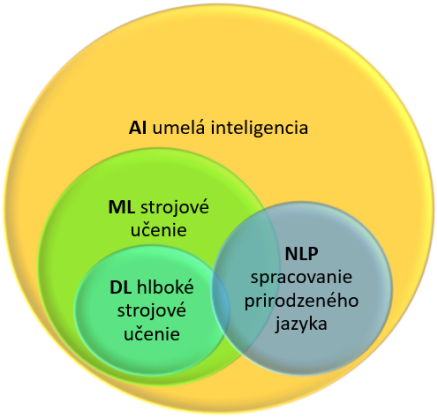
Techniky spracovania prirodzeného jazyka (Natual Language Processing, NLP), syntaktická analýza, token, rozdelenie vety na kúsky, stemovanie, lematizácia, parsovanie, sémantická analýza :)

Oblasť umelej inteligencie, ktorá sa zaoberá analýzou, interpretáciou a generovaním ľudskej reči počítačmi. Je kľúčová pre aplikácie ako chatboty, strojový preklad a analýzu sentimentu.



Cieľom NLP je s jeho použitím spracovať neštruktúrované dáta tak, aby s nimi vedeli pracovať relačné databázy, kde budú k dispozícii pre ich ďalšie spracovanie.

Pozrime sa bližšie na to, ako NLP, spracovanie prirodzeného jazyka, funguje: Bežná hovorená komunikácia či správy na sociálnych sieťach, predstavujú pre program neštruktúrované dáta. Na zachytenie významu týchto slov vieme veľmi efektívne použiť strojové učenie. Cieľom NLP je s jeho použitím spracovať neštruktúrované dáta tak, aby s nimi vedeli pracovať relačné databázy, kde budú k dispozícii pre ich ďalšie spracovanie.

Syntaktická analýza

Keď program potrebuje vyhodnotiť vstup napr. písaného textu, potrebuje si ho upraviť. Sami dobre vieme, že určité veci vieme opísať viacerými spôsobmi. Používame homonymá[1] a synonymá[2]. Čo urobí systém s takou vetou?

Rozdelenie vety na kúsky (tokenization)

S vetou na tri riadky si systém neporadí. Rovnako pre neho môže byť zložitá aj veta so štyrmi slovami. Pomocou tejto techniky si celú vetu rozdelí na samostatné slová, **tokeny**, s ktorými bude ďalej pracovať. Napríklad veta *Hladný kuchár varí polievku* bude vyzerať nasledovne.

Stemovanie (stemming)

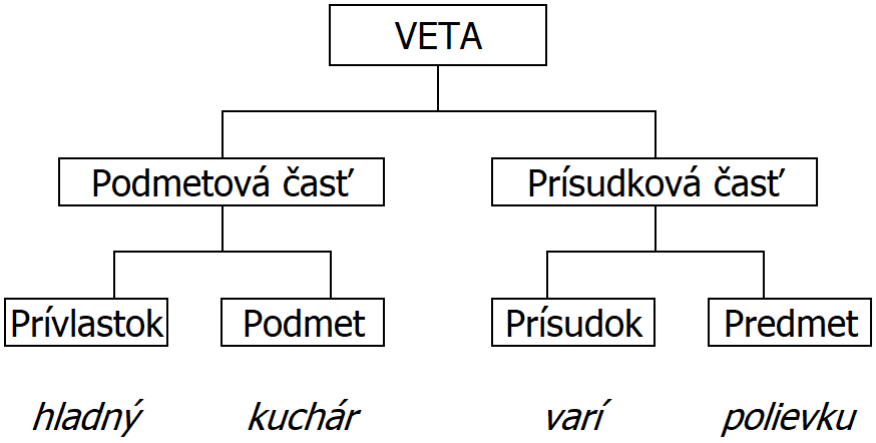
V tomto kroku sa odstránia predpony a prípony, rovnako sa slovo aj normalizuje. V praxi to znamená, že slovo odlet sa upraví na tvar let.

Lematizácia (lemmatization)

Tokeny sa upravujú na základný tvar, ako ho nájdeme napríklad v slovníku. To znamená, že napríklad vyčasované slovesá či vyskloňované podstatné mená, si program nastaví na základný tvar, napríklad šiel => *ísť*, jablkami => *jablko*.

Parsovanie (parsing)

Okrem úpravy slov na základný tvar sa slová rozoberú z hľadiska vetnej štruktúry, to znamená, rozdelenie podmetovej a prísudkovej časti a následne klasifikácia týchto častí. Systém robí v tomto bode tzv. parsovací strom.



Sémantická analýza

Moduly pracujú aj samostatne, avšak ich prepojenie môže ušetriť čas. Sémantika dokáže rozlúštiť význam vety niekedy skôr ako syntax, ktorý by rozoberal slovo za slovom[3].

Na to, aby sme predložený text dostali dokonale preložený, alebo sa audio záznam prekonvertoval na 100 % korektný text, si budeme musieť ešte počkať. Ľudský jazyk je totiž veľmi zložitý a obsahuje rôzne výnimky, dialekty a podobne.

[1] Rovnako znejúce slová s rôznym významom.
[2] Rozdielne slová s rovnakým významom.
[3] Napríklad ak by vo vete bolo slovo pero alebo *kohútik*.

Zdroje
Prevzaté a upravené z:
- **https://umelainteligencia.sk/techniky-spracovania-prirodzeneho-jazyka-nlp/,**
- **https://umelainteligencia.sk/ako-je-mozne-ze-nam-siri-rozumie/.**