Modul Acrea Text Mining 3.0

Textminingový nástroj Acrea Text Mining 3.0 (ATM) je modulem dataminingového softwaru IBM SPSS Modeler případně softwaru PS Clementine, který SPSS Modeler v sobě zahrnuje. Modul ATM je sestaven z procedur umožňujících transformovat nestrukturovaná textová data z dokumentů psaných v přirozeném jazyce do strukturované (tabulkové) podoby vhodné pro další strojové zpracování. Modul podporuje strojovou, jazykově závislou extrakci atributů a pojmenovaných entit především z českých a slovenských textů a klasifikaci dokumentů podle jejich sentimentu.

Kromě jazykově závislých procedur nabízí modul i stringologické procedury. Pomocí regulárních výrazů lze z volného textu extrahovat n-gramy, tokeny, emaily, URL, datum, čas a mnoho dalších entit, které můžeme regulárními výrazy specifikovat. Podobnost textových řetězců je možné měřit pomocí editačních metrik.

ATM nabízí i vizualizaci a přehledné zobrazování dokumentů. Textové atributy jako jsou slova nebo pojmenované entity je možné nakreslit do obrázku (Word cloud), kde se jednotlivé pojmy nepřekrývají a jejich velikost či barva indikuje četnost výskytu. Pokud je třeba zobrazit přímo zdrojové dokumenty např. po jejich výběru strojovým modelem, modul ATM nabízí prohlížeč, kde se texty dokumentů zkrátí a volitelně se k nim připojí celá řada metadat jako např. autor, datum nebo klíčová slova.

Integrace do SPSS Modeler

Textminingové procedury jsou integrovány do prostředí IBM SPSS Modeler jako uzly. Při práci se uzly modulu ATM standardním způsobem zařazují do proudů, jež slouží jako vizuální záznam postupu přípravy dat, analýzy, modelování a predikce. Textminingové uzly mají vlastní paletu nástrojů a jsou do dataminingového softwaru plně integrované. Lze s nimi pracovat nejen v proudech ale i pomocí skriptů v jazyce Jython.

Textová data se do textminingových uzlů dostávají v textových proměnných. Je možné je čerpat z běžných datových formátů jako XML, databáze, textový soubor či Excel, které běžně načítá SPSS Modeler. Modul ATM navíc nabízí speciální uzel pro načítání textů z nestrukturovaných textových souborů.

Modul ATM slouží primárně pro extrakci strukturovaných atributů z volných textů. Po načtení textových dokumentů a extrakci atributů se zpravidla k dalšímu zpracování používají standardní uzly SPSS Modeler. Pomocí nich můžeme získané atributy filtrovat, spojovat, restrukturalizovat apod. tak, abychom získali strukturovanou reprezentaci kolekce dokumentů ve tvaru vhodném pro zamýšlenou analýzu nebo modelování. Díky vnoření modulu ATM do SPSS Modeler můžeme analyzovat a prezentovat obsah kolekce, sestavovat predikční textminingové modely nebo třeba studovat časové řady témat a sentimentů. Velmi oblíbené je i zkoumání a vizualizace vazeb mezi nalezenými entitami a termíny.

Lokální a cloudové zpracování textových dat

Jazykově závislé zpracování textu se neprovádí na serveru či klientské stanici, kde je nainstalován IBM SPSS Modeler, nýbrž textová data se zabezpečeně posílají na vzdálený textminingový server, kde jsou umístěny rozsáhlé lingvistické zdroje a výkonné procedury pro zpracování textů v přirozených jazycích. Vzdálenému zpracování textů odpovídá i licencování textminingového modulu vycházející z rozsahu textových dokumentů zaslaných ke zpracování. Před každým spuštěním proudu s jazykově závislými textminingovými uzly je uživatel informován o aktuální dostupné kvótě a po ukončení výpočtu o aktuálním čerpání. Kvóta je udávána v počtu znaků a v počtu dokumentů.

Jazykově nezávislé procedury, jako například extrakce řetězců pomocí regulárních výrazů nebo měření editační vzdálenosti, se vykonávají lokálně na Modeler Client, případně na Modeler Server. Jazykově nezávislé uzly nečerpají kvótu, nevyžadují internetové připojení a jejich použití není omezeno předplacenou licencí.

Podporované jazyky

ATM je vyvíjen především pro zpracování českých a slovenských textů, avšak podporuje i práci s dokumenty v jazycích angličtina, němčina, francouzština, španělština, portugalština, polština a nizozemština. Pomocí jazykově nezávislých procedur lze zpracovat jakýkoli jazyk, neboť pro tyto procedury je dokument sekvencí znaků a zpracování nevyžaduje jazykově závislé lingvistické zdroje.

Pokud potřebujete zahrnout do svého analytického a predikčního postupu dokumenty v různých jazycích, můžete zapnout automatickou detekci jazyka u jednotlivých dokumentů. Před zpracováním každého dokumentu se nejprve automaticky detekuje jeho jazyk a na základě identifikovaného jazyka se použijí příslušné lingvistické zdroje jako například slovníky, pravidla či znalostní báze. Po zapnutí automatické detekce je rozpoznaný jazyk vrácen uživateli jako nový atribut dokumentu spolu s požadovanými výstupy.

Rozpoznání jazyka je spolehlivé u dlouhých dokumentů, jazyk u krátkých odpovědí, textových zpráv či komentářů se nemusí rozpoznat dobře. Běžně se stává, že chybně identifikovaný jazyk krátkého dokumentu patří mezi nepodporované jazyky, a výsledky jsou založeny pouze na obecných pravidlech nebo nejsou pro chybně identifikovaný dokument dostupné vůbec. Aby se omezilo riziko chybného rozpoznávaní jazyků, je možné automatickou detekci jazyka omezit pouze na vybrané jazyky. Například u multijazyčné kolekce komentářů můžeme detekci omezit na češtinu, slovenštinu a angličtinu, jiné jazyky se detekovat nebudou.

Diakritizace

Volné texty v přirozeném jazyce mohou reprezentovat emaily, záznamy z telefonních center, záznamy o interakci se zákazníky, technické reporty, názory respondentů, žádosti a mnoho dalších. Často se však stává, že v textech úplně nebo částečně chybí diakritika. Před jazykově závislé procedury je možné předřadit proceduru pro automatickou diakritizaci. Diakritizace je podporována pouze pro češtinu. Pokud je zapnuta i automatické detekce jazyka, diakritizace se provádí až po rozpoznání českého jazyka.

Diakritizace nijak nezvyšuje započítávaný objem dat zaslaných na textminingový server, stačí pouze při nastavování uzlů v proudu požádat o potřebnou úpravu. Zpracovávaná kolekce může být směsí dokumentů s diakritikou a bez diakritiky, diakritizace se provádí u každého dokumentu podle potřeby.

Strukturovaná reprezentace volných textů

Hlavním cílem modulu ATM je transformace volných textů do strukturované podoby, aby bylo možné s dokumenty pracovat podobně jako se zákazníky, žádostmi či jinými entitami v běžných dataminingových projektech. Proto se z dokumentů extrahují termíny, jež po vhodné restrukturalizaci slouží jako atributy strukturované datové matice.

Z dokumentů psaných v přirozeném jazyce je možné pomocí procedur ATM extrahovat různé typy atributů. Díky univerzálním regulárním výrazům lze například dokumenty tokenizovat, rozdělit je na n-gramy nebo v nich vyhledávat specifické entity jako jsou čísla, emaily, časy apod.

Jazykově závislé uzly umožní v textu najít informativní atributy podobně, jako by to dělal lidský čtenář. Kromě vyhledávání všech pojmenovaných entit (osoby, místa, organizace apod.) je k dispozici uzel, jež extrahuje z každého dokumentu několik klíčových termínů. Termíny mohou být jednoslovné (např. motor) i víceslovné (např. spalovací motor). Během extrakce jsou termíny automaticky převáděny do základního mluvnického tvaru (lemmatizace), aby nedocházelo ke zbytečnému zvyšování dimenzionality datové matice dokumentů v důsledku ohýbání slov.

Analýza sentimentu dokumentů a pojmenovaných entit

Rozdělení a řazení textových dokumentů podle sentimentu představuje specifickou klasifikační úlohu. Atributy extrahované z dokumentů standardním způsobem v sobě zpravidla nezahrnují dostatečně silnou informaci o pozitivních či negativních postojích autora. Na klasifikaci dokumentů dle sentimentu proto ATM nabízí speciálně předem sestavený model, který ohodnotí každý dokument dle postoje autora. Nejenže se každý dokument zařadí do pozitivní, negativní či neutrální kategorie, ale k dispozici jsou i číselné kvantifikace pozitivního a negativního náboje dokumentu.

Pokud je třeba stanovit sentiment menších částí textových dokumentů, je vhodné před klasifikací sentimentu každý dokument rozdělit například na odstavce. To lze provést pomocí regulárních výrazů nebo hned při načítání textů z textových souborů.

Specifickou úlohou je detekce sentimentu spojeného s pojmenovanými entitami. Hledaný sentiment není vázán na dokument, odstavec či větu, ale kontextově je spjat s konkrétní osobu, firmou či místem. ATM nabízí tuto klasifikaci sentimentu společně s extrakcí pojmenovaných entit. Každou entitu je možné klasifikovat do pozitivní, negativní nebo neutrální kategorie podle toho, jak se o ní v kontextu hovoří. K dispozici jsou i číselná skóre pozitivního a negativního sentimentu entit.

Obohacení strukturovaných dat

Modul ATM umožňuje uživatelům dataminingového softwaru IBM SPSS Modeler zahrnout do svých standardních predikčních postupů další zdroje nestrukturovaných textových dat a využít jejich informační potenciál ke zkvalitnění predikčních modelů. Ačkoli lze nástroji textminingového modulu zpracovávat textové dokumenty samostatně, hlavním přínosem modulu ATM je možnost kombinovat textová data s běžnými strukturovanými daty z databází a datových souborů. Atributy extrahované z textu pomocí uzlů modulu ATM lze snadno restrukturalizovat do datové matice potřebné struktury a granularity a připojit je ke strukturovaným datům. Uživatelé tak získávají informativnější data pro hledání skrytých vzorů chování svých zákazníků, pacientů, strojů atp. a mohou budovat přesnější modely pro řešení svých dataminingových úloh jako jsou detekce podvodů, řízení kreditního rizika, zamezení odchodu ke konkurenci, doporučování produktů, prediktivní údržba a další.

IBM SPSS Modeler Professional svými nástroji pokrývá všechny kroky dataminingového projektu, kdy uživatelé pracují se strukturovanými daty. Například nabízí celou řadu uzlů na realizaci datových manipulací. Díky komplexní podpoře celého procesu od převzetí dat až po export řešení mohou uživatelé ATM využívat velké množství procedur také při práci s nestrukturovanými textovými daty. Textová data lze před odesláním na textminingový server například předzpracovat pomocí standardních funkcí pro modifikaci textových řetězců. Převzaté extrahované atributy z textminingového serveru můžete v Modeleru dále standardními uzly restrukturalizovat, transformovat na jiné veličiny, redukovat jejich dimenzionalitu či napojovat na jiné datové zdroje.

Pokud chcete provádět jen analýzu textu nebo v průběhu přípravy dat vizualizovat vlastnosti zpracovávaných dokumentů, můžete využívat grafy a další výstupní uzly SPSS Modeler. Grafy v SPSS Modeler navíc nabízí interaktivitu spočívající v ad-hoc generování manipulačních uzlů jako jsou filtry nebo odvozování na základě výběru v grafu. Pestrá paleta grafických nativních nástrojů je doplněna výstupními uzly modulu ATM o oblak slov (Word cloud) a zobrazovač náhledu na dokumenty.

Rychlost a paralelizace výpočtů

Výkonný kód uzlů modulu ATM je naprogramován v jazyce C++. Uzly modulu jsou integrovány do SPSS Modeler, tak že pro své fungování nevyžadují další zdroje a technologie a chovají se v SPSS Modeler obdobně jako nativní uzly. To zajišťuje vysokou rychlost výpočtu na stroji, kde je ATM nainstalován. Jazykově závislé uzly však komunikují se vzdáleným textminingovým serverem pomocí zabezpečených webových služeb, což může prodloužit dobu zpracování.

Rychlost zpracování textových dat jazykově závislými uzly lze zvýšit dávkovým a paralelním zpracováním. Při dávkovém zpracování se v jednom dotazu odesílá na textminingový server více dokumentů najednou a také výsledky jsou vráceny v jedné dávce. Tím se sníží čas potřebný na komunikační režii. Při paralelním zpracování se dokumenty na textminingový server posílají paralelně po několika nezávislých vláknech. Nemusí se tak čekat na dokončení zpracování předchozího dokumentu, jednotlivé dotazy jsou na sobě nezávislé. Navíc pokud na textminingový server přichází více současných dotazů, automaticky se pro něj v cloudu alokují dodatečné hardwarové zdroje, a tak se zvyšuje jeho výkon.

Dávkové a paralelní zpracování je implementováno ve všech jazykově závislých uzlech, které komunikují s textminingovým serverem. Velikost dávky a počet paralelních vláken není třeba nastavovat, modul tyto parametry volí automaticky tak, aby bylo dosaženo, co nejvyššího výkonu. Dávkové a paralelní zpracování se spolu kombinuje, každé vlákno zasílá na server dokumenty po dávkách.

Implementované funkce

Jazykově závislé uzly (NLP)

Extrakce termínů (Tags)

Volný text zapsaný v přirozeném jazyce ukrývá množství informace. Aby tato informace mohla být vytěžena pomocí běžných metod strojového učení, je třeba dokumenty popsat sadou strukturovaných atributů. Z každého dokumentu jsou extrahovány termíny vypovídající o jeho obsahu. Termíny je možné využít jako atributy pro strukturovanou reprezentaci textových dokumentů jak v úlohách na zpracování samotných dokumentů, jako jsou klasifikace či seskupování dokumentů, tak v komplexních dataminingových úlohách, jako je například prevence odchodu zákazníka, křížový prodej či detekce podvodů.

Počet potřebných extrahovaných termínů závisí na délce a variabilitě každého dokumentu. Zpravidla není nutné extrahovat všechny termíny, ale omezit se jen na ty nejvíce relevantní. Při extrakci všech termínů by dimenzionalita strukturované reprezentace dokumentů byla příliš vysoká, což se může negativně projevit především při strojovém učení. Uživatel má při extrakci termínu možnost zvolit si jednu ze čtyř variant filtrování termínů podle relevance.

Extrahované termíny zahrnují klíčová slova a názvy pojmenovaných entit uvedená v základním tvaru (lemma). Extrahují se jednoslovné termíny i sousloví. Díky specifickým lingvistickým zdrojům se nemusí jednat o přesné termíny z textu, ale do jednoho termínu mohou být zahrnuta jeho synonyma nebo termín může vyjadřovat plné znění zkratky vyskytující se v textu.

Extrahované termíny se ukládají do datové matice tak, že každý termín nalezený v dokumentu se uloží do nového řádku. Jeden dokument tak generuje několik řádků v nové matici dat, každý řádek odpovídá jednoznačné kombinaci dokumentu a termínu. Dlouhý datový formát je výhodný pro uložení a reportování. Pomocí standardních manipulačních uzlů Modeleru je možné provést výběr a restrukturalizaci nalezených termínů do široké datové matice tak, jak to vyžadují algoritmy strojového učení. Tím vytvoříme strukturovanou reprezentaci celé kolekce dokumentů.

Volitelně lze spolu s termíny extrahovat i jejich lokální číselná skóre kvantifikující relevanci termínu v rámci dokumentu. Skóre může být použito namísto binárních indikátorů termínů v restrukturalizované datové matici dokumentů. Při restrukturalizaci je možné zkonstruovat nebo převést extrahované lokální skóre termínů na jiná běžně používaná skóre, jako je například TF-IDF. Strukturovanou matici dokumentů lze pak snadno napojit na další zdroje strukturovaných dat a při následném modelování tak využít informaci ukrytou jak v databázových datech, tak ve volných textech.

Pokud nejsou všechny dokumenty ve zpracovávané kolekci v českém nebo slovenském jazyce, je možné v uzlu pro extrakci termínů zapnout automatickou detekci jazyka pro každý dokument. Případně lze i omezit množinu jazyků z níž se jazyk při automatické detekci vybírá. Rozpoznaný jazyk se při automatické detekci zaznamenává do nového atributu spolu s každým extrahovaným termínem.

Podobně jako rozpoznání jazyka lze přímo v uzlu před extrakci hesel předřadit i proceduru na obnovení diakritiky pro české texty zapsané bez diakritiky. Diakritizace si poradí i s texty, kde diakritika chybí jen částečně.

Klasifikace a skórování sentimentu (Sentiment)

Určení sentimentu textového dokumentu je jednou z textminingových klasifikačních úloh, kdy dokumenty rozřazujeme do kategorií s pozitivním nebo negativním nábojem. Rozpoznání sentimentu vyžaduje specifické lingvistické zdroje, a proto je vhodné ho realizovat jako samostatnou proceduru a nespoléhat se na obecné klasifikátory pracující s běžnou strukturovanou reprezentací textových dokumentů. Mnohé dokumenty neobsahují sentiment vůbec. Klasifikátor by je měl rozpoznat a zařadit je do speciální kategorie dokumentů bez sentimentu. Modul ATM zařazuje dokumenty do následujících kategorií: velmi negativní, negativní, neutrální, pozitivní, velmi pozitivní a ambivalentní. Ambivalentní dokumenty v sobě ukrývají zároveň pozitivní i negativní sentiment.

Kromě zařazení dokumentu do příslušné kategorie je často potřeba sentiment obsažený v textu kvantifikovat. Číselné skóre úměrné pozitivnímu či negativnímu náboji dokumentu umožní dokumenty řadit a soustředit se pouze na ty nejvíce emotivní. Díky tomu, že k dispozici jsou kromě celkového skóre i samostatná skóre pro pozitivní a negativní sentiment, můžeme identifikovat i ambivalentní dokumenty. Ty by na základě celkového skóre mohly být vyhodnoceny jako neutrální, avšak od dokumentů bez sentimentu se odlišují a často bývají pro uživatele cenným zdrojem informace.

Přiřazení sentimentu k dokumentům nevyžaduje restrukturalizaci datové matice, kategorie sentimentu a jeho skóre se zaznamenávají do nových proměnných. Každý dokument se zařadí do jedné kategorie sentimentu na základě získaného skóre. Celkové skóre se pohybuje na reálné škále od mínus jedné do plus jedné. Dílčímu pozitivnímu skóre je vyhrazena škála od nuly do plus jedné, dílčí negativní skóre nabývá hodnot mezi mínus jedna a nula.

Při současném využití detekce sentimentu a extrakce termínů z dokumentů je možné dokumenty podrobněji klasifikovat do specifických pozitivních a negativních kategorií. Buď se dokumenty nejprve roztřídí podle sentimentu na pozitivní a negativní a pak se pro každou skupinu sestaví klasifikační model, nebo se vytvoří jeden souhrnný klasifikační model, do kterého kromě extrahovaných hesel vstoupí i rozpoznaný sentiment, a výsledné kategorie se interpretují s přihlédnutím k převládajícímu sentimentu v kategorii.

Pokud nejsou všechny dokumenty ve zpracovávané kolekci v českém nebo slovenském jazyce, je možné v uzlu pro rozpoznání sentimentu zapnout automatickou detekci jazyka pro každý dokument. Případně lze i omezit množinu jazyků z nichž se jazyk při automatické detekci vybírá. Rozpoznaný jazyk se při automatické detekci zaznamenává spolu se sentimentem do nových atributů ke každému dokumentu.

Podobně jako rozpoznání jazyka lze přímo v uzlu před analýzu sentimentu předřadit i proceduru na obnovení diakritiky pro české texty zapsané bez diakritiky. Diakritizace si poradí i s texty, kde diakritika chybí jen částečně.

Rozpoznávání pojmenovaných entit (Entities)

Mezi pojmenované entity se v modulu ATM řadí především jména osob, organizací a lokalit. Pojmenované entity v textu určují, kdo něco vykonal, kde se stala nějaká událost apod. Identifikaci pojmenovaných entit není možné provést fulltextovým vyhledáváním, neboť předem nevíme, které konkrétní entity se budou v dokumentech vyskytovat.

Extrahovaná jména entit se uvádí v základním tvaru (lemma). Pojmenované entity mohou být použity podobným způsobem jako termíny extrahované uzlem Tags pro strukturovanou reprezentaci dokumentů nebo pro obohacení datové matice v komplexních dataminingových úlohách. Na rozdíl od termínů se však v uzlu Entities nevybírají pouze ty nejdůležitější entity, ale procedura najde v každém dokumentu vždy všechny entity a je na uživateli, aby si zvolil, které z nich si ponechá. Extrakci entit lze však omezit pouze na osoby, firmy a lokality. Typ pojmenované entity je spolu s názvem nalezené entity indikován v nové kategorizované proměnné.

Rozpoznané entity se ukládají do datové matice tak, že každá entita se uloží do nového řádku. Jeden dokument tak generuje několik řádků v nové matici dat, každý řádek odpovídá jednoznačné kombinaci dokumentu a entity. Dlouhý datový formát je výhodný pro uložení a reportování. Pomocí standardních manipulačních uzlů Modeleru je možné provést výběr a restrukturalizaci nalezených entit do široké datové matice tak, jak to vyžadují algoritmy strojového učení.

Z extrahovaných entit je možné sestavit sociální síť. Entity vyskytující se v textu blízko sebe nebo ve specifickém kontextu utvoří příslušné vazby. Sítě menšího rozsahu lze graficky znázornit spojnicovým grafem. Malé i rozsáhlejší sítě lze zpracovávat a analyzovat standardními manipulačními a analytickými uzly SPSS Modeler nebo využít jeho speciální modul pro analýzu sociálních sítí.

Kromě jména entity a jejího typu lze získat i sentiment entity. Sentiment entity určuje, zda se o osobě, organizaci či lokalitě píše v dokumentu pozitivně, neutrálně nebo negativně. Kromě kategorií sentimentu entit je možné si nechat spočítat i skóre kvantifikující míru sentimentu. Číselné skóre úměrné pozitivnímu či negativnímu náboji výpovědi umožní entity řadit. K dispozici jsou kromě celkového skóre i samostatná skóre pro pozitivní a negativní sentiment entit. Díky dílčím skóre lze identifikovat i ambivalentní výpovědi o pojmenovaných entitách. Zatímco celkové skóre sentimentu entit se pohybuje na škále od mínus jedné do plus jedné, dílčímu pozitivnímu skóre je vyhrazena škála od nuly do jedné a dílčí negativní skóre nabývá hodnot mezi mínus jedna a nula.

Pokud nejsou všechny dokumenty ve zpracovávané kolekci v českém nebo slovenském jazyce, je možné v uzlu pro rozpoznávání pojmenovaných entit zapnout automatickou detekci jazyka pro každý dokument. Případně lze i omezit množinu jazyků z nichž se jazyk při automatické detekci vybírá. Rozpoznaný jazyk se při automatické detekci zaznamenává do nového atributu spolu s každou extrahovanou entitou.

Podobně jako rozpoznání jazyka lze přímo v uzlu před extrakci entit předřadit i proceduru na obnovení diakritiky pro české texty zapsané bez diakritiky. Diakritizace si poradí i s texty, kde diakritika chybí jen částečně.

Jazykově nezávislé uzly (stringologie)

Regulární výrazy (Regular expressions)

Při automatické analýze textu je často potřeba v textových dokumentech vyhledat specifické podřetězce. Při fulltextovém vyhledávání je nezbytné hledaný řetězec přesně zadat. Často však potřebujeme v textu vyhledat řetězce, jež není možné všechny explicitně zadat pro fulltextové hledání. Například pokud chceme vyhledat všechny emailové adresy, není možné je předem všechny vyjmenovat. K vyhledání řetězců, které musí splňovat určitá pravidla, ale specifikace všech jejich variant je obtížná, se používají regulární výrazy. Regulární výraz je řetězec obsahující speciální znaky, jimiž lze popsat širší množinu řetězců. Speciální znaky mohou nahrazovat množinu běžných znaků, specifikovat opakování znaků či podřetězců nebo vymezovat abstraktní pozice v textu.

Regulární výrazy je vhodné použít nejen k vyhledání řetězců speciálního typu, jako jsou například zmiňované emaily, ale i k rozdělení textového dokumentu na složky, jako jsou věty, slova, tokeny či n-gramy. Například tokeny nebo slova extrahovaná z dokumentů za pomoci vhodného regulárního výrazu se mohou stát základem pro strukturovanou reprezentaci dokumentů, kterou je možné v softwaru realizovat bez jazykově závislých uzlů.

Syntaxe regulárních výrazů zahrnující používání speciálních znaků má svá pravidla. Pro vytváření a editaci regulárního výrazu je k dispozici kalkulačka. Kalkulačka regulárních výrazů umožňuje do výrazů intuitivně vkládat speciální znaky a kontrolovat syntaktickou správnost výrazů. Uživatel nemusí speciální znaky znát, stačí na kalkulačce zvolit tlačítko s popisem funkce speciálního znaku. Syntaktická správnost zadávaného výrazu je po stisknutí kontrolního tlačítka znázorněna barvou textu.

Řetězce odpovídající zadanému regulárnímu výrazu se ukládají do datové matice tak, že každý řetězec se uloží do nového řádku. Jsou-li nalezeny dva stejné řetězce v jednom dokumentu, každý se uloží do jednoho řádku. Jeden dokument tak generuje několik řádků v nové matici dat, každý řádek odpovídá jednoznačné kombinaci dokumentu, nalezeného řetězce a pozice nalezeného řetězce. Pozici nalezeného řetězce je možné na vyžádání do datové matice také přidat. Vzniknou tak dva nové celočíselné atributy udávající pozici prvního a posledního znaku nalezeného řetězce v rámci dokumentu.

Editační vzdálenost (Edit distance)

Při zpracování krátkých textových výpovědí bývá úkolem vyhodnotit shodu dvou odpovědí nebo shodu odpovědi s konkrétním řetězcem. Například při ztotožňování jmen firem se snadno může stát, že názvy obsahují překlepy. Proto není vhodné vyhodnocovat přesnou shodu názvů, stačí když si uvedené názvy firem budou podobné.

Podobnost, resp. nepodobnost dvou textových řetězců se hodnotí podle počtu editačních operací nutných k tomu, aby z jednoho řetězce vznikl druhý. Podle druhu přípustných editačních operací, jako jsou smazání, vložení či záměna sousedních znaků, rozlišujeme různé celočíselné i neceločíselné míry nepodobnosti. Zatímco celočíselné metriky vyjadřují přímo počet nutných editačních operací, u reálných metrik se hodnoty standardizují na omezenou škálu od nuly do jedné. Uzel na výpočet editační vzdálenosti nabízí tyto celočíselné míry nepodobnosti: Levenhsteinovu, Damerau-Levenhsteinovu a LCS (Longest Common Subsequence) Dále uzel nabízí následující reálné míry nepodobnosti: Jarovu, Jaro-Winklerovu, Fuzzy Wuzzy, Jaccardovu a Dice-Sørensenovu.

Uzel pro výpočet editační vzdálenosti nemění strukturu datové matice, ale jednoduše přidá novou proměnnou s vypočtenou vzdáleností mezi dvěma řetězci uloženými v textových proměnných. Hledáme-li podobnost textu s předem daným konstantním řetězcem, předřadíme před výpočet editační vzdálenosti standardní uzel pro výpočet nového atributu, který vloží konstantní řetězec do nové porovnávané proměnné.

Při současném použití uzlu editační vzdálenosti s uzlem pro spojování datových matic lze snadno implementovat hledání podobnosti textového řetězce s libovolným řetězcem z předem daného seznamu či slovníku. Například můžeme hodnotit, jak se u respondentů shoduje spontánní znalost značek aut se značkami z daného soupisu.

Vstupy a výstupy

Extrakce dokumentů z textových souborů (Text files)

Uzly modulu ATM na zpracování textových dokumentů akceptují textová data vložená do textových proměnných. Textové proměnné s dokumenty mohou do proudů vstoupit standardními vstupními uzly softwaru SPSS Modeler. Pokud jsou však dokumenty určené na zpracování uloženy do separátních nestrukturovaných textových souborů, je možné je hromadně načíst speciálním uzlem modulu ATM.

Uzel ve vybrané složce hledá soubory určitého typu. Prohledávat lze všechny typy souborů, textové soubory nebo soubory XML a HTML. Prohledávání je možné omezit pouze na vybranou složku nebo prohledávat i podsložky. XML a HTML soubory se při načítání nemodifikují a nefiltrují.

Z nalezených souborů se extrahují textové dokumenty do zvolené textové proměnné. Kratší dokumenty se ukládají celé do jedné datové buňky, u delších dokumentů je vhodné zvolit odstavcový mód, kdy se každý neprázdný odstavec vloží jako samostatný řádek do datové matice. Každý dokument je v datové matici identifikován názvem souboru, z něhož byl extrahován. U odstavcového módu se přidá do datové matice i proměnná s pořadím odstavce v rámci dokumentu.

Jednotlivé znaky tvořící dokumenty jsou v textových souborech kódovány. K zakódování textů se používá Unicode nebo jednobytové kódovací tabulky zpravidla určené pro konkrétní jazyk nebo skupinu jazyků. Například kódovací tabulka CP-1250 je určena pro středoevropské jazyky včetně češtiny a slovenštiny.

Pokud je k zakódování použit celosvětový standard Unicode, kódy vzhledem ke své délce nejsou zpravidla ukládány do textového souboru přímo, ale používá se některý ze způsobů komprese, kdy jeden znak může zabírat různý počet bytů. Například pokud je text komprimován jako UTF-8, většina písmen české abecedy zabírá pouze jeden byte, písmena s diakritikou dva bajty. Uzel pro načítání textových souborů umožňuje zvolit dekódování z libovolného způsobu komprese Unicode, a navíc nabízí dekódování i z několika běžných kódovacích tabulek.

Oblak slov (Word cloud)

Extrahované atributy z textových dokumentů jako jsou slova, termíny nebo pojmenované entity nesou v sobě informaci o obsahu kolekce dokumentů a používají se jako vstupní atributy do textminingových modelů. Díky rozmanitosti přirozených jazyků nalezneme ve volných textech velké množství různých slov, termínů, entit apod. Výběr klíčových atributů je zásadní jak pro pochopení, o čem se v dokumentech píše, tak pro vybudování robustních predikčních modelů.

Na extrahované texty můžeme nahlížet jako na kategoriální proměnnou s velkým množstvím možných kategorií. Pro vizualizaci rozdělení takové proměnné lze použít například sloupcový graf četností, ale pro rychlé zjištění hlavních témat v dokumentech se lépe hodí obrázek nazývaný oblak slov. V oblaku slov jsou zobrazovaným termínům přiděleny náhodné pozice v obrázku při respektování omezení, že termíny se nesmí překrývat a vyčnívat mimo hranice obrázku. Termíny se v oblaku vykreslují s různou velikostí písma podle četnosti výskytu. Oblak je možné dále vylepšit tak, že texty mohou být horizontální i vertikální, různě obarvené, mít různý font, řez písma apod. Často se také používají oblaky slov různého tvaru.

Oblak slov se v modulu ATM zobrazuje v samostatném výstupovém okně v grafickém formátu. Termíny, které se umísťují do obrázku, je třeba předem z textových dokumentů extrahovat např. pomocí uzlu Tags, Entities nebo Regular expression. Velikost písma v oblaku závisí na četnosti nalezených termínů, volitelně lze k termínům připojit jejich lokální váhy pro výpočet vážených četností. Škálu velikosti písma uživatel volí pomocí minimální a maximální hodnoty a převodní funkce mezi četností a velikostí písma. Převodní funkce může být lineární, logaritmická nebo odmocnina. Pro zvýraznění četnostních rozdílů je vhodné i texty obarvit. Barevná škála se také stanovuje pomocí barvy pro nejméně četné termíny, barvy pro nejvíce četné termíny a převodní funkce mezi barvou a četností. Převodní funkce jsou stejné jako pro škálu velikosti písma.

Oblak slov může mít různý tvar. Na výběr jsou: obdélník, obdélník se zakulacenými rohy, ovál, půlkruh a kosočtverec. Směr písma může být pouze horizontální nebo kombinace horizontálního a vertikálního.

Při vykreslování oblaku slov se do obrázku nejdříve umísťují termíny zapsané velkým písmem a zbylé plochy obrázku se postupně zaplňují termíny zapsanými menším písmem. Takto se postupuje, dokud v obrázku nedojde místo. Je běžné, že se do obrázku nevejdou všechny extrahované termíny. Může nastat i opačná situace, že v obrázku po zobrazení všech termínů zbydou prázdné plochy. Estetický dojem a informační potenciál oblaku slov pak ovlivňuje nejen volba škály velikosti písma, ale i rozměry obrázku. Velikost obrázku lze měnit interaktivně změnou velikosti výstupového okna s oblakem slov. Při každé změně velikosti okna se oblak slov automaticky překreslí.

Při umísťování textu do obrázku hraje roli náhoda. Do stejně velkého okna při neměnném nastavení škál můžeme vygenerovat velké množství různých variant oblaků slov. Pokud chceme zkusit jinou variantu oblaku stačí jen mírně změnit velikost okna nebo jednoduše stisknout tlačítko pro překreslení na panelu vlevo nad obrázkem. Kromě překreslení můžeme pomocí tlačítek obrázek zkopírovat do schránky, uložit ve formátu PNG nebo uložit termíny a jejich četnosti jako CSV datový soubor.

Náhled na dokumenty (Document viewer)

Při řešení textminingových úloh se zpravidla zpracovávají rozsáhlé kolekce textových dokumentů. Ačkoli manuální pročítání všech textů bývá nemožné nebo neefektivní, během analýzy je vhodné si některé dokumenty přečíst. Například po rozdělení recenzí podle sentimentu a témat si přečteme jen negativní recenze na zákaznický servis.

Do modulu ATM vstupují dokumenty jako textové proměnné v datové matici. Standardní zobrazení datové matice uzly SPSS Modeler je pro čtení dokumentů většinou nepohodlné, protože dokumenty přetékají datové buňky a nedají se zkrátit. Uzel pro náhled textových dokumentů umožňuje prohlížené dokumenty nejen omezit na zadaný počet znaků, ale i rozdělit na části a doplnit metadaty a hypertextovými odkazy.

Textové dokumenty se zobrazují v samostatném výstupovém okně v podobném HTML formátu, na jaký jsme zvyklí u vyhledávačů. Zobrazené dokumenty lze rozdělit na nadpis, perex a tělo dokumentu. Délku perexu a těla dokumentu je možné omezit na zadaný počet znaků, nadpis může být spojen s hypertextovým odkazem. Počet zobrazených dokumentů lze též limitovat. Omezení na počet zobrazených dokumentů se hodí při prohlížení rozsáhlých kolekcí, abychom nevygenerovali příliš velký HTML soubor, se kterým by měl prohlížeč výkonnostní potíže.

Kromě samotných textových dat můžeme před každý náhled dokumentu připojit celou řadu metadat: identifikátor dokumentu, zdroj, autor, datum, jazyk, kategorie, sentiment, klíčová slova, skóre, počet znaků. Každý z těchto údajů má editovatelnou legendu, proto může být použit i pro zobrazení jiných méně obvyklých metadat. Identifikátor dokumentu, zdroj a autora můžeme podobně jako nadpis spojit s hypertextovým odkazem.

HW a SW požadavky

- 64bitový operační systém Windows
- nainstalovaný software IBM SPSS Modeler Professional v lokální nebo serverové verzi
- trvalé připojení k internetu
- 40 MB volné místo na disku pro instalaci