تحلیل داده های متنی
گزارش1 هوش مصنوعی دکتر قطعی

امیررضا رادجو
دانشجوی علوم کامپیوتر دانشگاه صنعتی امیرکبیر

11 اسفند 99

تحلیل داده‌ی متنی

متن كاوی و داده كاوي

داده كاوی روشي بسيار كارا براي كشف اطلاعات از داده هاي ساخت يافته است. متن كاوي مشابه داده كاوي است، اما ابزارهاي داده كاوي طراحي شده اند تا داده هاي ساخت يافته از پايگاه داده را به كار ببرند. ميتوان گفت، متن كاوي راه حل بهتری براي شركتها است. پس تفاوت متن كاوي و داده كاوي اين است كه داده كاوي بر روي داده هاي ساخت يافته پايگاه داده كار مي كند و متن كاوي، بر روي داده هاي غير ساخت يافته و نيم ساخت يافته مانند Email و مستندات تمام متني كار مي كند. در متن كاوي سعي مي گردد از همان تكنيكهاي داده كاوي استفاده گردد. براي اين منظور به تكنولوژيهايي ديگري مانند پردازش زبان طبيعي، يادگيري ماشين و ... نياز است تا به صورت اتوماتيك آمارهايي را جمع آوري نموده و ساختار و معنای مناسبي از متن استخراج گردد. در اين موارد، ديدگاه عمومي استخراج ويژگيهاي كليدي از متن است. ويژگيهاي استخراج شده بعنوان داده براي تحليل استفاده مي گردد.

متن كاوی و بازيابي اطلاعات

معمولاً در بازيابي اطلاعات با توجه به نياز مطرح شده از سوي كاربر، مرتبط ترين متون و مستندات و يا در واقع «كيسه كلمه» از ميان ديگر مستندات يك مجموعه بيرون كشيده ميشود. بازيابي اطلاعات يافتن دانش نيست بلكه تنها آن مستنداتي را كه مرتبط تر به نياز اطلاعاتي جستجوگر تشخيص داده به او تحويل ميدهد. اين روش در واقع هيچ دانش و حتي هيچ اطلاعاتي را به ارمغان نمي آورد.
متن كاوي ربطي به جستجوي كلمات كليدي در وب ندارد. اين عمل در حوزه بازيابي اطلاعات گنجانده مي شود. به عبارتی بازيابی اطلاعات جستجو، كاوش، طبقه بندي و فيلتر نمودن اطلاعاتي است كه در حال حاضر شناخته شده اند و در متن قرار داده شده است. ولی در متن كاوی مجموعه اي از مستندات بررسي شده و اطلاعاتي كه در هيچيك از مستندات به صورت مجرد يا صريح وجود ندارد، استخراج مي گردد.

پردازش زبان طبيعي يا NLP

هدف كلي آن رسيدن به يك درك بهتر از زبان طبيعي توسط كامپيوترهاست. تكنيك هاي مستحكم و ساده اي را براي پردازش سريع متن به كار مي برد. همچنين از تكنيكهاي آناليز زبان شناسي نيز براي پردازش متن استفاده مي كند. نقش NLP در متن كاوي فراهم كردن يك سيستم در مرحله استخراج اطلاعات با داده هاي زباني است.

.در اینجا یک آشنایی مروری با مباحث موجود در پردازش زبان‌های طبیعی می‌پردازیم

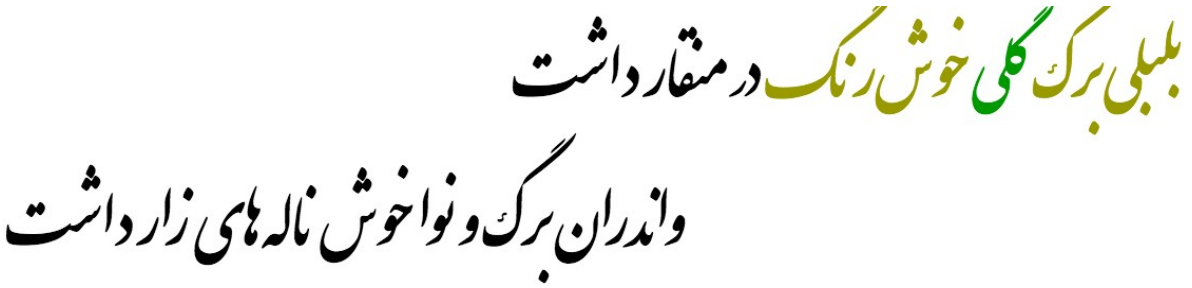
به طور کلی می‌توان گفت که **زبان** از لحاظ مفهومی نسبت به **تصویر** دارای پیچیدگی‌های متعددی است که فهم و پردازش آن توسط هوش مصنوعی را سخت‌تر می‌کند. از جمله‌ی این ویژگی‌ها می‌توان به پیچیدگی‌های حافظه‌ای و شبکه‌های معنایی اشاره کرد که در پردازش تصویر معمولاً چنین پیچیدگی‌هایی یافت نمی‌شود.

با ظهور یادگیری ژرف در فضای یادگیری ماشین، ایده‌های جدیدی در زمینه‌ی مسائل حوزه پردازش زبان‌های طبیعی نیز به وجود آمد که به مرور اجمالی آن‌ها می‌پردازیم.

واژه‌نهفت‌ها (*word embeddings*)

پس از برخورد با داده‌های متنی برای استفاده از تکنیک‌های یادگیری ژرف ابتدا لازم است تا نمایشی از آن‌ها در فضای برداری ایجاد کنیم. به همین منظور در سال ۲۰۱۳ توسط آقای میکولوف و همکاران روشی برای نشاندن واژگان در فضای چند بعدی ارائه کردند.

ایده‌ی اصلی این روش برای هر کلمه استفاده از کلماتی است که در نزدیکی آن قرار دارند. برای مثال در شعر زیر:



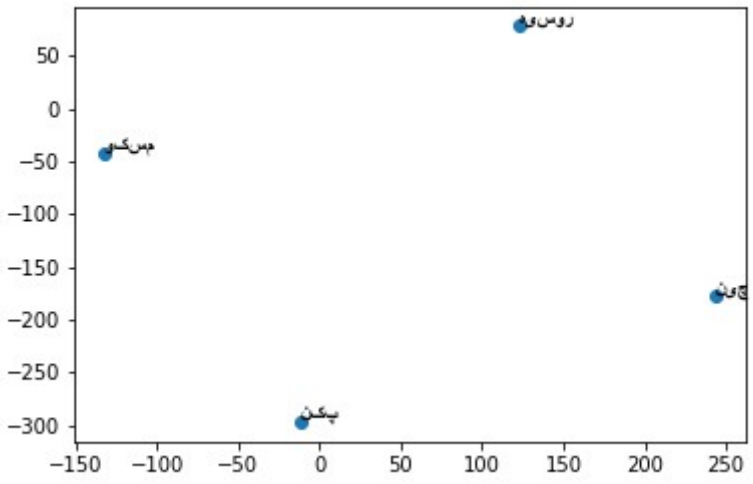
واژه «گل» در همسایگی واژگان «برگ» و «بلبل» و «خوش رنگ» آمده‌است. در این روش به نوعی سعی بر آن است تا با استفاده از فراوانی کلمات موجود در همسایگی یک کلمه شبکه‌ی معنایی آن را تخمین بزنند. این روش به کلمه به بردار (*word2vec*) نیز معروف است.

چند ویژگی جالب در این روش مشاهده شده‌است.

واژگانی که معنای نزدیک به یکدیگر دارند به جهت آن‌که در همسایگی آن‌ها واژگان مشترک با احتمال خوبی به تعداد زیادی یافت می‌شوند، نزدیک به یکدیگر قرار می‌گیرند.

ویژگی دیگر این است که اختلاف بردارهایی که ارتباط معنایی مشخصی با یکدیگر دارند تا حد قابل توجهی به یکدیگر شبیه هستند. برای مثال مشاهده شده که بردارهای کشورها به پایتخت آن‌ها تا حد قابل توجهی به یکدیگر شبیه اند.

برای مثال به شکل زیر توجه کنید



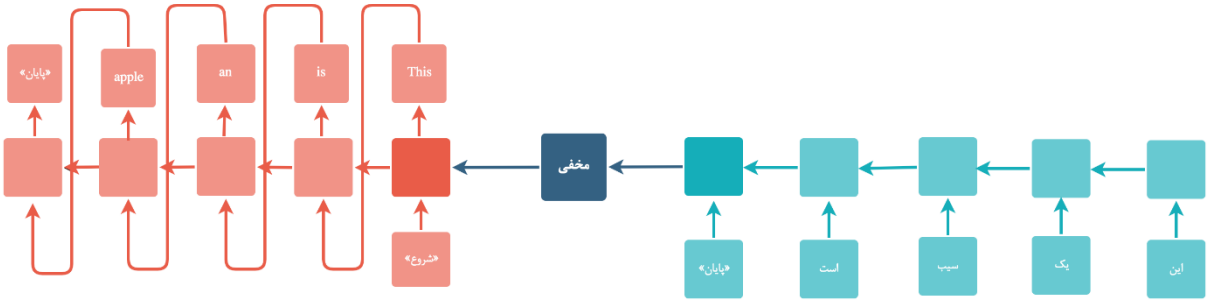
این ایده سنگ بنای بسیاری از ایده‌های آتی در زمینه یادگیری ماشین حوزه پردازش زبان‌های طبیعی است

ترجمه ماشینی و دنباله به دنباله

ترجمه ماشینی تبدیل خودکار جملات از یک زبان به زبان دیگر است که تا پیش از به کارگیری یادگیری ژرف در زمینه پردازش زبان‌های طبیعی به صورت آماری انجام می‌گرفت. اما با به وجود آمدن روش‌های نوین یادگیری ماشین و استفاده از واژه‌نهفت‌ها راه برای به وجود آمدن ایده‌ی ترجمه ماشینی عصبی (*Neural Machine Translation*) و به طور کلی‌تر دنباله به دنباله (*Sequence to Sequence*) هموار شد.

در روش دنباله به دنباله از دو شبکه‌ی عصبی بازگشتی (شبکه عصبی بازگشتی) نوعی شبکه عصبی است که یکی برای کدگذاری و دیگری برای کدگشایی داده‌ها استفاده *LSTM* حاوی گره‌هایی با یال به خود است) مانند می‌شود.

در این معماری ابتدا متن توسط واژه‌نهفت‌ها تبدیل به تعدادی بردار می‌شود. سپس این بردار به ترتیب وارد کدگذار شده و پس از پردازش، وضعیت مخفی (وضعیت فعلی شبکه عصبی بازگشتی پس از ورود واژه‌نهفت‌ها به کدگذار) به عنوان ورودی به کدگشا داده می‌شود و کدگشا پس از دریافت وضعیت مخفی شروع به تولید دنباله خروجی می‌کند. نحوه کارکرد آن را می‌توانید در شکل زیر مشاهده کنید:



مدل‌های از پیش‌آموزش‌دیده

اخیراً استفاده از مدل‌های از پیش‌آموزش‌دیده (مدل‌هایی که وزن‌های آن‌ها را طی فرایند آموزش ذخیره می‌کنند و در زمان اجرا از آن وزن‌ها استفاده می‌کنند) به شکل گسترده‌ای مورد استفاده قرار می‌گیرند.

یکی از معماری‌هایی است که از کدگذارهای تبدیل‌گر برای ایجاد یک مدل زبانی *BERT* برای مثال معماری دوطرفه استفاده کرده‌است. این مدل از پیش‌آموزش‌دیده بر روی مسأله‌های مختلف به خوبی جواب داده‌است. نیاز دارد. تدقیق به آموزش مدل موجود با (*fine-tuning*) همچنین برای مسائل بسیار دیگری تنها به تدقیق تعداد محدودی داده در فضای مسأله جدید گفته می‌شود که سبب می‌شود تا مدل از پیش‌آموزش‌دیده با اندکی تغییر در وزن‌هایش برای مسأله جدید قابل استفاده باشد.

:References

/https://quera.ir/college

https://www.coursera.org/specializations/deep-learning

https://virgool.io/@sadeqhasan

/https://www.kaggle.com/learn/natural-language-processing