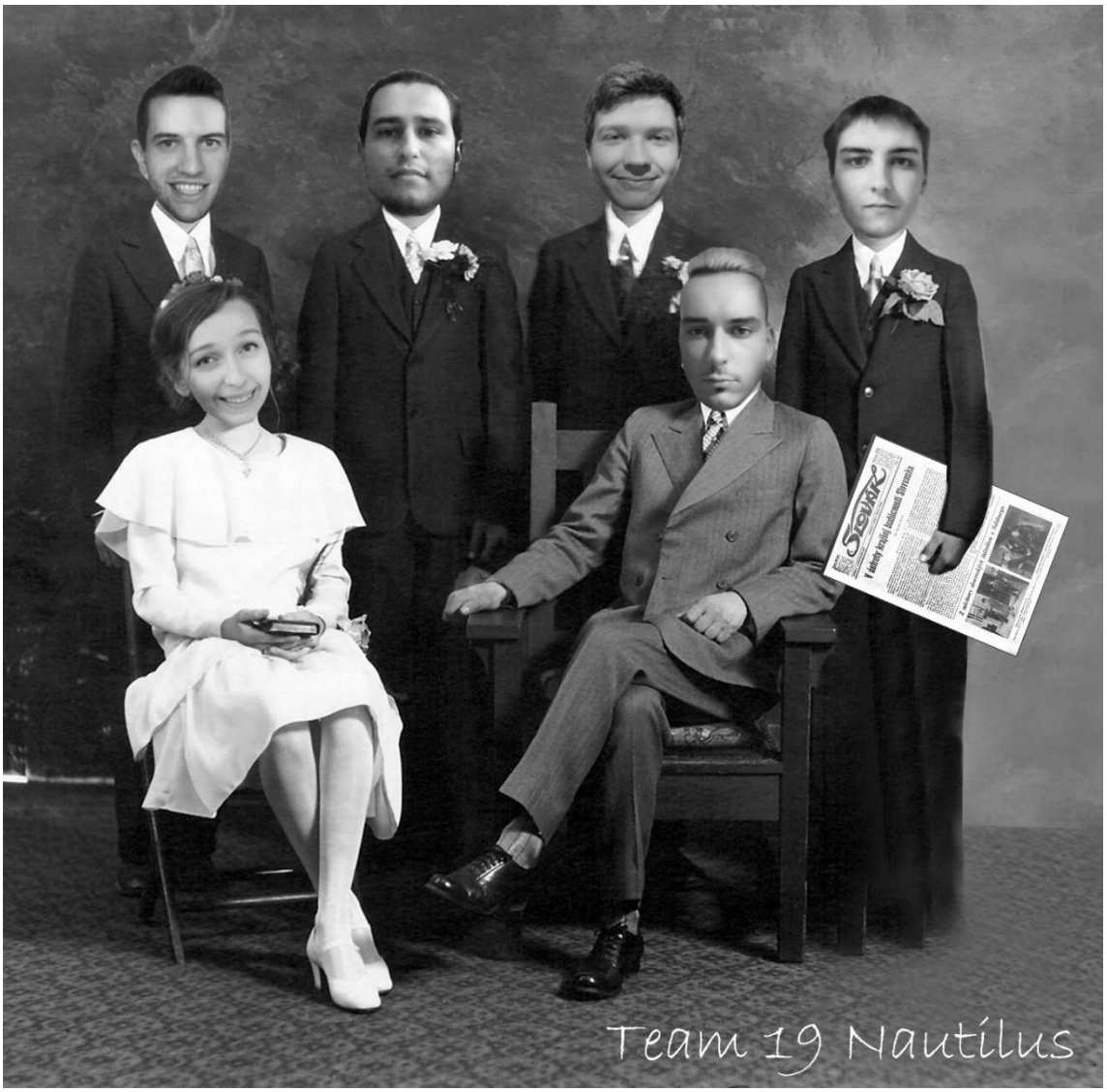
TÍM č. 19 IS-SI

Nautilus



Názov projektu:

Sémantické vyhľadávanie v starej tlači

Členovia tímu (študenti):
Bc. Jakub Hagara, Bc. Adam Rafajdus, Bc. Martina Redajová, Bc. Tomáš Repiský, Bc. Martin Vaško

Ved. tímu (pedagóg):
Dr. Nadežda Andrejčíková

Motto tímu:
Pomáhame sprístupniť informácie

O ČOM JE NÁŠ PROJEKT?

Sémantické vyhľadávanie, či inak povedané vyhľadávanie na základe významu použitého termínu, sa v súvislosti s nárastom počtu informácií, ale často aj potrebou nájsť rýchlo tie správne informácie, stáva nevyhnutnosťou. Úspešnosť vyhľadávania však závisí od dostupnosti zdrojov. Periodické dokumenty predstavujú bohatý a nenahraditeľný zdroj informácií. V súčasnosti, keď ich príprava, či samotné publikovanie je realizované v elektronickej podobe to nie je problém, čo sa však nedá povedať o tých starších. Digitalizácia nám umožňuje ich prevod do elektronickej podoby, ale ani to nie je dostatočné. Pre používateľa je dôležité, aby ho vyhľadávanie naviedlo na konkrétny článok a nie na celé číslo časopisu. Preto sme sa rozhodli v našom projekte zautomatizovať proces spracovania takto zdigitalizovaných periodík až na úroveň analytického rozpisu článkov.

Súčasťou procesu digitalizácie je rozpoznávanie znakov, k čomu sa využíva Adobe Recognition Software a digitalizované obrazy transformuje do špeciálnych XML dokumentov. Takéto súbory, ale aj bežné txt dokumenty, či word dokumenty sú tie, na ktoré sa zameriavame a ktoré sú na vstupe do nášho procesu. Tieto špeciálne XML súbory síce obsahujú množstvo informácií, ale sú to zväčša informácie týkajúce sa formátovania a úpravy textu, čo pre účely vyhľadávania nemá dostatočný význam. Našim hlavným cieľom v tomto projekte je, ako sme už uviedli, identifikovať tie informácie, ktoré nám umožnia správne rozpoznať názvy a k nim prislúchajúce texty konkrétnych článkov. Následne tieto texty spracovať a identifikovať v nich kľúčové slová, neskôr aj význam, v akom boli tieto slová v danom texte použité. Na základe takto získaných údajov generovať bibliografické záznamy pre jednotlivé čísla periodík, ako aj pre konkrétne články, vo formáte MARC21 podľa platných katalogizačných pravidiel, aby s nimi mohli priamo pracovať aj knižnično-informačné systémy. K jednotlivým bibliografickým

záznamom tiež v našom digitálnom repozitári ukladáme aj plné texty a výrez zo zdigitalizovaného obrazu strany, ktorý zachytáva daný článok.

Keďže kvalita zdrojov a schopnosti OCR nástrojov rozpoznávať znaky nemusí byť stopercentná, je súčasťou nášho projektu tiež návrh a realizácia aplikácie, ktorá umožňuje zamestnancom inštitúcií pre archiváciu takýchto zdrojov editovať výsledky OCR spracovania, teda napríklad upravovať nepresnosti pri rozpoznávaní textu. Rovnako pomocou tejto aplikácie umožňujeme týmto správcom upravovať výsledky nášho algoritmu pre rozdelenie dokumentu na samostatné dokumenty - články, pretože vplyvom nekonzistencie tlače nie je možné garantovať 100%-nú úspešnosť rozpoznania článkov z týchto údajov.

ČO NÁM DÁVA PRÁCA NA TOMTO PROJEKTE?

Keďže sme pri práci na tomto projekte riadili agilnou metódou vývoja Scrum, získali sme cenné skúsenosti ako v budúcnosti postupovať pri práci na projektoch riadených touto metódou vývoja. Na vlastnej koži sme si odskúšali, aké prínosné je napr. rozdelenie rozsiahlej úlohy na menšie podúlohy, keďže pre nich vedeli lepšie odhadnúť ich zložitosť, vďaka čomu sa nám podarilo znížiť niektoré riziká a zefektívniť deľbu práce v tíme.

Prácou na tomto projekte sme tiež kolektívne prišli k zisteniu, že základným kameňom práce v tímu je častá tímová komunikácia, a zlyhanie komunikácie, resp. nedostatočná komunikácia vychádza pre tím veľmi draho, čo sa týka času. Taktiež nám práca ukázala, aké účinné je mať ozdobný kameň tímovej práce, ktorým sú metodiky. Zadefinovanie rôznych metodík, hoci sa spočiatku javí ako nadbytočná úloha, sa vyfarbí ako pomôcka, vďaka ktorej vie tím pracovať efektívne a rýchlo napredovať bez zbytočného zmätku či chaosu, ktorý často nastáva pri neexistencii, či nedodržiavaní zadefinovaných metodík.

Vďaka netriviálnosti problematiky nám projekt umožňuje hlboko rozšíriť naše znalosti pri práci s rôznymi jazykmi a knižnicami, pomocou ktorých je napísaná aplikácia nášho projektu. Na spoluprácu pri vytváraní kódu sme použili verziovací nástroj Git, ktorý je v dnešnej

dobe základom pre každého informatika. Tiež sme si vyskúšali spoločne pracovať na väčšom projekte a riešiť problémy, ktoré s takouto úlohou prichádzajú. Jadro nášho algoritmu je práca s XML súbormi, preto zlepšenie sa pri práci s takýmito dátami bude pre nás využiteľné aj v budúcnosti. Okrem iného nám projekt taktiež ponúka veľké množstvo historických dokumentov spracovaných pomocou OCR, takže máme skúsenosti so spracovanými dokumentami v tejto forme. Rovnako sme získali skúsenosti s vytváraním MARC21 záznamov.

PREČO JE NÁŠ PROJEKT ZAUJÍMAVÝ?

V periodických dokumentoch, ktoré uchovávajú v svojich fondoch pamäťové a fondové inštitúcie je ukrytých mnoho zaujímavých informácií. Takéto informácie môžu priniesť nový pohľad na udalosti, ktoré sa v minulosti odohrali, ale aj objasniť prečo sa tieto udalosti stali. Avšak v dnešnej dobe neexistuje spoľahlivý spôsob, ako takéto dáta získať v elektronickej podobe až po úroveň analytického rozpisu jednotlivých čísiel na konkrétne články. Existujú spôsoby, ako získať texty celého vydania časopisu, čo však do značnej miery zahlcuje výsledky vyhľadávania. Naše riešenie prináša metódu, ako automatizovať proces extrakcie konkrétnych článkov ako aj analytického rozpisu jednotlivých čísiel periodík s vytvorením relevantných bibliografických záznamov s prepojením na plné texty článkov a ich vizualizované obrazy. Vďaka výsledkom nášho projektu bude možné značne spresniť výsledky vyhľadávania ako aj identifikovať a odhaľovať nové poznatky a vzťahy medzi základnými entitami v týchto dátach. Používatelia tak získajú nový plnohodnotný zdroj širokého spektra informácií.

POUŽITÉ TECHNOLÓGIE:

Python 3.x (lxml), ElasticSearch, RubyOnRails, Javascript

O ČOM TO VLASTNE JE?

Cieľom nášho projektu je vytvoriť nástroj pre spracovanie zdrojov získaných ako výsledok pri spracovaní tlače pomocou nástrojov pre spracovanie znakov. Tieto zdroje obohacujeme o relevantné

informácie potrebné pri sémantickom vyhľadávaní. Výsledky tohto spracovania zobrazujeme v aplikácii, ktorá ich umožňuje nielen prezerať, ale aj editovať a extrahovať v podobe MARC záznamu alebo obrázkov.