Natural language processing

نام: میلاد عزیزی شماره دانشجویی: ۴۰۲۱۵۴۱۵۰۶

استاد: دکتر محمدرضا جعفریانی

خرداد ماه سال ۱۴۰۳

دانشگاه صداوسیمای جمهوری اسلامی ایران

I. مقدمه

پردازش زبان طبیعی یا به اختصار NLP(Natural Language Process) یکی از شاخه های هوش مصنوعی است که به ماشین‌ها اجازه می‌دهد تا زبان انسان را درک کند و با ادراکی که برای ماشین صورت میگیرد بتواند آن را تفسر کند و برای مقاصد و کاربردهای مختلفی از آن بهره ببرد.

این فناوری از الگوریتم های پیچیده‌ای برای تحلیل و درک متون استفاده می‌کند، و موارد کاربردی زیادی دارد و در کارهای مختلفی می‌توان از آن بهره برد که در بخش هایی به تحلیل هر کدام از این وظایف که با استفاده از NLP میتوانیم انجام دهیم می‌پدازیم که از جمله آن ها عبارتند از

- Information Retrieval
- Keywords Extraction
- Information Extraction
- Dialogue Systems
- Sentiment Analysis
- Text Classification
- Token Classification
- Table Questions Answering
- Question Answering
- Zero-Shot Classification
- Translation
- Summarization
- Feature Extraction
- Text Generation
- Text2Text Generation
- Fill-Mask
- Sentence Similarity

حال به اختصار هر کدام از موارد بالا را توضیح خواهیم داد:

II. *Information Retrieval*

بازیابی اطلاعات، در این task ما قرار است که یک موضوع یا query را برای جستوجو داشته باشیم و به کمک یک ابزار بتوانیم اطلاعات مربوط را از دنیایی از اطلاعات استخراج کنیم و از حجم زیادی از مستندات به تعداد انگشت شماری از مستندات برسیم که این کاربرد در گوگل کاربرد زیادی دارد.

III. *Summerization*

خلاصه‌سازی، این مورد را داخل search engine ها زیاد می‌بینیم که خلاصه ای از متون را در صفحه اصلی search برای ما نشان می‌دهد که اصطلاحا به اسم google snippet هم شناخته می‌شوند

IV. Keyword Extraction

به خلاصه سازی نزدیک است با این تفاوت که در خلاصه‌سازی ما جملات کلیدی را استخراج می‌کردیم اما در keyword extraction ما کلمات کلیدی را استخراج می‌کنیم

V. Information Extraction

در اینجا ما پا را فراتر می‌گذاریم و اطلاعات ساختار یافته را از متن استخراج می‌کنیم. مثلا در ویکی‌پدیا فقط اطلاعات تاریخ تولد و کتاب نوشته شده، مکان تولد و ..... را از یک شخص بخواهیم استخراج کنیم، این اطلاعات در یک جدول مستطیلی در ویکی‌پدیا در قالب info box موجود است که به کاربر کمک می‌کند تا اطلاعات مهم را استخراج کند.

VI. Question Answering

بازیابی اطلاعات به ما کمک می‌کند تا اطلاعات را از حجم زیادی از اسناد استخراج کنیم. اما باز در خیلی از کاربری که عمل search را انجام می‌دهد مایل است اطلاعات خیلی مختصر و کلیدی از داده های متنی استخراج کند. مثلا اطلاعات مربوط به محل تولد یک شخص را بخواهیم داشته باشیم، اگر نام آن شخص را search کنیم قاعدتا به صفحات متنوع و زیادی می‌رسیم که باید از آن ها اطلاعات موردنظرمان را استخراج کنیم. که این کار قاعدتا زمان بر است. اما سیستم های پرسش و پاسخ این قابلیت را دارند که به صورت خودکار نه تنها اطلاعات را در قالب صفحات مهم استخراج کنند بلکه اطلاعات موردنظر را از آن صفحات هم استخراج کند، و در آخر به کاربر تنها یک پاسخ کلیدی بدهد.

VII. Dialogue systems

شبیه سیستم پرسش و پاسخ است، سوالی از سیستم می‌پرسیم و انتظار داریم سیستم پاسخگوی ما باشد. در دایالوگ سیستم‌ها بر خلاف سیستم پرسش و پاسخ تعامل ۲ طرفه خواهد بود، یعنی کاربر یک سوال از سیستم می‌پرسد و سیستم در پاسخ به آن سوال، سوال دیگری را مطرح می‌کند، در واقع یک مکالمه بین کاربر و سیستم رخ می‌دهد. و کاربرد آن در chatbot ها و helpcenter ها خواهد بود.

VIII. Sentiment Analysis

تحلیل نظرات، هدف تحلیل یک متن و استخراج اینکه آیا نظر مثبتی در این متن نهفته است یا نظر منفی، مثلا در review که کاربران در صفحات مختلف مثل هتل، رستوران و کالاهای مختلف در online shapping اختصاص داده شده‌اند. در یک نوشته طولانی از کاربر با استفاده از sentiment analysis بتواند جنبه های مهم این نظر را استخراج کند و برای هر جنبه مشخص کند که این نظر حس مثبتی داشته یا نه و با استفاده از آن تصمیم گیری درستی را انجام دهد.

IX. Text Clssification

یکی از مهم‌ترین و کاربردی ترین مسائل در حوزه های NLP است. که این فرآیند شامل طبقه‌بندی یک متن (مانند سند، جمله و ....) به یکی از دسته‌های از پیش تعریف شده بر اساس محتوا یا موضوع آن است. از کاربردهای آن میتوان به دسته بندی ایمیل، طبقه بندی مقالات و اخبار به موضوعاتی مانند سیاست، ورزش، اقتصاد و ...، تشخیص نوع متن(مقاله، داستانی، شعر و ...) که برای این کار از الگوریتم های یادگیری ماشین مانند SVM, Decision Tres, neural networks, KNN می‌توان استفاده کرد

X. Token Classification

یکی از کارهای مهمی که در پردازش زبان طبیعی صورت می‌گیرد بحث token classification است که هدف از انجام این کار دسته‌بندی توکن‌های موجود(کلمات، عبارات و کاراکترها) در یک متن به دسته های مشخص شده می‌باشد. این وظیفه می‌تواند در کاربردهای متنوعی مانند تحلیل متن، درک زبان طبیعی و بازیابی اطلاعات استفاده شود. یکی از رایج ترین توکن کلاس بندی NER(Named Entity Recongintion) است که هدف از انجام آن این است که موجودیت های مهم یعنی همان اسم های خاص را در متن مانند نام‌های افراد، مکان، سازمان‌ها، تاریخ‌ها و ... را شناسایی و دسته بندی کند.برای

XI. Table Question Answering

این روش به کاربران کمک می‌کند تا اطلاعات موردنیاز خود را از جداول استخراج کنند. و به طور گسترده‌ای در کاربردهایی نظیر تحلیل داده‌ها و سیستم های پرسش و پاسخ استفاده می شود.

روش TQA یا همان پاسخگویی به پرسش های جدول به معنای استخراج و ارائه پاسخ دقیق به سوالاتی است که به یک جدول یا چند جدول داده مرتبط هستند. این پرسش‌ها می‌توانند پیچیده باشند و نیازمند این باشند که درک عمیقی از ساختار جدول و روابط بین داده های مختلف داشته باشیم. که از روش های متداول آن می‌توان به روش Rule-Based-Methods و که برای سوالات ساده‌ای کاربرد دارند چون به صورت دستی تنظیم می‌شوند و manual هستند.

XII. Zero-Shot Classification

به معنای طبقه‌بندی داده ها به کلاس‌هایی است که مدل هرگز قبل از این فرآیند آن‌ها را ندیده است. برخلاف روش‌های سنتی که نیاز به داده‌های آموزشی برای هر کلاس دارند. در این رویکرد مدل از داده‌های متنی عمومی و دانش قبلی خود برای انجام این طبقه‌بندی استفاده می‌کند. که در این روش از embedding کردن کلمات در مدل های از پیش آموزش داده شده مثل chat GPT استفاده می‌کند تا ویژگی های برداری برای متن ورودی و کلاس ایجاد کنند و سپس کلاس‌ها به جملات تبدیل می‌شوند که مدل یادگیری آن را به خوبی درک کند.

این روش معمولا در سناریو هایی به کار می‌رود که تنوع کلاس‌ها بالا هستند و تهیه داده های آموزشی به طور کامل برای هر کلاس ممکن نباشد. این روش از قدرت مدل‌های زبانی پیش‌آموزش دیده و توانایی آن ها در فهم معنای عمومی متن استفاده می‌کند.

XIII. Translation

یک فرآیندی است که به طور خودکار یک متن را ای یک زبان به زبان دیگر تبدیل می‌کند. که این موضوع شامل مفاهیم، تکنیک ها و روش‌های بسیاری است که برای ترجمه خودکار استفاده می‌شود. در این task با استفاده از مدل‌های یادگیری عمیق و شبکه‌های عصبی پیشرفته، به دقت و کیفیت بالاتری می‌توان دست یافت و باعث شده این حوزه همواره در حال بهبود و پیشرفت باشد.

XIV. Feature Extraction

در اینجا هدف استخراج و انتخاب اطلاعات مفید از داده های متنی است. این ویژگی ها به عنوان ورودی به مدل‌های یادگیری ماشین و یادگیری عمیق استفاده می‌شود. در واقع استخراج ویژگی ها فرآیندی است که طی آن، اطلاعات خام متنی به مجموعه‌ای از ویژگی‌های قابل استفاده برای مدل‌های یادگیری تبدیل می‌شود. هدف این است که

نمایشی از داده‌های متنی ایجاد شود که مدل بتواند بهتر آن ها را تحلیل و پردازش کند. این روش باعث بهبود عملکرد مدل می‌شود و نقش مهمی را در به دست آوردن نتایج دقیق‌تر ایفا می‌کند.

XV. Text Generation

تولید متن به معنای ایجاد متنی جدید به صورت خودکار که از نظر گرامری و معنایی صحیح و منطقی باشد. مدل های تولید متن می‌توانند در کاربردهای مختلفی مانند تکمیل جمله، پاسخ به سوالات، خلاصه سازی، تولید داستان، و حتی تولید کدهای برنامه نویسی استفاده شوند.

XVI. Text2Text Generation

این مدل به معنای تبدیل یک متن ورودی به یک متن خروجی مرتبط و از نظر معنایی معنی‌دار است. این مدل‌ها با در نظر گرفتن متن ورودی، متن جدیدی را تولید می‌کنند که ممکن است ترجمه،خلاصه، جواب به یک سوال، بازنویسی و یا حتی تولید یک متنی باشد که به صورت خلاقانه طرح شده باشد

XVII. Fill-Mask

مدل های Fill-Mask که به معنای Maked Language Modeling نیز شناخته می‌شوند به این صورت عمل می‌کنند که بخشی از جمله (معمولا یک یا چند جمله) با یک توکن خاص مانند mask جایگزین می‌شوند و سپس مدل سعی می‌کند کلمات mask شده را بر اساس متن باقی مانده پیش‌بینی کند.

XVIII. Sentence Similarity

این مدل به معنای تعیین میزان شباهت بین دو جلمه بر اساس محتوا و مفهوم آن‌ها است. این شباهت می‌تواند در مقیاس عددی، به صورت درصد یا به عنوان یک طبقه‌بندی دودویی(مشابه بودن یا مشابه نبودن) نشان داده شود.