هدف از علم متن‌کاوی، استخراج اطلاعات از متن است. برای این کار از رایانه‌ها و نرم‌افزار استفاده می‌شود. زیرا در اکثر مواقع حجم متن زیادتر از آن که بتوان این کار را به چند نفر سپرد و البته استفاده از منابع انسانی برای خواندن و استخراج اطلاعات زیاد است و زمان زیادی لازم دارد. از سوی دیگر صحت‌سنجی عملکرد منابع انسانی نیز دشوار است. در علم متن‌کاوی سعی می‌شود تا با انجام محاسبات بر روی متن، اطلاعات مورد نیاز از متن استخراج شود. برای انجام محاسبه بر روی متن لازم است تا اجزای متن به‌صورت عددی نمایش داده‌ شوند. به‌طور مثال به‌جای اینکه کلمه Mine با چهار حرف M، i، n و e نمایش داده شود، از یک معادل عددی همانند 13 برای نمایش آن استفاده شود. حرف M یا کلمه Mine، هیچ معنای خاصی برای پردازنده ندارد البته در این مثال اختصاص یک کد مثل 13 به کلمه Mine هم مشکل خاصی را حل نمی‌کند و شاید فقط کمی محاسبات آماری را ساده‌تر کند. اما اگر به‌جای عدد 13، یک کد معنادار به کلمه Mine نسبت داده می‌شد، ارزش پردازشی بیشتری به‌وجود می‌آمد. به‌طور نمونه اگر این عدد میزان اهمیت این واژه در یک متن را نمایش می‌داد، به راحتی همه کلمات یک کتاب بر حسب عدد میزان اهمیت مرتب شده و ده واژه مهم‌تر کتاب استخراج می‌شد. ده واژه مهم یک متن، بلافاصله درک معنی‌داری از آن متن فراهم می‌کنند. اما انتظار بیشتری که وجود داشت، امکان نمایش عددی یک عبارت به نحوی بود که معنا و مفهوم عبارت هم در آن نهفته باشد، چیزی که شاید دو دهه قبل یک رویا به حساب می‌آمد. اما امروزه این رویا به حقیقت پیوسته است.

در حدود سال های 2000 تا 2010، از روش‌های آماری برای محاسباتی متنی استفاده می‌شد که به الگوریتم‌های زیر می‌توان به عنوان الگوریتم‌های معروف در این حوزه اشاره نمود:

Term Frequency-Inverse Document Frequency (TF-IDF)

C-Value

NC-Value

N-Grams

Bag of Words (BoW)

این روش‌های آماری در داده‌های یکدست و برای استخراج اطلاعات ساده از متن، عملکرد قابل قبولی داشتند. به طور نمونه روش TF-IDF میزان اهمیت یک واژه در متن را مشخص می‌کند. روش‌های آماری همیشه دقیق نبوده و در مواقعی بر روی داده‌های متنوع متنی در دنیای واقعی، عملکرد مناسبی ندارند. به‌طور قطع، یک عدد ساده مثل 13 نمی‌توانست تمامی ویژگی‌های معنایی و مفهومی یک واژه را بیان کند. از دهه دوم قرن بیستم، این ایده مطرح شد که برای نمایش هر یک از اجزای متن (مثلا کلمات) از بردار استفاده شود. اما منظور از بردار و اجزای متن چیست؟

**اجزای متن:**

منظور از اجزای متن تقریبا همان کلمات است اما در متن یک حرف مستقل نیز واحد متنی حساب شده و عبارت‌های متشکل از چند کلمه نیز یک واحد (Token) محسوب می‌شوند.

They use ServiceTitan as a HVAC real-time monitoring software

به طور نمونه جمله فوق از نه واحد تشکیل یافته است که برخی یک حرفی و برخی دو کلمه‌ای هستند.

**بردار:**

مجموعه‌ای از چندین عدد است که هر یک از این اعداد، یک بُعد از بردار محسوب می‌شوند. یک بردار 10 بعدی می‌تواند به این صورت باشد:

[0.12, 0.23, 0.13, 0.82, 0.02, 0.74, 0.55, 0.84, 0.72, 0.17]

برای انجام محاسبات بر روی متن، لازم بود تا هر واحد متنی به‌صورت یک بردار نمایش داده شود. در ابتدا روش‌هایی همچون مدل فضای بردار یا (VSM) Vector Space Model برای ساخت بردار از واحدهای متنی ابداع شدند. در این روش اندازه بردار معادل تعداد کلمات یکتا در متن بود. به طور مثال یک متن 1000 کلمه‌ای ممکن است از 300 کلمه یکتا ساخت شده باشد و برخی از چند کلمات چند بار در متن تکرار شده باشند. در همین متنی که شما اکنون مشغول خواندن آن هستند، واژه‌ی متن، بارها تکرار شده است اما برای تهیه فهرست واژه‌های یکتای این نوشتار، کلمه متن فقط یکبار شمرده می‌شود. مشکل اصلی روش‌های اولیه این بود که اندازه بردارها دقیقا با متناسب با تعداد واحد‌های متنی یکتا بود و با افزایش حجم متن، اندازه بردارها نیز افزایش پیدا می‌کرد و ممکن بود اندازه یک بردار به چند صدهزار بعد نیز برسد. در این زمان روش‌های ریاضی مختلفی برای کاهش ابعاد بردار ابداع شدند که روش Random indexing یکی از معروف‌ترین آنهاست.

یکی از مفاهیم اصلی حوزه متن کاوی، Word embedding است که در فارسی به آن جاسازی کلمه گفته می‌شود. جاسازی کلمه همان مفهوم نمایش یک واحد متنی در قالب بردار است. در سال 2013 شرکت گوگل روش word2vec را معرفی نمود که در آن از شبکه‌های عصبی و نوعی هوش مصنوعی برای ساخت بردارهای متنی استفاده می‌شد. روش word2vec را می‌توان یک ربات هوش مصنوعی فرض کرد که با خواندن متن و یادگیری آن، ذهنیت و درک خود را در قالب بردار بیان می‌کند. بدیهی است که یادگیری و اظهار نظر دقیق در مورد کلیه موضوعات یک رشته تخصصی با خواندن چند صفحه متن تخصصی، ممکن نیست. به همین دلیل ربات هوش مصنوعی نیز برای یادگیری صحیح و ساخت بردارهای دقیق به متنی در حدود چند میلیون کلمه نیاز دارد. بردارهای word2vec در اندازه‌های مختلفی قابل ساخت هستند اما اندازه‌هایی بین 50 تا 300 بعد، رایج‌تر هستند.

word2vec مجموعه‌ای از الگوریتم‌ها است که معرف‌ترین آنها Skip-gram و CBoW هستند.

دقت بالای بردارهای ساخته شده توسط روش Word2vec در مقالات مختلفی مورد بحث قرار گرفته است که به طور نمونه به دو مورد از آنها در انتهای این متن اشاره شده است. بردار ساخت شده برای هر واحد متنی، حاوی اطلاعات معنایی و مفهومی از آن واحد متنی است. در یک آزمایش واقعی برای چهار کلمه King، Man، Women و Queen با استفاده از word2vec بردارهای لازم ساخته شد. اگر هر بردار را با نام کلمه متناظر آن نشان دهیم، از بردار King، بردار Man کم شده و بردار Woman با آن جمع شد. نتیجه حاصل، بسیار نزدیک به بردار Queen بود:

King-Man+Woman ≈ Queen

بنابراین یکی از مهم‌ترین ابزارهای مورد استفاده در متن کاوی، استفاده از مشابهت بردارها، برای یافتن ارتباط بین واحد‌های متنی است. به این مفهوم که اگر دو بردار شبیه هم باشند، می‌توان نتیجه گرفت که واحدهای متنی متناظر آنها به یک مفهوم مشابه اشاره دارند. روش شباهت کسینوسی، یکی از روش‌های رایج برای محاسبه شباهت دو بردار است. اگر برای مجموعه‌ای از عبارت‌های فناوری، بردارهای لازم ساخته شود، برداری مشابه، فناوری‌های مرتبط را نشان می‌دهد.