Vývin extrakcie informácií

Metódy inžinierskej práce 2023/2024

Nikolas Kristín

Fakulta informatiky a informačných technológií
Slovenská technická univerzita v Bratislave

26. november 2023

V súčasnosti sa denne vygeneruje mimoriadne množstvo neštrukturovaných dát na internete. S narastajúcim objemom týchto neštrukturovaných dát vzniká potreba efektívneho spracovania a získania cenných informácií Preto je dnes nevyhnutnosťou, tieto dáta spracovávať pomocou extrakčných algoritmov. S rastúcou digitalizáciou je spracovanie neštrukturovaných dát kľúčovým prvkom pre úspešné využívanie potenciálu internetu a digitálnej éry.

Prehľad

Message Understanding Conference (MUC)

Conference	*Year*	*Text Source*	*Topic (Domain)*
MUC-1	1987	Mil. reports	Fleet Operations
MUC-2	1989	Mil. reports	Fleet Operations
MUC-3	1991	News reports	Terrorist activities in Latin America
MUC-4	1992	News reports	Terrorist activities in Latin America
MUC-5	1993	News reports	Corporate Joint Ventures, Microelectronic production
MUC-6	1995	News reports	Negotiation of Labor Disputes and Corporate Management Succession
MUC-7	1997	News reports	Airplane crashes, and Rocket/Missile Launches

Informácie jednotlivých konferencií, Zdroj: en.wikipedia.org/wiki/MessageUnderstandingConference

Regex

Named Entity Recognition

Relation Extraction

Customized Data Pipelines

Jasper

Rosoka