*پردازش زبان طبیعی *(NLP)* به‌عنوان توانایی سیستم‌هایی تعریف می‌شود که می‌توانند زبان انسانی، شامل گفتار و متن، را تحلیل، درک و تولید کنند. *NLP* بخشی از زبان‌شناسی محاسباتی *(مطالعه زبان‌شناسی با استفاده از علوم کامپیوتر)* است و با توجه به اینکه اکثر داده‌های تولیدشده به صورت داده‌های بدون ساختار هستند، *NLP* ابزاری قدرتمند برای تفسیر و درک زبان طبیعی محسوب می‌شود.

*چندین اصطلاح وجود دارد که قبل از ادامه باید آن‌ها را درک کرد :

توکن‌سازی و توکن‌ها *(Tokenization & Tokens)*:این به معنای شکستن یک متن بزرگ به قطعات کوچک‌تر به نام توکن‌ها است.تصور کنید یک پاراگراف به جملات تقسیم شود، جملات به کلمات، یا حتی واحدهای کوچک‌تری مانند حروف.شیء متن*(text object)* یک شیء متن می‌تواند هر چیزی باشد، از یک کلمه، یک جمله، یک پاراگراف، یا حتی یک مقاله کامل.ریشه‌یابی و ریشه‌یابی کلمات *(Stem & Stemming)* این روشی برای ساده‌سازی کلمات با حذف پسوندهای رایج مانند «-ly» ،«ing» یا «-s» است.این روش همیشه دقیق نیست، اما به گروه‌بندی کلماتی با معانی مشابه کمک می‌کند.واژه‌بنیادی‌سازی *(Lemmatization)*برخلاف ریشه‌یابی، واژه‌بنیادی‌سازی به دنبال ریشه واقعی کلمه *(یا شکل واژه‌نامه‌ای آن)* با استفاده از قوانین و تحلیل زبانی است.این روش نتایج دقیق‌تری نسبت به ریشه‌یابی ارائه می‌دهد.واج‌معنا *(Morpheme)*واج‌معنا کوچک‌ترین بخش یک کلمه است که دارای معنا است.نحو *(Syntax)*نحو مربوط به چگونگی چیدمان کلمات برای ساخت جملات معنادار است.این مانند قوانین دستوری است که کلمات را در ترتیب صحیح فاعل ،مفعول و فعل قرار می‌دهد.معناشناسی و کاربردشناسی *(Semantics and Pragmatics)*:معناشناسی درباره معنای کلمات و چگونگی ترکیب آن‌ها برای ساخت عبارات یا جملات معنادار است. این علم بر روی معنای کلمات تمرکز دارد.کاربردشناسی مربوط به نحوه استفاده و تفسیر جملات در موقعیت‌های واقعی است. این علم لحن، زمینه و چگونگی تغییر معنا در شرایط مختلف را در نظر می‌گیرد. کاربردشناسی کمتر به معنای لغوی کلمات و بیشتر به معنای مورد نظر گوینده می‌پردازد.معناشناسی کمک می‌کند بفهمیم چه چیزی گفته شده است، در حالی که کاربردشناسی کمک می‌کند بفهمیم چرا این جمله گفته شده است.

***مقدمه‌ای بر پردازش زبان طبیعی *(NLP)*:برای پردازش جمله ورودی و دریافت خروجی مفید از یک مدل *NLP*، چندین مرحله کلیدی وجود دارد : *lexical analysis , syntatic analysis , semantic analysis*

1. **تحلیل واژگانی *(Lexical Analysis)***:در هر مجموعه داده، یک متن که بافت مرتبط ندارد، به عنوان نویز در نظر گرفته می‌شود. اولین مرحله در *NLP* تمیز کردن و استانداردسازی متن ورودی برای حذف نویز و آماده‌سازی برای تحلیل است.که شامل چند مرحله میباشد :

الف. حذف نویز *(Noise Removal)*:این مرحله شامل تهیه یک دیکشنری از کلمات نویزی *(مانند: the, a, of, this, that)* است.متن پردازش شده و کلماتی که در این دیکشنری نویز قرار دارند، حذف می‌شوند.

ب. نرمال‌سازی واژگان *(Lexicon Normalization)*:کلماتی مانند *follow, following, followed, follower* تغییرات مختلف یک واژه هستند.هدف نرمال‌سازی کاهش پیچیدگی از طریق:1*(ریشه‌یابی (Stemming)*: حذف پسوندها و پیشوندها.2*(واژه‌بنیادی‌سازی (Lemmatization)*: استفاده از ساختار و روابط دستوری کلمات برای یافتن ریشه دقیق‌تر است.

ج. الگوریتم *Porter Stemmer*:*(کتابخانه NLTK و متد PorterStemmer)* الگوریتم محبوبی برای بهبود بازیابی اطلاعات با حذف انتهای کلمات *(پسوندها)* و کاهش آن‌ها به ریشه مشترک میباشد.مثال :کلمات *troubled, trouble, troubling* به ریشه *trouble* کاهش داده می‌شوند.مزیت آن این است که کلمات مشابه به یک ریشه مشترک گروه‌بندی شده و نمایش آماری دقیق‌تری از تعداد تکرار یک کلمه خاص به وجود میاید اما گاهی ممکن است معنای کلمه از دست برود.

د.استانداردسازی شیء *(Object Standardization)*:متن ممکن است شامل کلماتی باشد که در دیکشنری واژگان موجود نیستند، مانند اصطلاحات عامیانه، اختصارات، هشتگ‌ها، یا کلمات خاص رسانه‌های اجتماعی *(مثلاً: RT، DM‌ing)*.این کلمات می‌توانند با استفاده از دیکشنری‌های آماده یا الگوهای منظم *(regular expressions)* حذف یا استانداردسازی شوند.

2. **تحلیل نحوی *(Syntactic Analysis)***:برای تحلیل متن، لازم است که متن به ویژگی‌ها تبدیل شود. تحلیل نحوی شامل بررسی جملات برای درک روابط بین کلمات و اختصاص یک ساختار نحوی به متن است. چندین الگوریتم برای تحلیل نحوی وجود دارد، اما گرامر مستقل از متن *(Context-Free Grammar)* به دلیل سادگی و استفاده گسترده، محبوب‌ترین است.در این جمله باید فاعل *(subject)*، مفعول‌ها *(objects)*، نویز *(noise)*، و ویژگی‌ها *(attributes)* را شناسایی کنیم تا ترتیب کلمات و وابستگی‌های آن‌ها را درک کنیم.

الف)تحلیل وابستگی *(Dependency Parsing)*(کتابخانه *NLTK* و متد *StanfordDependencyParser)* : گرامر پایه‌ای روابط یا وابستگی‌ها بین اجزای ساختار را مشخص می‌کند.تحلیل وابستگی این روابط را در قالب یک ساختار درختی نمایش می‌دهد که گرامر و ترتیب کلمات را نشان می‌دهد.گرامر وابستگی روابط دوتایی نامتقارن بین توکن‌ها را تحلیل می‌کند.

ب)برچسب‌گذاری اجزای سخن *(Part of Speech Tagging))*(کتابخانه *NLTK* و متد *word_tokenize)*: شامل اتصال هر کلمه یا توکن در جمله به یک برچسب اجزای سخن *(POS)* است. این برچسب‌ها شامل اسم‌ها، افعال، صفات، قیدها، اعداد و غیره هستند.کاربردها:1. ساخت درخت‌های تجزیه *(Parse Trees)* برای درک ساختار جمله2.تحلیل احساسات *(3 .Sentiment Analysis)*. پاسخ به سوالات. 4. شناسایی موجودیت‌های مشابه میباشد.

ج)شناسایی ویژگی‌ها و نرمال‌سازی*(Normalization)* :با شناسایی انواع سخن در کنار زمینه‌های مختلف یک کلمه، برچسب‌های *POS* به تمایز کمک کرده و ویژگی‌های قوی‌تری ایجاد می‌کنند.برچسب‌های *POS* پایه‌ای برای نرمال‌سازی *(Normalization)* و واژه‌بنیادی‌سازی *(Lemmatization)* هستند تا ساختار جملات و وابستگی‌ها بهتر درک شوند.

د) حذف کلمات *(Stopword Removal)*:برچسب‌های *POS* برای حذف کلمات پرتکرار یا غیرضروری *(مانند: the، a، of)* از متن بسیار مفید هستند.

3. **تحلیل معنایی *(Semantic Analysis)*** : تحلیل معنایی پیچیده‌ترین مرحله در پردازش زبان طبیعی *(NLP)* است. این مرحله به استخراج معنای دقیق یا معنای واژگانی متن می‌پردازد. با استفاده از دانش مربوط به ساختار کلمات و جملات در بافت متن، معنای کلمات، عبارات، جملات و متون مشخص می‌شود و به دنبال آن هدف و پیامدهای آن‌ها نیز تحلیل می‌گردد.

***پس از پیش‌پردازش داده، تحلیل واژگانی، نحوی و معنایی، لازم است که **متن به بازنمایی‌های ریاضی** تبدیل شود تا بتوان آن را ارزیابی، مقایسه و بازیابی کرد. برای تبدیل متن به زبان ریاضی از تکنیک های مختلفی می توان استفاد کرد.تکنیک‌هایی مانند *N-grams*، بردارهای *TF-IDF*، تحلیل معنایی پنهان، شباهت کسینوسی، طبقه‌بندی‌کننده بیزی، و الگوریتم‌های ژنتیک ابزارهای قدرتمندی برای پردازش زبان طبیعی، جستجو، و طبقه‌بندی هستند. هر کدام از این تکنیک‌ها با توجه به نیازهای خاص کاربردی انتخاب می‌شوند :

1) *N-grams (کتابخانه NLTK و متد ngrams)*: ترتیب *N* کلمه از متن ورودی را حفظ می‌کند. و در اصلاح املایی ، تجزیه کلمات ، خلاصه‌سازی متون کاربرد دارد.
2) بردارهای *TF-IDF* : اسناد به‌صورت بردارهای با ابعاد بالا نمایش داده می‌شوند که هر ورودی نمایانگر کلمه‌ای در سند و تعداد وقوع آن است.وزن هر کلمه بر اساس تعداد تکرار کلمه ،تعداد کل اسناد و تعداد اسنادی که کلمه حداقل یک‌بار در آن‌ها وجود دارد محاسبه می‌شود.بردارهای *TF-IDF* برای یافتن اسناد مرتبط با یک کلمه خاص استفاده می‌شوند.

3) تحلیل معنایی پنهان *(Latent Semantic Analysis)*: در مجموعه‌های بزرگی از اسناد، تحلیل معنایی پنهان روابط بین کلمات مرتبط را شناسایی می‌کند.برای مثال، اگر سندی شامل کلمه "sea" باشد، احتمالاً شامل کلمه "beach" نیز خواهد بود.این روش روابط معنایی بین کلمات و اهمیت آن‌ها در سند را شناسایی می‌کند.

4) شباهت کسینوسی *(Cosine Similarity)*(کتابخانه SciPy و متد cosine) : شباهت کسینوسی شباهت بین دو بردار را اندازه‌گیری می‌کند و می‌تواند برای مقایسه اسناد یا تطابق بین سند و یک جستجو استفاده شود.هرچه زاویه بین دو بردار بزرگ‌تر باشد، شباهت کم‌تر است.می‌توان بردارهای TF-IDF را با شباهت کسینوسی ترکیب کرد تا شباهت بین یک پرسش و مجموعه‌ای از اسناد را محاسبه کرد.

5) طبقه‌بندی‌کننده بیزی ساده *(Naive Bayesian Classifier)*: این روش مبتنی بر قضیه بیز است و برای ورودی‌هایی با ابعاد بالا مناسب است. با وجود سادگی، در بسیاری از موارد بهتر از روش‌های پیچیده‌تر عمل می‌کند.احتمال تعلق یک نمونه به کلاس خاص را محاسبه می‌کند. در طبقه‌بندی اسناد بر اساس احتمال تعلق آن‌ها به کلاس مورد نظر کاربرد دارد.

* **الگوریتم‌های ژنتیک *(Genetic Algorithms)*:** این الگوریتم‌ها با الهام از تکامل طبیعی برای کاهش نرخ خطا طراحی شده‌اند. مانند شبکه‌های عصبی که از مغز انسان الهام گرفته‌اند، الگوریتم‌های ژنتیک تلاش می‌کنند عملکرد کروموزوم‌ها را شبیه‌سازی کنند.این الگوریتم انتخاب طبیعی را شبیه‌سازی می‌کند، جایی که بهترین ویژگی‌ها به نسل‌های بعد منتقل می‌شوند تا مسائل حل شوند یا راه‌حل‌های بهینه پیدا شوند.
اما این الگوریتم چگونه کار میکند ؟

1. **مقداردهی اولیه *(Initialization)***:یک جمعیت تصادفی از راه‌حل‌های ممکن *(به نام "افراد" یا "individuals")* ایجاد می‌شود.هر فرد یک پاسخ بالقوه است که به‌صورت رشته‌ای از "ژن‌ها" *(مانند اعداد یا مقادیر باینری)* کدگذاری شده است
2. **ارزیابی شایستگی *(Fitness Evaluation)***:میزان "خوبی" هر فرد در حل مسئله اندازه‌گیری می‌شود که به آن امتیاز شایستگی *(fitness score)* می‌گویند.مثال: اگر مسئله پیدا کردن کوتاه‌ترین مسیر بین شهرها باشد، امتیاز شایستگی می‌تواند طول مسیر باشد *(هرچه کوتاه‌تر، بهتر)*
3. **انتخاب *(Selection)***:"شایسته‌ترین" افراد *(بهترین راه‌حل‌های فعلی)* برای ایجاد نسل بعد انتخاب می‌شوند.این فرآیند شبیه "بقای اصلح" در طبیعت است.
4. **ترکیب *(Crossover)* یا *Recombination***:بخش‌هایی از دو فرد *(والدین)* برای ایجاد افراد جدید *(فرزندان)* ترکیب می‌شوند.مثال:اگر والد 1 = "1010" و والد 2 = "0101" باشد، فرزند ممکن است "1011" باشد
5. **جهش *(Mutation)***:تغییرات تصادفی کوچک روی برخی افراد اعمال می‌شود تا تنوع ایجاد شود و از گیر افتادن در راه‌حل‌های نامناسب جلوگیری شود.مثال: تغییر یک "ژن" در "1010" به "1110".فرآیند جهش *(Mutation)* برای جلوگیری از بیش‌برازش *(Overfitting)* استفاده می‌شود.
6. **تکرار *(Repeat)***:این فرآیند چندین بار تکرار می‌شود تا بهترین راه‌حل یا یک راه‌حل قابل قبول پیدا شود.

*** به طور کلی تسلط بر هنر یادگیری ماشین نیاز به زمان و تجربه دارد. چندین نکته مهم وجود دارد که می‌تواند به استفاده بهینه از زمان و کارایی مدل کمک کند :

1. **مدیریت داده‌ها**:مدیریت صحیح داده با فرآیندها، سیاست‌ها، و رویه‌های مناسب ضروری است.
2. **ایجاد خط مبنای عملکرد**:اجرای چندین الگوریتم و مقایسه عملکرد آن‌ها لازم است، زیرا هیچ الگوریتمی برای همه مسائل بهینه نیست.
3. **پاکسازی داده‌ها**:داده‌های تمیز و استاندارد پیش‌نیاز مدل‌های مؤثر هستند.
4. **زمان آموزش**:ابعاد بزرگ داده و محدودیت قدرت محاسباتی می‌تواند زمان آموزش را افزایش دهد. برای مثال، شبکه‌های عصبی ممکن است برای وظایفی که زمان محدود دارند مناسب نباشند.
5. **انتخاب مدل مناسب**: ارزیابی مدل‌های مختلف رویکرد مفیدی است. برای انتخاب مدل، اغلب از اصل تیغ اوکام *(Occam's Razor)* استفاده می‌شود، که ساده‌ترین مدل را پیشنهاد می‌دهد.
6. **انتخاب متغیرها**: اگرچه داده‌های بیشتر معمولاً مطلوب هستند، استفاده از تعداد کمتری از متغیرهای پیش‌بینی‌کننده برای بسیاری از مسائل یادگیری ماشین ارجح است.
7. **حذف متغیرهای زائد**: متغیرهای غیرضروری دقت مدل را کاهش می‌دهند.
8. **جلوگیری از بیش‌برازش**: مدل‌هایی با متغیرهای زیاد ممکن است روی داده‌های واقعی عملکرد ضعیفی داشته باشند.
9. **بهره‌وری** :حجم داده، منابع محاسباتی، هزینه‌ها، و زمان مورد نیاز برای یادگیری و اعتبارسنجی باید در نظر گرفته شوند.
10. **قابلیت درک**:مدل‌هایی با متغیرهای پیش‌بینی‌کننده کمتر قابل فهم‌تر و توضیح‌پذیرتر هستند.
11. **دقت و تعمیم‌پذیری**:تعمیم‌پذیری خوب هدف اصلی مدل‌های یادگیری ماشین است؛ گاهی تقریب دقیق‌تر از دقت بالا مفیدتر است.
12. **تأثیر نتایج منفی کاذب *(False Negatives)***: هنگام ارزیابی مدل پیش از استقرار آن در محیط واقعی، باید تأثیر نتایج منفی کاذب در نظر گرفته شود.**مثال سرطان پستان**: نتیجه مثبت کاذب کمتر خطرناک است، اما نتیجه منفی کاذب می‌تواند مرگبار و پرهزینه باشد.
13. **رابطه خطی و تبدیل داده‌ها**:داده‌های غیرخطی ممکن است نیاز به تبدیل داشته باشند تا مدل‌های خطی بتوانند روی آن‌ها کار کنند *(مانند لگاریتم‌گیری از داده‌هایی که رابطه نمایی دارند)*.
14. **پارامترها و ابرپارامترها**:تنظیم دقیق آن‌ها برای بهینه‌سازی مدل ضروری است.
15. **روش‌های مجموعه‌ای *(Ensembles)***:ترکیب طبقه‌بندی ها با تکنیک‌هایی مانند رأی‌گیری و وزن‌دهی می‌تواند دقت مدل را بهبود دهد.