است. در حقیقت جستجوگرها برای سازماندهی هر چه بهتر محتوای اینترنت طراحی شده‌اند.

¹ Information Retrieval
² Search engine

2-7 گراف دانش¹

گراف دانش‌ها شبکه‌های معنایی بسیار بزرگی هستند که منابع اطلاعاتی ناهمگون را برای نمایش دانش در مورد حوزه‌های خاص گفتمان ادغام می‌کنند. در یک گراف دانش، دانش در یک گراف نمایش داده می‌شود تا به ماشین اجازه دهد تا از طریق استدلال و استنتاج، پاسخ‌های معناداری به پرسش‌ها ارائه دهد. استفاده ترکیبی از گراف دانش و مدل‌های یادگیری ماشینی می‌تواند سیستم‌های هوش مصنوعی را شفاف‌تر و قابل تفسیرتر کند. از گراف دانش‌ها می‌توان برای پاسخ به سوالات، درک تصاویر و بازیابی اطلاعات استفاده کرد که همگی جنبه‌های مرتبط بسیاری از انواع تحقیقات هستند. در سال‌های اخیر، گراف دانش‌ها به طور گسترده در حوزه‌های مختلف برای اهداف مختلف استفاده شده‌اند که شامل استخراج ویژگی‌ها، استخراج روابط، ساخت گراف دانش و استدلال مبتنی بر گراف دانش هستند.

1. استخراج ویژگی: بسیاری از روش‌های تعبیه دانش از گراف دانش برای استخراج ویژگی‌ها در حوزه‌های مختلف استفاده می‌کنند. به عنوان مثال، گراف دانش‌ها در تجزیه و تحلیل متن برای شناسایی موجودیت‌ها و ویژگی‌های یک متن استفاده می‌شوند، که برای ایجاد خلاصه‌های ساخت یافته معنایی به منظور غنی‌سازی نتایج جستجو استفاده می‌شود. گراف دانش‌ها همچنین برای ارائه پیوندهایی به نهادهای مرتبط، بهبود قابلیت‌های موتورهای جستجو و بهبود تجربه جستجوی کاربر استفاده می‌شوند.

2. استخراج روابط: یک گراف دانش می‌تواند جهت نمایش و ادغام دانش گذشته در استانداردهای مختلف استفاده شود. این رویکرد، روابط معنایی بین موجودیت‌های مختلف ایجاد می‌کند و اجازه می‌دهد تضادهای قابلیت همکاری بین استانداردهای ساز چارچوب‌ها و استانداردها حل شود. علاوه بر این، گراف دانش‌ها را می‌توان در موتورهای جستجوی یکپارچه به کار برد تا خلاصه‌ای از اطلاعات مرتبط در مورد پرس و جوهای مورد جستجو و فهرستی از موضوعات مرتبط را ارائه دهد.

3. ساخت گراف دانش: یک گراف دانش می‌تواند از مفاهیم مختلف (مانند داروها، علائم)، مترادف‌های مفاهیم، روابط سلسله مراتبی و انجمنی بین مفاهیم و نگاشت به مفاهیم دیگر در اصطلاحات خارجی که در سیستم‌های دانش محور مختلف استفاده می‌شود، تشکیل شده باشد. بنابراین، بسیاری از ناشران و ارائه دهندگان داده، گراف دانش‌ها را برای سازماندهی داده‌های خود، حاشیه نویسی و ارائه انواع مختلف اطلاعات به شیوه‌ای معنادار و افزودن برچسب‌ها یا برچسب‌های معنایی به مجموعه‌ای از اطلاعات، می‌سازند.

4. استدلال گراف دانش: از گراف دانش می‌توان در استدلال برای پیش‌بینی دقیق حقیقت یا عبارات بر اساس اطلاعات یا حقایق موجود استفاده کرد. حقایق و نتایج جدید را می‌توان از موجودیت‌ها، مفاهیم و روابط موجود در گراف دانش استنباط کرد.

¹ Knowledge Graph

2-8 پردازش تصویر¹

در حوزه هوش مصنوعی، پردازش تصویر به مجموعه‌ای از پردازش‌ها و عملیات اعمال شده بر روی تصاویر دیجیتالی اطلاق می‌شود. پردازش تصویر گام مهمی در پیش پردازش داده‌های تصویری است تا در نهایت از خروجی آن، در سایر پژوهش‌ها و پروژه‌های هوش مصنوعی نظیر تشخیص چهره، تشخیص اشیاء و فشرده‌سازی تصاویر استفاده شود. از پردازش تصویر می‌توان به منظور افزایش سطح کیفیت تصاویر دیجیتالی استفاده کرد. همچنین، الگوریتم‌های حوزه پردازش تصویر را می‌توان برای استخراج اطلاعات از داده‌های تصویری نیز به کار برد.

¹ Image Processing

3 دانايار؛ باهوش‌ترين همكار شما

فرض كنيد شخصي وجود دارد كه بيش از ۸۰ سال در تمامي دولت‌ها، يا وزارتخانه‌ها، يا سازمان‌ها حضور داشته است. اين فرد تمام تغييرات در دولت، در وزارتخانه مربوطه، در سازمان مربوطه را در طول اين سال‌ها ديده است: تمام مديراني كه آمده اند و رفته اند، تصميماتي كه گرفته شده، قوانيني كه وضع و اصلاح شده، برنامه‌هايي كه توسط دولت‌هاي مختلف تدوين شده است اما به دلایلي اجرايي نشده‌اند. تمامي اتفاقات و تغييراتي را كه در اين سال‌ها رخ داده است، در حافظه خود دارد و به عنوان مشاوري بي طرف و متخصص در تمام جلسات حضور دارد. از او در حوزه خاصي سوال مي‌پرسيد و او با توجه به تاريخچه، به پرسش‌هاي شما با دقت پاسخ مي‌دهد و شما را آگاه مي‌كند كه وضعيت كلي از گذشته تاكنون به چه شكل بوده است. چه اقداماتي انجام شده است و چه اقداماتي در اين حوزه به چه دلایلي به نتيجه نرسيده‌اند.

«دانايار» اين فرد خبره است كه هيچ‌گاه بازنشسته نمي‌شود از همه افراد با سابقه‌تر است، بر همه حوزه‌ها اشراف دارد و از هر انساني باهوش‌تر و سريع‌تر است. در جلسات شركت مي‌كند، چيزي را فراموش نمي‌كند و با ارجاعي دقيق به داده‌ها، كمك مي‌كند رئيس جمهور، وزرا و مديران به ديد بهتر و وسيع‌تري برسند. او در هيات دولت، مشاور ارشد بي طرف همه روساي جمهور است، و در وزارتخانه‌ها مشاور ارشد وزرا است.

«دانايار» يك عامل هوش مصنوعي است كه با بهره‌گيري از تكنولوژي‌هاي پردازش زبان طبيعي، OCR، بازيابي اطلاعات، تحليل كلان داده‌ها، پردازش تصوير و با استفاده از مدل‌هاي زباني بزرگ، با تشخيص و استخراج اطلاعات از انواع اسناد و داده‌ها، مانند فرم‌ها، رسيدها، صورتحساب‌ها، جداول و ... نقش يك مشاور آگاه و باسابقه را كه بر تاريخچه عملکرد دولت‌ها مسلط است، بازي مي‌كند. در چند سال اخير، هوش مصنوعي به يكي از پيشروترين فناوري‌ها در جهان تبديل شده و روند رشد آن به طور چشمگيري افزايش يافته است. حجم زياد داده‌ها و اطلاعاتي كه در يك كشور توليد مي‌شود، آنقدر زياد است كه حتي باهوش‌ترين انسان‌ها هم امكان مهار كردن كسر يك هزارم آن را هم ندارند تا چه برسد كه بتواند آن را تحليل كنند. به همين علت، تصميم‌هايي كه بهينه هستند گرفته نمي‌شود. اما هوش مصنوعي اكنون قادر است حجم زيادي از اسناد و داده‌ها را با سرعت و دقت بي‌نظيري تحليل كند، كه اين امر به ويژه در مقايسه با ظرفيت محدود ذهن انسان براي پردازش اطلاعات، بسيار چشمگير است. در حالي كه انسان‌ها در تحليل و درك حجم عظيم داده‌ها با محدوديت‌هاي ذاتي روبرو هستند، هوش مصنوعي مي‌تواند در كمترين زمان ممكن، تحليل‌هاي پيچيده‌اي را بر روي حجم زيادي از اسناد و داده‌ها انجام دهد.

«دانايار» به عنوان يك فناوري چند رشته‌اي، قادر است در زمينه‌هاي مختلفي متخصص باشد. در مقابل، يك انسان معمولاً نمي‌تواند در بيش از يك رشته تخصص داشته باشد و وزاري يك كشور نمي‌تواند به طور همزمان در تمامي زمينه‌هاي سياست‌گذاري و

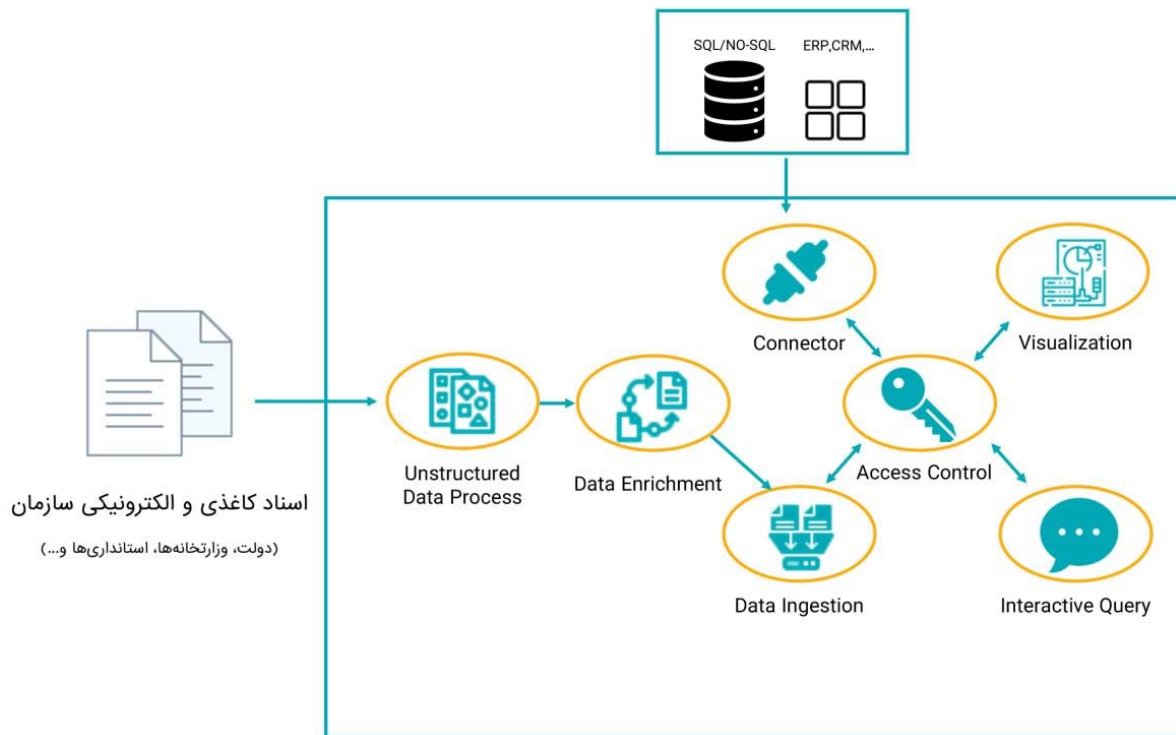
اجرائی موجود در کشور تخصص داشته باشند. این تفاوت اساسی نشان می‌دهد که هوش مصنوعی، با امکانات محاسباتی و الگوریتم‌های پیچیده‌ای که دارد، می‌تواند در زمینه‌های مختلف به طور همزمان و با کیفیت بالا عمل کند، در حالی که انسان‌ها به دلیل محدودیت‌های ذاتی، نمی‌توانند حتی یک هزارم آن را عمل کنند.

برای نمونه، وقتی هیات دولت بخواهد آیین نامه جدیدی تصویب کند، ضرورت دارد که اشراف کاملی بر آیین نامه‌های مرتبط و نتایج آنها داشته باشد. آیا این آیین نامه جدید با آیین نامه‌های قبلی سازگار است؟ همپوشانی دارد؟ تا چه اندازه؟ آیین نامه‌های قبلی چه میزان بهره‌وری را افزایش یا کاهش داده است؟ به چه علت؟ کفایت از دانايار پرسیده شود و دانايار نه تنها به زبان انسانی، بلکه با استفاده از نمودارها و همچنین ارجاع صحیح به آیین نامه‌ها و داده‌ها، پاسخ بدهد.

در سازمان برنامه و بودجه، دانايار، هنگام تدوین بودجه سالانه کشور، می‌تواند با تحلیل داده‌های گسترده و استفاده از الگوریتم‌های پیشرفته، الگوهای پیچیده‌تری را که به طور معمول توسط کارشناسان تدوین بودجه لحاظ نمیشوند، شناسایی کند. این امر به دولت‌ها کمک می‌کند تا از نظر مالی بهینه‌ترین برنامه‌ریزی را داشته باشند، هزینه‌های اضافی را کاهش دهند و استفاده بهینه‌تری از منابع مالی خود داشته باشند.

3-1 معماری دانايار

دانايار یک عامل هوش مصنوعی است که با پردازش و درک اسناد و داده‌ها، اطلاعات ساختاریافته، نیمه ساختاریافته و بدون ساختار را از اسناد و منابع داده‌ای گوناگون استخراج کرده و در حافظه خود نگاه می‌دارد. با تحلیل این اطلاعات، به عنوان همراهی که بر تمام اطلاعات موجود مسلط است، در سازمان حضور دارد و می‌تواند به هر سوالی از هر حوزه‌ای از سازمان با ارجاع دقیق به اطلاعات، پاسخ دهد. همانطور که در شکل 1 مشاهده می‌کنید دانايار از هفت ماژول اصلی تشکیل شده است.



شکل 1. معماری سیستم دانیار

در ادامه به شرح این ماژول‌ها می‌پردازیم.

3-1-1 ماژول پردازش داده‌های بدون ساختار (Unstructured Data Process)

این ماژول به جمع‌آوری اسناد الکترونیکی و کاغذی موجود در سازمان می‌پردازد. اسناد الکترونیکی متنی متنوعی مانند pdf، txt، xlsx، docx و غیره توسط این ماژول پردازش می‌شود. همچنین، اسناد تصویری و اسکن با استفاده از تکنولوژی OCR و پردازش تصویر که پیش از این توضیح داده شد، توسط این ماژول جمع‌آوری می‌شود. به این ترتیب، اسناد کاغذی و الکترونیکی موجود در سازمان با هدف ساخت متون قابل جست‌وجو و ویرایش توسط رایانه‌ها گردآوری می‌شود. اسناد سازمان مانند آیین‌نامه‌ها، بخش‌نامه‌ها، قراردادها، صورت‌حساب‌های مالی و غیره اسکن شده و به فرمت قابل فهم توسط کامپیوتر تبدیل می‌شوند. به این ترتیب، فایل تصویری مورد نظر تبدیل به فایل متنی شده و قابل استفاده برای ماژول‌های بعدی می‌باشد.

3-1-2 ماژول غنی‌سازی داده (Data Enrichment)

وظیفه ماژول Data Enrichment، غنی‌سازی داده‌های ورودی به صورت‌های مختلف است. این ماژول خود شامل تعدادی زیرماژول است که در ادامه شرح داده می‌شود:

زیرماژول دسته‌بندی و تگ‌گذاری اسناد

هدف از زیرماژول دسته‌بندی و برچسب‌گذاری اسناد به برچسب‌های مختلفی مانند آموزشی، اداری، امور انسانی، توسعه‌ای و غیره می‌باشد. برای پیاده‌سازی این ماژول از تکنیک Zero-Shot یا Few-Shot در مدل‌های زبانی بزرگ بهره گرفته می‌شود. یادگیری Few-Shot یک تکنیک مورد استفاده توسط LLM است که در آن برای دسته‌بندی مفاهیم و متون به کلاس‌های مختلف، ابتدا نمونه‌هایی از آن کلاس‌ها در طی فرآیند یادگیری به مدل داده می‌شود. در تکنیک Zero-Shot فرآیند دسته‌بندی بدون کمک از نمونه‌های متعلق به هر کلاس انجام می‌شود و خود مدل LLM به تشخیص دسته‌ها می‌پردازد.

زیرماژول شناسایی موجودیت‌های نام‌دار¹

شناسایی موجودیت‌های نام‌دار (NER) یک وظیفه پردازش زبان طبیعی (NLP) است که شامل شناسایی و طبقه‌بندی موجودیت‌های نام‌گذاری شده در متن به دسته‌های از پیش تعریف‌شده مانند افراد، سازمان‌ها، مکان‌ها، تاریخ‌ها و انواع دیگر است. نحوه کار این زیرماژول به این ترتیب است که ابتدا متن به کلمات یا عبارات منحصربفردی به نام «توکن» تقسیم می‌شود. سپس، از یک مدل Encoder-Decoder از قبل آموزش دیده برای شناسایی توکن‌های نام‌دار استفاده می‌شود. به علاوه، می‌توان برای این کار از مدل‌های Only-Decoder نیز استفاده کرد که دقت بالاتری دارند اما هزینه محاسباتی آنها به مراتب بیشتر خواهد بود.

زیرماژول استخراج کلمات کلیدی

استخراج کلمات کلیدی زیرشاخه‌ای از پردازش زبان طبیعی (NLP) است که بر شناسایی خودکار مهم‌ترین کلمات یا عبارات در یک متن تمرکز دارد. این کلمات کلیدی مفاهیم و ایده‌های اصلی بیان شده در متن را نشان می‌دهند. کلمات کلیدی فراداده‌های بسیار ارزشمندی هستند که از آنها در ماژول‌های دیگر سیستم «دانايار» استفاده می‌شود. نحوه کار این زیرماژول به ترتیب زیر است:

- پیش پردازش: اولین گام، پیش پردازش متن ورودی است که شامل وظیفه‌هایی مانند توکن‌گذاری، برچسب‌گذاری بخشی از گفتار² و تقسیم جمله است. این به LLM کمک می‌کند تا ساختار و دستور زبان متن را درک کند.
- ایجاد کلمات کلیدی کاندیدا: گام بعدی ایجاد لیستی از کلمات کلیدی یا عبارات کلیدی کاندیدا از متن از پیش پردازش شده است. این را می‌توان با استفاده از تکنیک‌های مختلفی مانند استخراج n-gram، فیلتر بخشی از گفتار و روش‌های مبتنی بر فرکانس و یا LLM انجام داد.
- امتیازدهی کلمات کلیدی: سپس از LLM برای امتیاز دادن به کلمات کلیدی کاندیدا بر اساس ارتباط و اهمیت آنها در متن داده شده استفاده می‌شود. این را می‌توان با تنظیم دقیق LLM روی مجموعه داده‌ای از کلمات کلیدی برچسب‌گذاری شده، یا با استفاده از توانایی ذاتی LLM برای درک و مفهوم‌سازی معنای کلمات انجام داد.

¹ Named-Entity Recognition

² part-of-speech tagging

- انتخاب کلمات کلیدی: در نهایت، کلمات کلیدی یا عبارات کلیدی با امتیاز بالا به عنوان خروجی نهایی فرآیند استخراج کلمه کلیدی انتخاب می شوند. تعداد کلمات کلیدی استخراج شده می تواند از پیش تعریف شده باشد یا بر اساس یک آستانه تعیین شود.

زیرماژول استخراج خلاصه سند

در سیستم دانیار، خلاصه سند می تواند کاربردهای مختلفی داشته باشد. یکی از مهم ترین کاربردها نمایش خلاصه ای از یک سند هنگام ارجاع به آن سند در خروجی تولید شده توسط دانیار است. همان طور که قبلا گفته شد، دانیار قابلیت ارجاع به سندها را دارد. به این ترتیب که در خروجی یک پرس و جو و یا در نمودار موجود در یک داشبورد، سند های مورد استفاده جهت تولید خروجی نمایش داده می شوند. کاربر می تواند با کلیک بر روی آن ارجاع به سند مدنظر دسترسی یابد. در اینجا مناسب است که علاوه بر نمایش کل سند، یک خلاصه از سند نیز نمایش داده شود.

یک کاربرد دیگر خلاصه سازی اسناد در بازایی اطلاعات است. در هنگام جستجو برای یافتن اسناد مرتبط با پرسش مطرح شده توسط کاربر، این خلاصه های اسناد مفید واقع خواهند بود. بدین منظور، علاوه بر ایجاد تعبیه¹ برای کل سند، برای خلاصه سند نیز تعبیه در نظر گرفته می شود.

به دو ترتیب می توان خلاصه سازی سند را انجام داد. یکی با استفاده از مدل های Encoder-Decoder و دیگری با استفاده از مدل Only-Decoder. بررسی و انتخاب مدل منبع باز مناسب برای این امر یکی از مهم ترین وظایف در این زیرماژول است.

زیرماژول همبسته سازی اسناد مبتنی بر مدل گرافی

می توان اسناد مختلف را بر اساس معیارهای مختلفی با یکدیگر همبسته نمود و ارتباطات بین آنها را ترسیم نمود. دو یا چند سند می توانند از جنبه های مختلف مانند زمان تدوین، نویسنده (گان)، دریافت کننده (گان)، نوع سند (مثلا اداری، امور انسانی و غیره)، سازمان تدوین کننده، سازمان های هدف و غیره با یکدیگر همبسته شوند. به علاوه، اسناد را می توان از لحاظ مفهومی نیز با یکدیگر همبسته سازی کرد. هم چنین، از آنجا که معمولا در اسناد به سند های دیگر ارجاع داده می شود، استخراج این ارجاعات نیز بخشی از وظایف این زیرماژول است.

پس از شناسایی و کشف این وابستگی ها، این ماژول یک آنتولوژی مفهومی بر روی این اسناد می سازد. خروجی این روابط کشف شده معمولا با یک مدل گرافی نشان داده می شود. گراف دانش با ارائه یک دید بصری عالی به کاربر این امکان را می دهد که اسناد مختلف را از جهات متنوع با یکدیگر ارتباط دهد. یال های بین گره ها در مدل گرافی میزان شدت یا ضعف همبستگی بین اسناد را نشان می دهد.

¹ Embedding

3-1-3 مازول دریافت داده (Data Ingestion)

وظیفه مازول Data Ingestion، ذخیره داده‌های ورودی به نحوی است که برای بازیابی اطلاعات بهینه باشد. پس از اینکه داده‌ها غنی شدند، آماده ذخیره‌سازی هستند. هر سند یا داده با توجه به نوع عملیات بازیابی اطلاعاتی که در آینده ممکن است بر روی آن صورت گیرد به یک یا چند روش از روش‌های زیر ذخیره می‌شود:

- پایگاه داده برداری: در این پایگاه داده‌ها ابتدا اسناد به بخش‌های کوچکی (Chunk) تقسیم شده و به ازای هر کدام از این بخش‌ها یک بردار عددی به کمک یک مدل زبانی تولید می‌شود و هر بخش در کنار بردار خود ذخیره می‌شود. این شیوه ذخیره‌سازی داده کمک می‌کند تا به صورت معنایی بر روی داده‌ها جستجو انجام شود.
- پایگاه داده گرافی: در پایگاه داده‌های گرافی از ساختار گراف برای نمایش روابط بین اشیاء و مفاهیم استفاده می‌شود. این ساختار شامل گره‌ها (اشیا) و یال‌ها (ارتباطات بین اشیاء) است. هر گره می‌تواند شامل اطلاعات مربوط به یک شیء مانند نام، توضیحات و ویژگی‌های آن باشد. یال‌ها نیز نوع و قدرت ارتباط بین دو گره را نشان می‌دهند. مزیت اصلی پایگاه داده‌های گرافی، توانایی مدل‌سازی روابط پیچیده بین اشیاء است.
- موتور جستجوی تمام متن¹: موتور جستجوی تمام متن، ابزاری برای جستجوی عبارات و کلمات در میان حجم عظیمی از متن است. این نوع موتورهای جستجو، به جای تکیه بر عناوین و کلمات کلیدی، کل محتوای متن را اسکن و بررسی می‌کنند و نتایج را علاوه بر کلمات کلیدی نیز بر اساس تناسب معنایی با عبارت جستجو شده ارائه می‌دهند.

این مازول قابلیت مدیریت نسخه را برای اسناد و داده‌ها فراهم می‌آورد. به این معنا که اسناد را می‌توان بعد از ذخیره کردن در این مازول بروزرسانی کرد. قابلیت مدیریت نسخه این اجازه را می‌دهد تا این تغییرات رصد شده و بتوان نسخه‌های مختلف از یک سند را با هم مقایسه کرد.

3-1-4 اتصال گر (Connector)

همان‌طور که شرح داده شد، سه مازول جمع‌آوری، غنی‌سازی و ذخیره‌سازی بر روی داده‌های بدون ساختار متمرکز هستند. در یک سازمان، علاوه بر اسنادی که در شاخه داده‌های بدون ساختار قرار می‌گیرند، حجم بسیار بالایی از داده‌ها در پایگاه داده‌های رابطه‌ای، پایگاه داده‌های غیررابطه‌ای، انبارهای داده، سامانه‌های ERP، CRM وجود دارند. نیاز است به طریقی به این داده‌ها متصل شده و دانش مورد نیاز را استخراج کرد.

¹ Full text search engine

ماژول اتصال گر قادر است به این Data source های مختلف متصل شود. اتصال گر ها عامل های مبتنی بر مدل های زبانی بزرگ¹ هستند که به شما اجازه می دهند اطلاعات را از منابع داده ای مختلف دریافت کنید.

3-1-5 کنترل دسترسی (Access Control)

کنترل دسترسی این امکان را می دهد که سطح دسترسی به افراد و گروه ها به فایل های خاص مشخص شود. حین فرآیند پرس و جو و یا نمایش داشبورد، ماژول کنترل دسترسی به طور فعال عمل می کند و بر روی خروجی ماژول های Connector و Data Ingestion نظارت کرده و سیاست های کنترل دسترسی تعیین شده را اعمال می کند. سیاست های کنترل دسترسی بر روی اسناد و فایل ها در یک سازمان، توسط مدیران میانی و ارشد سازمان تدوین می گردند.

3-1-6 ماژول پرس و جوی تعاملی (Interactive Query)

استفاده از LLM یا مدل های زبانی بزرگ، باعث می شود فرد بتواند با زبان طبیعی به پرس و جوی متنی و یا صوتی مدنظر بپردازد. ماژول پرس و جو یک Agent مبتنی بر LLM است که پس از دریافت سوال کاربر به زبان طبیعی، ابتدا این سوال را به تعدادی سوال دیگر تبدیل می کند. سپس پاسخ این سوالات را از ماژول های Connector و Data Ingestion به شکل تعدادی مستند دریافت می کند. ماژول پرس و جو بر اساس مستندات دریافتی از مرحله قبل، مجددا پرس و جوهای جدیدی تولید می کند. این فرآیند آنقدر تکرار می شود که سوال اولیه کاربر از تمامی جوانب مورد تحلیل و بررسی قرار گرفته و پاسخ مناسب به کاربر برگردانده شود.

پرس و جوی کاربر می تواند علاوه بر شکل نوشتاری (به زبان طبیعی) به صورت گفتاری نیز مطرح شود و ماژول پرس و جو قادر به پاسخ به صورت گفتاری می باشد.

3-1-7 ماژول مصورسازی (Visual Generator)

ماژول مصورسازی، وظیفه تبدیل داده های ورودی به نمودارها، چارت ها و ویژوال های دیگری که به صورت گرافیکی و تصویری نمایش داده می شوند، دارد. با دریافت داده ها از منابع مختلف و با استفاده از پردازش زبان طبیعی، این ماژول قادر است اطلاعات را تحلیل

¹ LLM data connectors

کرده و آن‌ها را به شکل قابل فهم و قابل دسترسی برای افراد سازمان تبدیل کند. با ایجاد یک داشبورد گرافیکی با تمامی ویژگی‌های مورد نیاز که توسط یک مدل‌های زبانی بزرگ ایجاد شده است، افراد سازمان می‌توانند تصمیم‌گیری‌های موثرتری انجام دهند. این ماژول با ارائه نمودارهای مختلف به افراد کمک می‌کند تا به سرعت اطلاعات را درک کرده و از طریق داده‌ها بهترین تصمیم‌ها را بگیرند. مشابه سایر ماژول‌های دانیار این ماژول نیز شامل سه زیرماژول Summarizer، Goal Explorer و Viz Generator است. زیرماژول Summarizer وظیفه تهیه خلاصه‌ای از داده‌ها (نوع داده‌ها و یا اطلاعات آماری مرتبط با داده‌ها) را دارد. زیرماژول Goal Explorer از این خلاصه استفاده کرده و چند نمونه از نمودارهایی را که مطابق با این مجموعه داده قابل ترسیم است به زبان طبیعی پیشنهاد می‌دهد و در نهایت زیرماژول Viz Generator وظیفه رسم نمودارها را دارد.

4 هزینه زمانی و مالی پروژه

جدول 1 زمان‌بندی توسعه و اجرای پروژه دانايار را نشان می‌دهد. همان‌طور که مشاهده می‌شود، توسعه محصول دانايار طی ده فاز (۱) مطالعات پژوهشی پیرامون طرح، (۲) فهم کسب‌وکار، (۳) آماده‌سازی محیط توسعه، (۴) پردازش داده‌های بدون ساختار، (۵) غنی‌سازی داده‌ها، (۶) ذخیره‌سازی داده‌ها، (۷) طراحی و پیاده‌سازی اتصال‌گر، (۸) طراحی و پیاده‌سازی ماژول Query، (۹) طراحی و پیاده‌سازی ماژول Visual Generator و (۱۰) طراحی و پیاده‌سازی ماژول Access Control انجام خواهد شد.

جدول 1. زمان‌بندی توسعه و اجرای محصول دانايار

شماره فاز	شرح فعالیت	ماه تقویمی																	
		۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱	۱۲	۱۳	۱۴	۱۵	۱۶	۱۷	۱۸
۱	مطالعات پژوهشی پیرامون طرح																		
	مطالعات پژوهشی مدل‌های زبانی بزرگ (LLM)																		
	مطالعه تکنیک‌های تزریق، ذخیره‌سازی و پرس‌وجوهای اسناد																		
	مطالعه تکنیک‌های NLP و بینایی ماشین																		
	ارزیابی تکنیک‌ها و انتخاب مدل‌های برتر																		
۲	فهم کسب‌وکار																		
	شناخت اسناد سازمانی																		
	شناخت سلسله مراتب سازمانی																		
	شناخت داده‌های بدون ساختار و داده‌های ساختارمند سازمان																		
۳	آماده‌سازی محیط توسعه																		
	تأمین و پیکربندی سخت‌افزار																		
	پیکربندی محیط توسعه، CI/CD و تست																		
	استقرار مدل‌های زبانی بزرگ (LLM) متن باز																		
۴	آماده‌سازی داده‌ها																		
	پردازش داده‌های بدون ساختار																		
	OCR																		
۵	تبدیل اسناد کاغذی به نسخه الکترونیکی																		
	غنی‌سازی داده‌ها																		
	دسته‌بندی و تگ‌گذاری اسناد																		
	استخراج موجودیت‌های نام‌دار و کلمات کلیدی																		
۶	استخراج خلاصه اسناد																		
	همبسته‌سازی اسناد با مدل گراف‌ی																		
	ذخیره‌سازی (Ingestion) داده‌ها																		
	پیاده‌سازی پایگاه داده برداری																		
	پیاده‌سازی پایگاه داده گراف‌ی																		

ماه تقویمی																		شرح فعالیت	شماره فاز
18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1		
																		پیکربندی موتور جستجوی تمام متن	
																		طراحی و پیاده‌سازی اتصال‌گر	7
																		طراحی و پیاده‌سازی اتصال‌گر برای SQL	
																		طراحی و پیاده‌سازی اتصال‌گر برای No-SQL	
																		طراحی و پیاده‌سازی اتصال‌گر برای سامانه‌های مختلف	
																		طراحی و پیاده‌سازی ماژول Query	
																		پیاده‌سازی ماژول پرس‌وجوی متنی مبتنی بر LLM	8
																		پیاده‌سازی ماژول پرس‌وجوی صوتی مبتنی بر LLM	
																		طراحی و پیاده‌سازی ماژول Visual Generator	9
																		پیاده‌سازی ماژول Summarizer	
																		پیاده‌سازی ماژول Goal Explorer	
																		پیاده‌سازی ماژول Viz Generator	
																		طراحی و پیاده‌سازی ماژول Access Control	10
																		شناسایی و استخراج سیاست‌های کنترل دسترسی سازمان	
																		اعمال سیاست‌های کنترل دسترسی به ماژول‌های مختلف	

بر اساس فازهای اصلی توسعه و اجرای محصول مطابق جدول 1 و پیچیدگی زمانی قابلیت‌های قابل توسعه، یک برآورد زمانی از هزینه توسعه و اجرای محصول صورت گرفته که جزئیات آن در جدول 2 ارائه شده است.

جدول 2. تخمین هزینه توسعه و اجرای پروژه دانايار

ردیف	نام فاز	گروه‌های کاری درگیر	هزینه زمانی ¹ (نفر ماه)
1	فاز اول: مطالعات پژوهشی پیرامون طرح	۲ نفر متخصص هوش مصنوعی ۲ نفر متخصص بینایی ماشین ۲ نفر متخصص NLP ۱ نفر مدیر پروژه	۵۰
2	فاز دوم: فهم کسب‌وکار	۲ نفر متخصص هوش مصنوعی ۲ نفر متخصص NLP ۴ نفر مشاور سازمان ۱ نفر مدیر پروژه	۴۹
3	فاز سوم: آماده‌سازی محیط توسعه	۲ نفر متخصص هوش مصنوعی ۲ نفر متخصص NLP ۲ نفر Backend Developer ۲ نفر DevOps ۱ نفر مدیر پروژه	۴۹
4	فاز چهارم: پردازش داده‌های بدون ساختار	۲ نفر متخصص هوش مصنوعی ۵ نفر متخصص بینایی ماشین ۲ نفر متخصص NLP ۲ نفر Frontend Developer ۲ نفر Backend Developer ۲ نفر DevOps ۱ نفر مدیر پروژه	۱۴۴
5	فاز پنجم: غنی‌سازی داده‌ها	۲ نفر متخصص هوش مصنوعی ۲ نفر متخصص NLP ۲ نفر Frontend Developer ۲ نفر Backend Developer ۲ نفر DevOps ۱ نفر مدیر پروژه	۵۹
6	فاز ششم: ذخیره‌سازی (Ingestion) داده‌ها	۲ نفر متخصص هوش مصنوعی ۲ نفر متخصص NLP ۲ نفر Frontend Developer ۲ نفر Backend Developer	۴۰

¹ این هزینه زمانی همراه با سربار محاسبه شده است.

	۲ نفر DevOps ۱ نفر مدیر پروژه		
۴۰	۲ نفر متخصص هوش مصنوعی ۲ نفر متخصص NLP ۲ نفر Frontend Developer ۲ نفر Backend Developer ۲ نفر DevOps ۱ نفر مدیر پروژه	۷ فاز هفتم: طراحی و پیاده‌سازی اتصال‌گر	
۵۹	۲ نفر متخصص هوش مصنوعی ۲ نفر متخصص NLP ۲ نفر Frontend Developer ۲ نفر Backend Developer ۲ نفر DevOps ۱ نفر مدیر پروژه	۸ فاز هشتم: طراحی و پیاده‌سازی ماژول Query	
۷۰	۲ نفر متخصص هوش مصنوعی ۲ نفر متخصص NLP ۲ نفر Frontend Developer ۲ نفر Backend Developer ۲ نفر DevOps ۲ نفر گرافیکست ۱ نفر مدیر پروژه	۹ فاز نهم: طراحی و پیاده‌سازی ماژول Visualization	
۴۹	۲ نفر متخصص هوش مصنوعی ۲ نفر متخصص NLP ۲ نفر Frontend Developer ۲ نفر Backend Developer ۲ نفر DevOps ۲ نفر متخصص امنیت ۱ نفر مدیر پروژه	۱۰ فاز دهم: طراحی و پیاده‌سازی ماژول Access Control	
۶۰۹		مجموع	