OD SUROVÝCH DÁT K PRVÝM NÁČRTOM TRENDOV VO VZDELÁVANÍ KNIŽNIČNO-INFORMAČNÝCH VIED

Marcela Katuščáková – Martin Záborský

Hlavný cieľ projektu ADAMIS realizovaného v rokoch 2013-2015 je uvedený v jeho názve – *Adaptácia študijného programu mediamatika a kultúrne dedičstvo na potreby vedomostnej spoločnosti*. Preto je prirodzené, že jednou zo štyroch hlavných aktivít projektu bol audit súčasného stavu a návrh zmien (aktivita 1.1) s cieľom „*konfrontovať existujúci študijný program s najlepšími programami v danom odbore i s aktuálnymi požiadavkami praxe*" (Žilinská univerzita, 2015). V rámci aktivity prebehlo porovnanie štruktúry a obsahu existujúceho študijného programu mediamatika a kultúrne dedičstvo realizovaného na Katedre mediamatiky a kultúrneho dedičstva s podobnými programami na iných univerzitách vo svete.

Výber podobných študijných programov na zahraničných univerzitách sa realizoval podľa rebríčkov univerzít QS World University Rankings (2013) a Academic Ranking of World Universities (2013) z údajov dostupných za rok 2013. Na základe definovania oblasti záujmu v spomínaných rebríčkoch a v nich uvedených univerzitách bolo identifikovaných 141 príbuzných študijných programov na 92 fakultách 76 univerzít, ktorých názvy predmetov sa porovnávali s obsahovou náplňou študijného odboru knižnično-informačné štúdiá v študijnom programe mediamatika a kultúrne dedičstvo.

Zber údajov z vybraných študijných programov

V prvej fáze sa realizoval potrebný zber údajov na webových stránkach vyššie uvedených univerzít v štruktúre: študijný program; fakulta; univerzita; stupeň; názvy predmetov z informačných listov; linky na tieto informačné listy a na kontaktnú osobu študijného programu, ktoré členovia projektu získali z verejne dostupných zdrojov (prevažne webstránky), prípadne komunikáciou so zodpovednými osobami priamo na cieľovej univerzite. Následne bolo nevyhnutné získané údaje štruktúrovať do zmysluplných celkov, na účely čoho bol vytvorený online Google formulár.

Cieľom formulára bolo získať prehľad o štruktúre sledovaných študijných programov a dať do vzťahu predmety zo sledovaných univerzít s jednotlivými predmetmi študijného programu mediamatika a kultúrne dedičstvo. Pokiaľ túto väzbu nebolo možné jednoznačne identifikovať, mal respondent možnosť predmet zaradiť do jednej alebo viacerých všeobecne definovaných kategórií predmetov.

Predmety študijného programu mediamatika a kultúrne dedičstvo boli v dotazníku pre lepšiu orientáciu a uľahčenie spracovania rozdelené do nasledujúcich kategórií:

- žurnalistika a médiá,
- informačno-komunikačné technológie,
- manažment a ekonomika,

- sociálne, psychologické, filozofické a etické aspekty,
- knižničné (aktuálne a tradičné),
- umelecké,
- matematika, logika a štatistika,
- ostatné predmety vyučované v rámci študijného programu mediamatika a kultúrne dedičstvo,
- iné, ktoré sa nevyučujú v rámci študijného programu mediamatika a kultúrne dedičstvo.

Okrem uvedeného základného členenia podľa obsahu, boli tiež špecificky vybrané a odlišne označené predmety nájdené na stránkach študijných odborov z knižničnej a informačnej vedy (Library and Information Science, ďalej LIS), aby sa oddelili od predmetov nájdených na príbuzných študijných odboroch so študijnými programami zameranými na *new media studies* a *information studies*.

Predspracovanie údajov z dotazníka

Napriek tomu, že dotazník mal štruktúrovanú podobu, jednotlivé polia určené na vkladanie obsahov boli prevažne textového typu. Tento aspekt na jednej strane umožnil, aby respondenti neboli zbytočne limitovaní pevne danou množinou odpovedí a mohli zachytiť všetky nájdené predmety, na druhej strane ale neštruktúrovaná podoba spôsobovala problémy pri následnom spracovaní a vyhodnocovaní údajov.

Výstupom z formulára boli údaje vo forme tabuľky roztriedené podľa jednotlivých študijných programov ale hlavne zatriedené do jednej z deviatich vyššie uvedených kategórií predmetov. Celkovo tabuľka obsahovala približne 3200 zaplnených buniek, ktoré bolo potrebné dodatočne čistiť a upraviť na jednotný tvar vhodný na následné automatické spracovanie textov, čo si vyžiadalo rozsiahle zásahy.

Vyššie bolo spomenuté, že vstupné textové polia formulára umožnili respondentom vložiť rôzne údaje v neštruktúrovanej podobe, čo mnohí realizátori projektu pôsobiaci mimo našej katedry využili nevhodným spôsobom. K názvom predmetov priložili aj ich opisy, prípadne ďalšie doplňujúce informácie zo stránok univerzít, ktoré sa do tohto formulára zameraného na kategorizáciu nájdených predmetov už vkladať nemali. Navyše tieto údaje sme našli vložené v rôznych jazykových podobách. Údaje získané z dotazníka vo forme tabuľky odpovedí mali nekonzistentný charakter a pred vlastnou frekvenčnou analýzou bolo nutné ich nahrubo očistiť, aby neprimerane neovplyvňovali výsledky analýzy. Toto čistenie prebiehalo "ručne" prechádzaním celého zoznamu (3200 názvov predmetov, ich prekladaním do anglického jazyka, mazaním, resp. úpravou irelevantných častí, nakoľko počítačové spracovanie by si vyžadovalo pokročilé nástroje spracovania prirodzeného jazyka a sémantickej analýzy, ktoré by pri danom rozsahu textu nemuseli priniesť želané výsledky, keďže pre efektívnu činnosť sa vyžaduje veľký rozsah textu a vhodne určené trénovacie údaje.

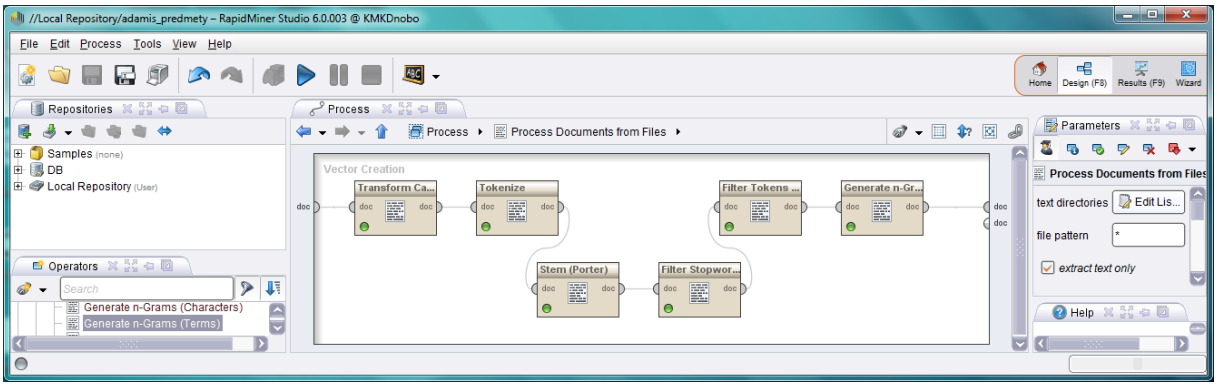
Za účelom strojového spracovania textu v analytickom nástroji a z dôvodu lepšej prehľadnosti a jednoduchosti manipulácie s údajmi bolo nevyhnutné umiestniť každý vložený predmet do samostatnej bunky tabuľky, preto bolo nevyhnutné predmety z textových polí formulára rozdeliť. Na tento účel sme do tabuľkového kalkulátora Microsoft Excel implementovali krátky VBA skript, ktorý ako oddeľovač jednotlivých predmetov vnímal interpunkčné znamienka ako bodky, čiarky, bodkočiarky alebo pevné ukončenie riadka. Každý nanovo oddelený predmet potom pridal na samostatný riadok v rámci daného stĺpca, čím ostali zachované pôvodné kategórie predmetov. Tento postup nebol stopercentný najmä z dôvodu rôznorodosti získaných vstupných údajov, ale vzhľadom na ich rozsah priniesol uspokojivé výsledky.

Frekvenčná analýza názvov nájdených predmetov

Pôvodne bol pre potreby projektu zamýšľaný nákup komerčného softvéru na dolovanie znalostí z textov, ale z dôvodu problémov pri verejnom obstarávaní a nedodania požadovaného softvéru sa celá analýza musela vykonať vo voľne dostupnom nástroji na strojové spracovanie prirodzeného jazyka. Na základe článku a výsledkov analýzy v elektronickom magazíne Predictive Analytics Today (2014), v ktorom odporúčajú najlepšie voľne dostupné aplikácie na analýzu textov, sme vybrali a otestovali niekoľko nástrojov na frekvenčnú a zhlukovú analýzu textov.

Po preskúmaní jednotlivých nástrojov a zohľadnení ich funkčných, prípadne praktických obmedzení, sme na frekvenčnú analýzu vybrali voľne dostupnú verziu aplikačného prostredia RapidMiner Studio (2015), ktorá primárne slúži na riešenie rôznych problémov a úloh dolovania znalostí zo štruktúrovaných údajov (štatistická analýza, predikčná analýza, strojové učenie a pod.), ale umožňuje tiež dodatočnú inštaláciu, t.j. voliteľné rozšírenie, ktorým sa dá funkcionalita rozšíriť aj o potrebné funkcie pre spracovanie neštruktúrovaných údajov s dôrazom na analýzu textov. Tento softvér umožňuje nie len tvorbu frekvenčných analýz pre jednoslovné, ale aj pre viacslovné tokeny (n-gramy), čo bolo pre prácu s názvami predmetov z jednotlivých univerzít nevyhnutnou podmienkou, s ohľadom na často sa opakujúce tokeny (napr. *information* alebo *library*) a nutnosť upresniť ich význam v danom kontexte.

Riešenie analytického problému sa v aplikácii RapidMiner definuje pomocou procesného modelu. V tomto modeli sa postupne pripájajú funkčné operátory, následne sa vhodne určia ich vzájomné väzby a nastavia sa ostatné parametre. Celý proces je potom interaktívne znázornený v prehľadnej vizuálnej podobe. Pre frekvenčnú analýzu sme použili model pozostávajúci z funkčných operátorov určených na: načítanie zdrojového súboru s predspracovaným textom (process document from Files), transformáciu všetkých písmen na malé (Transform Cases), tokenizáciu na základe medzery (Tokenize/non-letters), Porterov stemmer na generovanie koreňov slov (Stem Porter), filtrovanie neplnovýznamových stop-slov (Filter Stopwords/ English), filtrovanie slov kratších ako 3 znaky (Filter Tokens) a z operátora pre určenie frekvencie jednotlivých termov (obr. 1).



Obr. 1: Model procesu v aplikácii RapidMiner

Výstupom procesu boli vygenerované frekvenčné zoznamy jednoslovných termov zo všetkých študijných programov a zvlášť z predmetov identifikovaných ako predmety z programov LIS. V ďalšej fáze sme proces rozšírili o operátor n-gramov nastavený na dvojice (Generate n-Grams/Terms), ktorý uvažoval s každou susediacou dvojicou tokenov ako jedným elementom a po ich spracovaní sme opäť dali vygenerovať frekvenčné zoznamy predmetov zo všetkých príbuzných odborov a len predmetov z odborov LIS.

Tab. 1: Ukážka prvých dvanástich pozícií najčastejšie sa vyskytujúcich slovných 2-gramov z predmetov zo všetkých aj príbuzných odborov

Poradie	2-gramy	Výskyt
1	information_system	49
2	information_science	35
3	Information_management	33
4	information_retrieval	29
5	knowledge_management	24
5	media_communication	24
5	social_media	24
6	digital_library	21
7	information_study	19
8	digital_media	18
8	information_technology	18
9	information_resource	17
10	media_communication	16
10	media_culture	16
10	media_study	16
11	human_computer	14
11	information_communication	14
11	information_service	14
12	cultural_heritage	13
12	information_society	13

Tab. 2: Ukážka prvých dvanástich pozícií najčastejšie sa vyskytujúcich slovných 2-gramov z predmetov len z LIS študijných odborov

Poradie	2-gramy	Výskyt
1	digital_library	43
2	information_science	33
3	information_system	31
4	information_retrieval	21
5	knowledge_management	19
6	information_resource	14
6	information_service	14
7	cultural_heritage	13
7	information_study	13
8	library_information_science	11
9	information_management	10
9	information_society	10
10	information_literacy	9
11	human_computer	8
11	information_architecture	8
12	business_intelligence	7
12	database_design	7
12	information_communication	7
12	information_ethics	7
12	knowledge_organization	7
12	library_service	7

Pri určovaní frekvencií dvojíc tokenov môže vzniknúť problém s jednoslovnými názvami predmetov, ktoré sa v tomto prípade kvôli vstupnej podmienke nezahrnú do analýzy, čím sa stratí časť dostupnej informácie. Uvedený problém sa dá riešiť zakomponovaním umelého tokenu s jedinečným tvarom, ktorý sa pridá na začiatok a koniec každého názvu predmetu. Výsledky tak nemusia byť ovplyvnené faktom, že sa berú do úvahy iba viacslovné názvy predmetov. Keďže však získané jednoslovné tokeny boli príliš všeobecné a nemali väčšiu výpovednú hodnotu, ďalej sme s nimi neuvažovali.

Za relevantnejšie považujeme frekvenčné zoznamy získané na základe dvojíc susediacich tokenov, lebo hlbšie charakterizujú dané predmety. Pokúšali sme sa realizovať aj frekvenčnú analýzu pre susediace trojice tokenov, ale v tomto prípade už boli výsledky skreslené veľkým množstvom kombinácii a nízkymi výskytmi aj najfrekventovanejších trojíc tokenov. Navyše veľká časť názvov predmetov tri tokeny ani neobsahovala.

Vizualizácia výsledkov

Údaje z frekvenčnej analýzy slov, resp. z párov po sebe nasledujúcich slov, sme použili pre vizualizáciu získaných výsledkov formu mraku tagov. Na jeho tvorbu sme použili voľne dostupný online nástroj wordle.net (Feinberg, 2014), ktorý ponúka pomerne široké možnosti nastavenia parametrov a požadovaných vlastností pri jeho tvorbe (rôzne veľkosti, fonty písma, farebné propozície, preferencie tvaru a iné).

Keďže nástroj wordle.net vyžaduje ako vstup čistý text bez údajov o frekvencii výskytu jednotlivých tokenov, výsledky získané pri frekvenčnej analýze sme upravili na tvar vhodný pre vizualizáciu v aplikácii Microsoft Excel jednoduchým nakopírovaním tokenov podľa počtu ich výskytu, aby mrak tagov adekvátne premietol veľkosť každého tokenu podľa frekvencie jeho výskytu. Pri pokuse vizualizovať dvojslovné termy nastal problém, keďže nástroj wordle.net dvojslovné termy považoval za dva rôzne tokeny a nie za jeden element. Keďže sme iný vhodný nástroj nenašli, dvojslovné termy sme spojením upravili na jednoslovný tvar, pričom pôvodné tokeny sme rozlišovali veľkosťou písma (prvé písmeno prvého tokenu veľké, ostatné písmená malé; rovnako pri druhom tokene).

Vygenerovali sme viaceré mraky tagov s rôznymi vizuálnymi parametrami, s rôznym počtom zahrnutých najfrekventovanejších výrazov z názvov všetkých predmetov, resp. výrazov len z názvov predmetov vyučovaných na programoch LIS – reprezentačný výber predstavujú obrázky 2 a 3.

Obr. 2: Mrak tagov slovných 2-gramov pre všetky predmety

Obr. 3: Mrak tagov slovných 2-gramov pre predmety jadra

Záver

Výsledkom našej práce bol prvý náčrt trendov vo vzdelávaní na príbuzných študijných programoch v zahraničí. Získané a spracované výsledky si vyžadovali následnú kvalitatívnu analýzu, ktorú realizovali garanti študijného programu mediamatika a kultúrne dedičstvo. Postupne pri realizácii jednotlivých krokov aktivity 1.1 projektu ADAMIS sme si uvedomili, že nie je v našich silách dopracovať sa k vyčerpávajúcim výsledkom, s ktorými by sme mohli byť úplne spokojní, hlavne z časových dôvodov. Zistili sme, že projekty mapujúce pokrytie len jednej oblasti (napr. manažment znalostí na LIS školách) boli predmetom samostatného výskumu odborníkov z niekoľkých univerzít. Preto pôvodný zámer vytvorenie akejsi mapy oblastí, ktoré dnes predstavujú trendy na LIS školách sme pozmenili na kvantitatívnu analýzu predmetov, vyučovaných na vybraných školách, t.j. na zber názvov predmetov a ich následnú frekvenčnú analýzu. Realizovaný postup nie je z niekoľkých hľadísk ideálny: niekto by mohol namietať voči veľkosti a zloženiu výberovej vzorky (študijné predmety), iný, že názov predmetu neodzrkadľuje jeho skutočný obsah, alebo že sme mali tiež zisťovať počet hodín alebo kreditov, teda váhu, ktorú jednotlivým predmetom školy prikladajú. Váhu predmetov nebolo možné objektívne zisťovať, nakoľko niektoré sledované školy mali zverejnené informačné listy aj s počtom hodín pre daný predmet, ale mnohé nie. Keďže mnohé školy svoje informačné listy nemajú uvedené dvojjazyčne, teda aj v anglickom jazyku, často sme sa dopracovali len k informačným listom v pôvodnom jazyku (napr. grécky, švédsky, maďarský a pod.), takže ich obsah sme určili za pomoci slovníkov, čo určite do istej miery tiež skreslilo výsledky výskumu.

Napriek tomu sa domnievame, že naše výsledky možno považovať na danej úrovni za použiteľné a odzrkadľujú v hrubých líniách oblasti, ktoré sa na významných univerzitách s LIS programom dnes vyučujú, nakoľko ich považujú za potrebné pre svojich študentov. Ide v prvom rade o predmety so zameraním na digitálne knižnice, informačnú vedu, informačné systémy, informačný prieskum, manažment znalostí, informačné zdroje, služby a pod.

Zoznam bibliografických odkazov

Feinberg, J., 2014. *Wordle.net*. [online]. [cit. 2.12.2015]. Dostupné na: <http://wordle.net/>.

Žilinská univerzita, 2015. *Projekt ADAMIS: Webová stránka k projektu ADAMIS*. [online]. [cit. 12.12.2015]. Dostupné na: <http://adamis.uniza.sk>.

Predictive Analytics Today, 2014. *Top Free Software for Text Analysis Text Mining Text Analytics*. [online]. [cit. 4.3.2015]. Dostupné na: <http://www.predictiveanalyticstoday.com/top-free-software-for-text-analysis-text-mining-text-analytics>.

Quacquarelli Symonds Limited, 2013. *QS World University Rankings*. [online]. [cit. 16.10.2013]. Dostupné na: <http://www.topuniversities.com/university-rankings>.

RapidMiner, 2015. *RapidMiner Studio*. [online]. [cit. 12.3.2015]. Dostupné na: <https://rapidminer.com>.

ShanghaiRanking Consultancy, 2013. *Academic Ranking of World Universities*. [online]. [cit. 17.10.2013]. Dostupné na: <http://www.shanghairanking.com>.