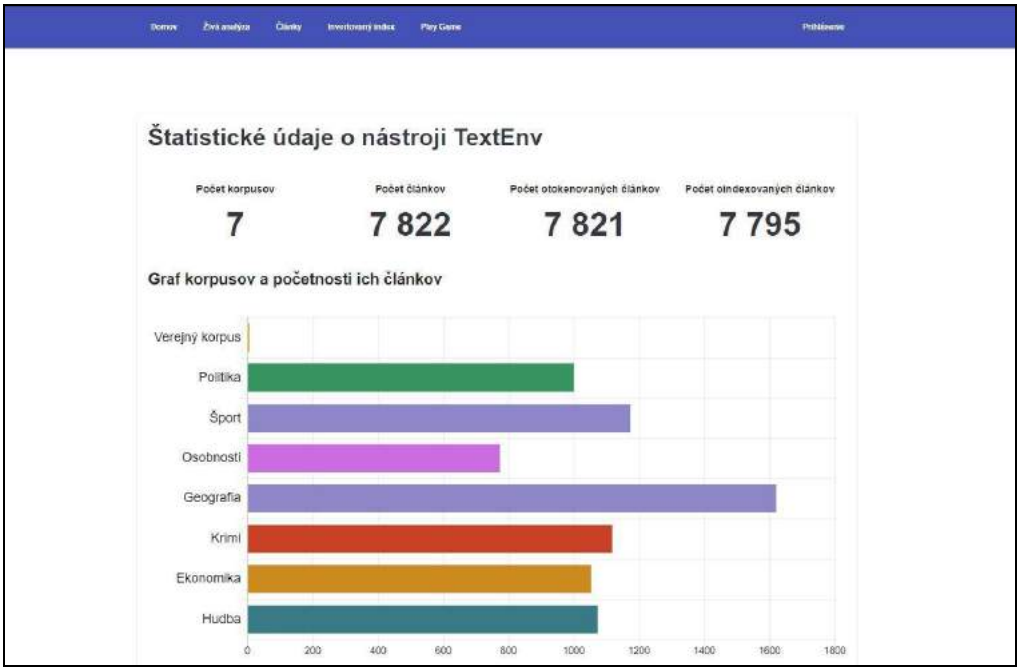
TextMania - Prostredie pre inteligentnú analýzu slovenského textu

O ČOM JE NÁŠ PROJEKT?

V súčasnosti nás obklopuje veľké množstvo textu v elektronickej podobe a inteligentná analýza textu je v dnešnej dobe veľmi zaujímavou témou v oblasti informačných technológií. Po schválení viacerých zákonov Európskou úniou týkajúcich sa textov a prejavov na internete, budú nástroje zaoberajúce sa automatizovaným spracovaním textu ešte viac žiadané.

Projekt TextMania vznikol ako odpoveď na problém nedostatku času manuálne analyzovať textové dokumenty, resp. pre potrebu analýzy textu v reálnom čase (segmentácia textu, detekcia tém, určovanie vhodnosti textu pre určité skupiny ľudí a podobne). Nespornou výhodou je podpora slovenského textu, čím v podstate vypĺňa „dieru na trhu“. Výsledky textovej analýzy využijú nielen používatelia pri triedení a filtrovaní svojich článkov, ale aj dátoví vedci pre vylepšovanie algoritmov.

Dnes už síce existujú nástroje zaoberajúce sa spracovaním a vizualizáciou textových dát, avšak žiaden z nich nie je vhodný pre texty v slovenskom jazyku a jeho špecifické črty (napr. rôzne tvary slov). Z tohto dôvodu sme sa v rámci nášho projektu rozhodli vytvoriť webové prostredie pre inteligentnú analýzu textu, ktoré by zjednodušilo vykonávať textovej analýzu nielen dátovým vedcom, ale aj bežným používateľom.



PONÚKANÉ RIEŠENIE?

Cieľom projektu TextMania je vytvoriť prostredie pre inteligentnú analýzu textu napísaného v slovenskom jazyku. Finálny produkt ponúkne možnosť importovať, analyzovať a automaticky spracovať články na základe ich obsahu pre rôzne úlohy klasifikácie textov či extrakcie čŕt. Systém inteligentnej analýzy textu sme sa rozhodli navrhnúť ako webovú aplikáciu.

Na začiatku získame vybrané články (podľa tématiky) zo stránok: wikipédia.sk a webnoviny.sk. Následne ich označíme podľa kategórie, z ktorej boli stiahnuté a uložíme ich do databázy. Ďalej sa vykoná lexikálna a syntaktická analýzu vložených textov, aby bolo možné následne aplikovať požadované metódy strojového učenia pre identifikáciu entít v texte či klasifikáciu textu z rôznych hľadísk (napr. určenie témy alebo vhodnosti textu pre určitú vekovú skupinu ľudí). Texty sa analyzujú pomocou pripravených metód spracovania prirodzeného jazyka, ktoré bude možné rozšíriť a vzájomne porovnávať. Ďalej sa vytvorí napr. invertovaný index pre urýchlenie a zjednodušenie vyhľadávania v článkoch a korpusoch, čo tiež umožní identifikovať kroky použitých algoritmov v prípade potreby ich vylepšenia. Pre určenie „relevantnosti“ je zatiaľ použitá extrakcia čŕt pomocou štatistickej metódy tf-idf (*term frequency–inverse document frequency*), avšak počíta sa s rozšírením o ďalšie algoritmy.

Naše webové riešenie poskytuje aj možnosť skúmať text formou hry, ktorá zobrazí používateľovi náhodne vybranú vetu z niektorého z článkov a používateľ má za úlohu vybrať prislúchajúci korpus a svoj výber podložiť stávkou z bodov (peňazí) pridelených na začiatku hry. Touto formou vieme získať dodatočné dáta, ktoré môžu poslúžiť na vylepšenie použitých algoritmov. Môže byť napríklad zaujímavé porovnávať dáta vypočítané „strojom“ s dátami od reálnych používateľov.

POUŽITÉ TECHNOLÓGIE:

- Node.js
- Python Django
- Angular Framework
- Express Framework
- MongoDB

Kto sme?

Náš tím pozostáva z 7 študentov prvého ročníka inžinierskeho štúdia na Slovenskej technickej univerzite - Fakulte informatiky a informačných technológií v Bratislave v odbore Inteligentné softvérové systémy. Projekt Textmania je výsledkom spolupráce nasledujúcich členov tímu: Dávid Csomor, Adam Ďuriš, Alan Kováč, Daniel Kováč, Peter Križan, Patrik Melicherík, Krištof Orlovský. Vedúcim projektu je Ing. Miroslav Blšták.