باسمه تعالی

پردازش زبان‌های طبیعی

تمرین دوم

ایجاد و به‌روزرسانی وظایف با عبارات منظم از مکالمات متنی

گروه ۴

پردیس زهرایی ۹۹۱۰۹۷۷۷

مهدی سعادت بخت ۹۹۱۰۵۴۷۵

محمدمهدی قیدی ۹۸۱۰۵۹۷۶

در این تمرین ما با کمک رجکس کلاس TaskExtractor را ایجاد کردیم. بخش های مختلف این کلاس به این شکل است.

استخراج نام اشخاص

برای این بخش یک دیتاست 2500 تایی از اسم های دختر و پسر زبان ایرانی شامل اسم های فارسی، ترکی، عربی را از سایت shadima.com کراول کردیم و با کمک آن اسم ها را استخراج می کنیم. برای اینکه بخش های بیشتری از اسم ها را شامل شویم و بتوانیم از فامیلی هم کمک بگیریم ولی تعداد فامیلی ها تقریبا نامحدود است، با کمک رجکس و استفاده از پیشوند و پسوند آنها را استخراج می کنیم مثلا بر اساس پیشوندهای خانم و آقا و دکتر و مهندس و ... می توانیم تمامی حالت های ممکن را برای اسم فرد استخراج کنیم. مثلا در مثال زیر:

"حتما حتما به المیرا و سوسن بگو که فایل ها رو کپی بگیرند و آقای علی مردانی و دکتر پردیس مومنی و آقای مهندس فراهانی هم تا شب فقط وقت دارند که زنگ بزنند و سوالات امتحان رو برای بچه ها آماده کنند و مهندس نیکبخت هم با آقای مهدی علیزاده باید هماهنگی جشن را انجام بدهند چون خانم لویزانی هم باید بادکنک ها را بیاورد و خانم دکتر محبی هم باید کیک درست کند و حال مریض را بپرسه"

اسم های زیر را استخراج می کند:

['المیرا', 'سوسن', 'آقای علی مردانی', 'دکتر پردیس مومنی', 'مهندس فراهانی', 'مهندس نیکبخت', 'آقای مهدی علیزاده', 'خانم لویزانی', 'دکتر محبی']

اگر نتواند اسمی را نیز استخراج کند، با توجه به فعل مخاطب را انتخاب می کند مثلا

'برو بقیه پروژه رو انجام بده'

اسم زیر را استخراج می کند:

['شنونده']

'اصلا من خودم باید اینکارو بکنم فقط'

اسم زیر را استخراج می کند:

['گوینده']

و اگر هیچ مخاطبی نداشت مثل زیر:

'هندونه خوشمزه است ها!'

به صورت زیر در آمده:

['نامعلوم']

استخراج وضعیت

با کمک چندین کلمه کلیدی مثل تمومه و تمام و پایان یافت و ... پایان را متوجه شده.

استخراج ضرورت

این بخش در صورت سوال نبوده و ما آن را خودمان اضافه کردیم که بر اساس کلمات کلیدی مثل فوری و فورا و ضروری و ... میزان تاکید و ضرورت انجام کار را پیدا می‌کند.

استخراج تاریخ

تکنیک‌ها و حالت‌های مختلفی برای این بخش در نظر گرفته شده است. تمام روزها، ماه‌ها، اسم‌های خاص (عید و محرم و...) و عددهای فارسی و انگلیسی و حالت‌های مختلف عدد مثل (1و یک و یکم و ۱) را در نظر می‌گیرد و با پسوند و پیشوند شروع و اتمام تسک تاریخ‌های مشخص شده را استخراج کرده
مثلا در : "این تسک را در ۲ آذر امسال شروع و تا 3 دی پایان رسونده." به صورت
['۲ آذر امسال']
['3 دی']
تشخیص داده شده و مثلا: "این تسک را از دوشنبه تا کریسمس وقت دارید که حل کنید."
['دوشنبه'] و ['کریسمس']
را تشخیص می دهد.
"این تسک را تا بیست و یکم تیر وقت دارید که حل کنید."
[]
['بیست و یکم تیر']
"لطفا این کار رو در سی و یکم خرداد شروع کنید و حتما تا پاییز کار تموم باشه دیگه."
['سی و یکم خرداد']
['پاییز']
"از امروز خداوکیلی تا کریسمس قراره که امتحان بدیم و سختی بکشیم فقط"
['امروز']
['کریسمس']
"بهش بگو از کیش تا قشم هم بره من زیر بار نمیرم ها"
[]
[]
در مثال بالا فانکشن پیاده شده توسط ما تشخیص داده که کیش و قشم تاریخ نیستند.

تغییر تاریخ و اسم

مثال های مختلف از تغییر تاریخ و اسم را هم تشخیص داده مثلا
"ددلاین به 6 مهر انتقال یافت"
را به درستی فهمیده یا
"مسئولیت تسک به مریم و محسن منتقل شد" همینطور.

تشخیص عنوان و زیروظایف

برای تشخیص عنوان از رجکس‌های مختلفی استفاده شده تا بتوان موارد متعددی را پوشش داد.
برای مثال اگر متن 'باید تسک حل تمرین دوم درس را در یک آذر شروع کنیم و تا ده آذر تمام کنیم.' را به عنوان ورودی به آن دهیم خروجی برابر خواهد بود با:
تسک حل تمرین دوم درس

برای تشخیص زیروظایف نیز از رجکس‌های مختلفی استفاده شده است. برای مثال اگر متن ورودی برابر باشد با:
برای اینکار باید اول موضوع را کنیم و بعد پیاده‌سازی را انجام دهیم.
خروجی برابر خواهد بود با:
('موضوع را مشخص کنیم', 'پیاده‌سازی را انجام دهیم')
که در حقیقت لیستی از زیروظایفی است که تشخیص داده است.

تغییر در عنوان

برای عوض کردن عنوان نیز از رجکس‌هایی استفاده شده که بتواند کلمات کلیدی مانند «تغییر کرد» و «تغییر یافت» و ...غیره را تشخیص دهد.