**متن کاوی**

اجزای یک متن و موجودیت‌های یک زبان عبارت از کلمات، عبارات و جملات کامل هستند و اگر این اجزا به صورت مفهومی و با ارتباط معنی دار در کنار هم قرار بگیرند به آن یک سند متنی گفته می‌شود. بخش قابل توجهی از اطلاعات قابل دسترس در پایگاه داده‌های متنی ذخیره شده‌اند که این پایگاه داده‌ها شامل مجموعه بزرگی از اسناد و منابع مختلف مانند مقالات خبری، علمی، کتاب‌ها، ایمیل‌ها و صفحات وب هستند. پایگاه‌های داده و اطلاعات موجود به فرم الکترونیکی سریعا در حال رشد بوده و امروزه بیشتر اطلاعات موجود در صنعت، کسب و کار و سایر سازمان‌ها از این نوع هستند.

یکی از مهمترین حوزه‌های تحقیقاتی پردازش زبان طبیعی، متن کاوی بوده که به معنای کشف اطلاعات جدید توسط کامپیوتر و استخراج خودکار آنها است. متن کاوی حوزه‌ای تحقیقاتی بوده که خود ترکیب چند فیلد تحقیقاتی دیگر است. در متن کاوی نوشته‌ها و متون تحلیل شده، دانش و الگوهای با ارزشی که در آنان مخفی هستند یافته شده و استخراج می‌شوند. متن کاوی شامل کشف ارتباط بین واژه‌ها و جملات، طبقه بندی متون و خلاصه سازی هستند.این حوزه‌ها شامل مواردی مانند داده کاوی، پردازش زبان طبیعی و بازیابی اطلاعات هستند. اما بین داده کاوی و متن کاوی تفاوت هایی وجود دارند که مهمترین آنها عبارتند از:

داده کاوی معمولا با داده‌های ساخت یافته سر و کار دارد اما برعکس متن کاوی با داده های ساخت نیافته یا شبه ساخت یافته درگیر بوده، خصوصا زمانی که این کاوش درون یک مقاله یا سند متنی عمومی انجام شود.

حتی اگر کمی ساخت یافتگی هم در متن وجود داشته باشد باز هم دلیلی دیگر وجود دارد که متن کاوی را بسیار مشکل تر از داده کاوی می‌کند. این دلیل انتزاعی بودن مفاهیم درون متن است که به سختی می‌توان آنها را با ساختارهای ارائه دانش مرسوم مدل کرد.

در متن کاوی با مشکلات دیگری مانند کلمات هم خانواده (کلماتی که دارای معنی مشابه اما دیکته و تلفظ متفاوت) و کلمات متشابه (کلماتی که دیکته یا تلفظ مشابه داشته اما معنی کاملا متفاوت دارند) در گیر بوده که کار با متن کاوی را به مراتب مشکلتر نسبت به داده کاوی می‌کند .

**کاربردهای متن کاوی**

با توجه به تعاریف گسترده‌ای از متن کاوی وجود دارد، تظریات متفاوتی نیز در مورد کاربردهای آن ارائه شده است. با اینکه متن کاوی یک فیلد تحقیقاتی جدید بوده اما نرم افزارهای آنالیز داده متنی از اواخر سال 1990 در دسترس بوده‌اند. از جمله متداول‌ترین کاربردهای متن کاوی می‌توان به موتورهای جستجو اشاره کرد که در آنان کاربر یک عبارت یا کلمه را حتی با غلط املایی تایپ کرده و موتور جستجو با وجود حجم بسیار بزرگ اطلاعات موجود در وب مرتبط ترین متون را یافته و لیست می‌کند. مهمترین کاربردهای دیگر متن کاوی موارد زیر هستند:

- **شناسایی اسپم:** آنالیز عنوان و محتوای ایمیل برای تشخیص جهت تشخیص اسپم.
- **نظارت:** نظارت کردن رفتار شخص یا گروهی از انسان‌ها به صورت پنهان. نرم افزارهایی وجود دارند که میتوانند تلفن، اینترنت و دیگر وسایل ارتباطی را برای شناسایی رفتار افراد کنترل کنند.
- **شناسایی نام‌های مستعار**
- **روابط میان مفاهیم:** از جمله واقعیت‌هایی که می‌توان از یک مجموعه متون دریافت، ارتباط و وابستگی برخی مفاهیم با مفاهیم دیگراست. به این صورت که پدیدار شدن برخی کلمات منجر به ظهور کلمات دیگری شده که با آن وابستگی شدید دارند. منظور این است که هرگاه مجموعه ای از یک سری واژه‌ها دیده شوند، انتظار می‌رود که مجموعه‌ای دیگر از لغات نیز پس از آنان دیده شوند .
- **جستجو و بازیابی:** هدف بازیابی اطلاعات این هدف است که از میان مجموعه‌ای از متون، آنهایی را که به نیاز اطلاعاتی کاربر ارتباط بیشتری داشته، جدا کرده و به کاربر نشان دهند. همیشه شرایطی وجود دارد که کاربر نتواند به نیاز اطلاعاتی خود از میان متن های موجود دسترسی پیدا کند. شرایط متفاوتی از قبیل تعداد زیاد اسناد مرتبط، تفاوت فاحش در نوع نگرش آنها به موضوع و مسائلی از این قبیل می‌توانند این چنین کاوش‌ها را به صورت دستی بسیار مشکل کنند. اگرچه این نوع از سیستم‌های بازیابی اطلاعات از تکنولوژی هایNLP و یادگیری ماشین استفاده می‌کنند اما آنچه در نهایت در این سیستم‌ها، نقش اصلی را دارا است، پایگاه دانشی بوده که از طریق روش‌های مبتنی بر NLP و یا روشهای آماری بر روی مدارک موجود در مجموعه ساخته شده است.