پیش نویس

پروژه درس هوش مصنوعی و سیستم های خبره

پاییز ۱۴۰۰

دانشگاه شهید بهشتی

مقدمه

آزمایشگاه پردازش زبان طبیعی دانشکده مهندسی و علوم کامپیوتر دانشگاه شهید بهشتی در حال توسعه سامانه ای به نام سها (سامانه هوشمند ارتباطات) به عنوان چت بات دانشکده و دستیار هوشمند دانشجویان است. تصمیم بر آن شده که پروژه این ترم هوش شکل ساده شده دو پیمانه از زیر سیستم پرسش و پاسخ سها باشد. به این ترتیب ضمن اینکه دانشجویان تجربه مشارکت در یک کار عملی واقعی را بدست می آورند، امکان بکارگیری آموخته های خود در طول ترم را خواهند داشت و ازسوی دیگر توسعه دهندگان پروژه های برتر می توانند جهت مشارکت در پروژه دعوت به همکاری شوند.

وظیفه سها گفتگو با کاربر، پاسخ به پرسشهای وی، انجام راهنمایی های لازم و یا ارجاع سوالات،انتقادات، پیشنهادات یا درخواستهای کاربر به مسئول مربوطه برای پاسخگویی است. بنابراین در یکی از ماجولهای کوچک لازم است سیستم، مخاطب یک گفته را تعیین نماید. و در یک ماجول دیگر به سوالات کاربر از جمله سوالات مکانی پاسخ دهد.

براین اساس دو پروژه برای درس به شرح زیر تعریف می گردد. هر پروژه در بخش اصلی دارای ۳ نمره بوده و می تواند تا ۱ نمره اضافه کسب نماید. هرکس می تواند یکی از دو پروژه را انتخاب کند و انجام دهد. کسانیکه هر دو پروژه را انجام دهند مشروط بر آنکه در یکی از پروژه ها حداقل ۷۰ درصد نمره را بگیرند می توانند تا سقف ۴.۵ از بخش اصلی پروژه نمره بگیرند.

گروههای پروژه تک نفری است.

پروژه اول : تعیین مخاطب گفتگو در سامانه سها

هدف از این پروژه تعیین مخاطب گفته کاربر است. برای ساده سازی فرض می کنیم ۵ دسته مخاطب زیر را داریم:

- **آموزش دانشکده (کد ۱) برای سوالاتی مثل "آخرین مهلت حذف چه موقع است؟، از کجا می تونم کارنامه مهر خورده بگیرم؟، شرایط حذف اضطراری چیست؟ سقف تعداد واحد ترم تابستون چنده؟...."**

- **میز اطلاعات (Information) (کد ۲) برای سوالاتی مثل "کتابخانه کجاست؟ کجا می توانم رییس دانشکده را ببینم؟ دانشکده پنجشنبه ها بازه؟ صبح ساعت چند در دانشکده باز می شه؟..."**
- **سایت/ کتابخانه (کد۳) برای سوالاتی مثل " چگونه می توانم روی سرور دانشکده اکانت بگیرم؟ روی کامپیوتر های سایت اوبونتو نصبه؟ اگر کتاب رو دیر بیارم چقدر جریمه میشم؟ اخرین چاپ کتاب هوش مال چه سالیه؟..."**
- **صندوق انتقادات و پیشنهادات (کد۴) برای درخواستها، انتقادات و پیشنهاداتی که باید توسط مسئولین دانشکده بررسی شود با ارسال گفته هایی مثل فلان استاد بد درس میده، لطفا تعداد دستشویی های در اختیار دانشجویان را در طبقه همکف زیاد کنید، نمیخواهید بالاخره کلاسها رو حضوری کنین؟ آسانسور خراب شده، لطفا رسیدگی شود....**
- **سایر (کد ۵) مواردی که در هیچکدام از دسته های بالا جا نمی شود.**
- **به هریک از نمونه ورودی های فوق یک "گفته" میگوییم که می تواند سوال یا جمله خبری باشد. فرض می کنیم هر گفته فقط حاوی یک جمله است و بین ۳ تا ۱۵ کلمه دارد.**
- **جملات می تونن هم رسمی و هم محاوره ای نوشته بشن. سعی کن بالانس مناسبی بین این و دسته ایجاد کنید و ترجیحا از هردو دسته به میزان مساوی داشته باشید.**

مرحله ۱) در این پروژه شما بایستی تا ۳ روز آینده به تعداد ۱۰۰ گفته همراه با برچسب دسته مناسب در یک فایل اکسل که ستون اول آن حاوی گفته ها و ستون دوم کد دسته است وارد کنید. در ستون سوم می توانید به عنوان کار اضافه یک دگر نویسی (پارافریز) گفته را نیز بیاورید. (دگرنویسی یعنی همان جمله را به بیان دیگر بنویسید) اکسل حاصله را به آدرس sbu.ai.course@gmail.com **ایمیل نمایید.**

دقت شود هر گفته فقط یکی از برچسبهای دسته(۱ تا۵) را داشته باشد

سپس تیم حل تمرین تمامی گفته های شما را به صورت یکپارچه در قالب یک فایل اکسل در می آورد و در کانال درسی در اختیار شما قرار میدهد.

مرحله ۲) شما باید ابتدا داده های خام را پردازش کنید تا اماده اموزش توسط الگوریتم های یادگیری ماشین شود. نرمال سازی متن فارسی و توکن بندی آن به کلمات از جمله پیش پردازش هایی است که باید انجام دهید. همچنین حذف داده های نادرست در این مرحله انجام می شود.

در این مرحله حذف گفته های تکراری، تشخیص تناقض در دسته بندی روی گفته های مشابه و افزایش تعداد گفته ها با روش های خودکار مشمول نمره اضافه است.

مرحله ۳) در این مرحله شما باید با حداقل ۲ الگوریتم یادگیری ماشین، به کلاس بندی داده ها بپردازید. یکی از این الگوریتم ها Naïve Bayes است که لازم است از پایه و بدون استفاده از کتابخانه های آماده، خودتان پیاده سازی کنید و برای الگوریتم (های) دیگر می توانید از کدهای آماده استفاده نمایید.

فراموش نکنید پیش از شروع اموزش، بخشی از داده ها را برای تست کنار بگذارید.

مرحله ۴) نتیجه کار را با معیارهای دقت، فراخوانی و معیار اف بسنجید، تست را بصورت ۳- فولد انجام دهید. (یعنی ۳ بار داده آموزش و آزمایش را بصورت تصادفی انتخاب کنید و بعد از ارزیابی بین نتایج میانگین بگیرید) نتایج به دست آمده را با یک دیگر مقایسه کرده و تحلیل کنید که دلیل اختلاف الگوریتم های مورد استفاده و دلایل خطای هریک چیست؟ این نتایج و تحلیل ها و شیوه کار خود را در قالب یک گزارش مختصر ارائه کنید.

پروژه دوم : پاسخ به سوالات مکانی

در این پروژه قرار است بخشی از پیمانه پرسش و پاسخ سها که مسئول پاسخگویی به پرسش های مکانی است را پیاده سازی کنید. به این منظور سیستم شما سوال کاربر را بصورت یک جمله سوالی دریافت می کند، اگر سوال در مورد مکانی نبود که متوقف میشود. در غیراینصورت به نوعی تشخیص می دهد سوال در مورد چه مکانی است و سپس براساس نقشه دانشکده کوتاهترین مسیر به آن مکان را تعیین می کند.

کسانیکه این پروژه را انجام میدهند از داده های تولید شده توسط گروه اول استفاده خواهند کرد ولی می توانند بعنوان نمره اضافه خودشان سوال مکانی نیز تولید نمایند.

انتخاب الگوریتم و شیوه تشخیص مکان مورد نظر کاربر از روی سوال بر عهده شماست و هرچه این روش بهتر و دقیقتر باشد نمره بیشتری کسب خواهد کرد. برای یافتن پاسخ به سوالات مکانی می توانید از طیفی از روشها از روشهای ساده مبتنی بر قاعده و انطباق الگو با کمک یک دیتابیس ساده تا روش های پیچیده یادگیری عمیق در سر دیگر طیف استفاده کنید. وقتی پاسخ کاربر را یافتید کوتاهترین مسیر از ورودی دانشکده تا مکان موردنظر را بعنوان خروجی برگردانید. انتخاب شیوه جستجو برعهده شماست استفاده از روشهای آگاهانه نمره بیشتری نسبت به روشهای کورکورانه دارد.

رسم گرافیکی نقشه و نمایش کوتاهترین مسیر روی آن نمره اضافه دارد.

مرحله ۱- نقشه دانشکده را دریافت و مسئله مسیریابی در آن را براساس روشهای تعریف مسئله آموزش داده شده در کلاس تعریف کنید (حالات، کنشها، هزینه ها، ...)

مرحله ۲- نمونه سوالات مکانی قابل پرسش از میز اطلاعات دانشکده را از گروه حل تمرین دریافت یا خودتان تولید کنید و آنها را پیش پردازش (نرمال سازی و توکن بندی) کنید. سوالاتی مثل کتابخانه کجاست؟ کجا می تونم معاون آموزشی را ببینم؟ درس هوش در کدام کلاس برگزار می شود؟ اتاق ۱۱۵ کدام طرف است؟ ...سوال مکانی بحساب می آیند. تولید سوال از قوانین نمره دهی مرحله اول پروژه ۱ تبعیت می کند یعنی با همان فرمت و داشتن نمره اضافی برای دگرنویسی ولی شرط تعداد وجود ندارد.

مرحله ۳- به ازای هر سوال، پاسخ آن را استخراج کنید. به این منظور می توانید از یکسری الگو کمک بگیرید و پس از فهمیدن اینکه سوال در مورد چیست پاسخ را از یک دیتابیس یا ساختمان داده اماده استخراج کنید (البته این صرفا پیشنهاد است). الگوریتم شما در این مرحله می تواند بسته به میزان پیچیدگی و هوشمندی نمره اضافه دریافت کند.

مرحله۴- نزدیکترین مسیر رسیدن به پاسخ یافت شده در مرحله قبل را بیابید (مبدا در ورودی دانشکده است). برای این منظور از حداقل ۲ الگوریتم جستجو استفاده کنید. پیاده سازی یکی از اینها باید از پایه و توسط شما انجام شود و برای دیگری می توانید از کدهای آماده استفاده کنید. (استفاده از الگوریتمهای آگاهانه نمره بیشتری دارد.)

مرحله ۵- نتایج را تحلیل کنید و نقاط قوت و ضعف روش (های) خود را شرح دهید. این نتایج و تحلیل ها و شیوه کار خود را در قالب یک گزارش مختصر ارائه کنید.

توضیح بیشتر پروژه دوم

شما یک جمله دریافت میکنید

- **۰- روی جمله پیش پردازشهای زبانی شامل نرمال سازی و توکن بندی را انجام میدهید. (برای این منظور می توانید از ابزارهای موجود مثلا هضم استفاده کنید.)**
- **۱- چک میکنید ایا سوال مکانی هست یا نه**
- **۲- اگر بود سوال را تحلیل میکنید که در مورد مکان چه کسی/چیزی سوال پرسیده**
- **۳- با جستجو کوتاهترین مسیر رسیدن به محل پاسخ را می یابید**

مثال ۱

سوال ورودی: نمازخونه کجاست؟

در خروجی باید به نوعی مسیر در ورودی به راه پله شرقی از راهروی سمت راست سپس حرکت از پله ها به زیرزمین و بعد حرکت به چپ را نشان دهید. نحوه نمایش شما براساس طرحتون می تونه متفاوت باشه. مثلا میتونید نقشه رو روی صفحه بیارید و این مسیر رو روش علامت بزنید. می تونید قدم های یک متری در نظر بگیرید و مثلا بنویسید گردش ۹۰ درجه به راست- مستقیم-مستقیم- گردش ۹۰ درجه به چپ- مستقیم مستقیم- مستقیم- ...- گردش ۹۰ درجه به چپ- مستقیم- مستقیم- ...- گردش ۹۰ درجه به چپ- حرکت از پله به پایین- ....

یا می توانید بجای گامهای یک متری حرکت به مکان بعدی را کنش درنظر بگیرید و مثلا بنویسید درورودی- کلاس ۱۰۱-کلاس ۱۰۲-کلاس ۱۰۳- ... (دم در کلاس منظوره) و ...

اینا چند نمونه طراحی بود که ممکنه خوب یا بد باشه. طراحی شما بخشی از نمره است و باید توجیهی براش داشته باشید و بسته به اون شکل نمایش خروجی عوض میشه.

فعالیتهای میانی برای رسیدن به خروجی در این مثال:

- **۱- سوال مکانی است؟ بله**
- **۲- محل مورد سوال: نمازخانه**
- **۳- جستجو برای کوتاهترین مسیر از درب ورودی به نمازخانه**

توجه: سوالات هم معنی این سوال مثل نمازخانه کجاست؟ یا کجا میتونم نماز بخوانم؟ هم باید همین نتیجه را داشته باشد. در مورد سوال کجا می تونم نماز بخونم که محل مورد سوال صراحتا در سوال نیست قبل از مسیریابی باید درقدم ۲ محل مورد سوال یعنی نمازخانه را بیابید و بعد مسیر را پیدا کنید

مثال۲؛

ورودی:از کجا بفهمم جواب اعتراض به نمره ام چیه؟

خروجی: سوال مکانی نیست (اتمام پردازش در مرحله ۱)

مثال ۳:

ورودی: محل جشن فارغ التحصیلی ۹۷ ایها؟

خروجی: مسیر در ورودی به امفی تاتر از راهروی راست (با همان توضیحات شیوه نمایش که در مثال ۱ دادم)

مثال ۴: میشه از محل بودجه ۱۴۰۰ لپ تاپ خرید؟

خروجی: سوال مکانی نیست (اتمام پردازش در مرحله ۱)

مثال ۵: رییس دانشکده را کجا می توان ملاقات کرد؟

خروجی: مسیر ورودی به دفتر ریاست

(در مرحله ۲ باید محل مورد نظر که دفتر ریاست هست را بیابید)

سوالات مشابه مثل دفتر رییس کجاست؟ اتاق رییس دانشکده کدوم طبقه است؟ رییس دانشکده کجاس؟ و ... به پاسخ مشابه برسند.

همانطور که دیده میشود این مسئله مراحلی دارد که هرکدام به تنهایی می توانند یک پروژه بزرگ باشند. در پروپوزال از شما خواسته می شود که بنویسید در هر مرحله چه می کنید و تمرکزتان روی کدام مرحله است؟

مثلا ممکن است شما در مرحله ۱ برای جداکردن سوالات مکانی از غیرمکانی از الگوریتم های رده بندی پیشرفته ای استفاده کنید و یا از یک مدل انطباق الگوی ساده بصورت قاعده بنیان (rul e based) بهره بگیرید. میزان پیچیدگی این مرحله نمره شما را تعیین می کند. به همین ترتیب برای مراحل بعدی

بصورت پیش فرض در این ۳ نمره هر مرحله ۱ نمره دارد. اگر یک مرحله را قویتر انجام دهید نمره ان اضافه می شود. بنابراین شما می توانید مثلا مرحله ۱ را با استفاده از رده بندی و یادگیری با داده مناسبی که تهیه کرده اید انجام دهید و بجایش اصلا مرحله ۳ را انجام ندهید و باز ۳ نمره بگیرید یا مرحله ۳ را هم ساده انجام دهید و ۴ نمره بگیرید. به همین ترتیب می توانید نمره اضافی را روی مرحله ۲ با پردازش انواع مختلف سوالات و افزایش سطح پوشش الگوریتمتان و یا بهره گیری از روشهای پیچیده تر از روش مبتنی بر کلیدواژه انجام دهید و نمره اش را دوبرابر کنید و یا در مرحله ۳ با افزودن خروجی گرافیکی یا بکارگیری

روشهای جستجوی آگاهانه نمره تان را افزایش دهید. هر مرحله ۱ نمره دارد که تا ۲ قابل افزایش است ولی سقف کل نمره حداکثر ۴.۵ است.

شکل ساده پروژه رسیدن به خروجی درست با هر الگوریتم دلخواه است. یعنی خروجی درست مستقل از روش انتخابی ۱ نمره دارد به شرط آنکه پوشش نسبتا مناسبی روی حالات مختلف ایجاد شود.