Modul Acrea Text Mining 2.0

Textminingový nástroj Acrea Text Mining 2.0 (ATM) je modulem dataminingového softwaru IBM SPSS Modeler. Tvoří ho soubor 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.

Integrace do SPSS Modeler

Textminingové procedury jsou integrovány do prostředí IBM SPSS Modeler jako uzly. Při práci se 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 Python.

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. Modul ATM navíc nabízí speciální uzel pro načítání textů z nestrukturovaných textových souborů.

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, 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.

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 takových textech chybí diakritika. Před jazykově závislé procedury je možné předřadit proceduru pro automatickou diakritizaci. Tento přípravný krok nijak nezvyšuje započítávaný objem dat zaslaných na textminingový server, stačí pouze při nastavování uzlů v proudu požádat o potřebnou úpravu.

ATM je vyvíjen především pro zpracování českých a slovenských textů, avšak podporuje i práci s dokumenty v jiných jazycích. Pokud potřebujete zahrnout do svého analytického a predikčního postupu dokumenty v různých jazycích, můžete zapnout automatickou detekci jazyka u jednotlivých dokumentů. Na základě identifikovaného jazyka se pro každý dokument použijí příslušné lingvistické zdroje jako například slovníky, pravidla či znalostní báze. Po zapnutí automatické detekce je rozpoznaný jazyk dokumentu vrácen uživateli jako nový atribut dokumentu.

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 (osob, míst a organizací) je k dispozici uzel, jež extrahuje z každého dokumentu několik klíčových termínů. Termíny mohou být jednoslovné i víceslovné. 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 a identifikace pojmenovaných entit

Třídění a řazení textových dokumentů podle sentimentu představuje specifickou klasifikační úlohu. Atributy extrahované z dokumentů standardním způsobem zpravidla v sobě nezahrnují dostatečně silnou informaci o pozitivních či negativních postojích autora. Na klasifikaci dokumentů dle sentimentu proto ATM nabízí speciálně připravenou proceduru, která ohodnotí každý dokument dle postoje pisatele. 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 každý dokument rozdělit 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 textu hovoří. K dispozici jsou i číselná skóre pozitivního a negativního sentimentu entit.

Obohacení strukturovaných dat

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 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 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 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 zdroje dat.

Implementované funkce

Jazykově závislé uzly (NLP)

Sémantické značkování dokumentů (Labels)

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. Počet extrahovaných termínů

závisí na délce a variabilitě každého dokumentu. Termíny je možné využít jako strukturovanou reprezentaci textových dokumentů jak v úlohách na zpracování samotných dokumentů, jako jsou klasifikace či klastrování dokumentů, tak v komplexních dataminingových úlohách, jako například prevence odchodu zákazníka, křížový prodej či detekce podvodů. Extrahované termíny zahrnují klíčová slova a názvy pojmenovaných entit uvedená v základním tvaru. 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. Kromě klíčových slov v termínech najdeme i vybraná jména osob, firem a míst. Typy termínů jako jsou klíčové slovo nebo pojmenovaná entita mohou být indikovány v nové kategorizované proměnné.

Pomocí standardních manipulačních uzlů Modeleru je možné provést výběr a restrukturalizaci hesel 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 číselná skóre kvantifikující důležitost 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 skóre termínů na jiná běžně používaná skóre jako je například TF-IDF. Strukturovanou matici dokumentů lze pak snadno napojit na jiné 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. Podobně jako rozpoznání jazyka lze přímo v uzlu před extrakci hesel předřadit i proceduru na obnovení diakritiky pro 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 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.

Kromě zařazení dokumentu do sentimentální 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. Ambivalentní dokumenty obsahují jak pozitivní, tak negativní náboj. Celkově by tedy 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 z kategorií: velmi pozitivní, pozitivní, neutrální, negativní, velmi negativní, ambivalentní. Celkové skóre se pohybuje na škále od mínus jedné do plus jedné. Dílčímu pozitivnímu skóre je vyhrazena škála od nuly do 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 v klasifikované kolekci v českém či slovenském jazyce, je možné v uzlu pro rozpoznání sentimentu zapnout automatickou detekci jazyka pro každý dokument. Rozpoznaný jazyk je pak zaznamenán 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 texty zapsané bez diakritiky. Diakritizace si poradí i s texty, kde diakritika chybí jen částečně.

Rozpoznávání pojmenovaných entit (Entities)

Mezi pojmenované entity se v AML řadí jména osob, organizací a míst. 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.

Pojmenované entity mohou být použity podobným způsobem jako termíny extrahované při sémantickém značkování dokumentů pro strukturovanou reprezentaci dokumentů nebo pro obohacení datové matice v komplexních dataminingových úlohách. Na rozdíl od termínů se však nevybírají pouze ty nejdůležitější entity, ale procedura najde v každém dokumentu všechny entitiy a je na uživateli, aby si zvolil, které z nich si ponechá.

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.

Extrahovaná jména entit se uvádí v základním tvaru (lemmatizace). Extrakce entit je omezena na osoby, firmy a místy. Typ pojmenované entity může být indikován v nové kategorizované proměnné.

Kromě jména entity a jejího typu lze získat i sentiment entity. Sentiment entity určuje, zda se o osobě, organizaci či místě 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 extrakci pojmenovaných entit zapnout automatickou detekci jazyka pro každý dokument. Podobně jako rozpoznání jazyka lze přímo v uzlu před extrakci hesel předřadit i proceduru na obnovení diakritiky pro texty zapsané bez diakritiky. Diakritizace si poradí i s texty, kde diakritika chybí jen částečně.

Jazykově nezávislé uzly (stringologie)

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.

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.

Jednotlivé znaky dokumentů 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. Uzel pro načítání textových souborů umožňuje zvolit dekódování z libovolného způsobu komprese Unicode, a navíc nabízí dekódování i z několika běžných kódovacích tabulek.

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 intuitivně do výrazů 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 extrahují do nové proměnné v restrukturalizované datové matici. Kromě proměnné s extrahovaným řetězcem se v restrukturalizované datové matici nachází identifikátor původního dokumentu. Na vyžádání je možné vložit i proměnné s pozicí začátku a konce vyhledaného řetězce v původním 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 název obsahuje překlepy, proto není vhodné vyhodnocovat přesnou shodu názvů, stačí když si uvedené názvy firem budou podobné.

Podobnost shody 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 řetězců. Některé z nich vyjadřují přímo počet nutných editačních operací, jiné míru standardizují na omezenou

škálu. Uzel na výpočet editační vzdálenosti nabízí tyto míry: Levenhsteinovu, Damerau-Levenhsteinovu, LCS a Jaro-Winklerovu. Zatímco první tři míry přímo počítají nutné editační operace, Jaro-Winklerova míra je omezena na interval od nuly do jedné a navíc přikládá vyšší důležitost shodám na začátku řetězců. 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.

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

Zvýšení výkonu

Výkonný kód uzlů modulu ATM je naprogramován v jazyce C++ a zkompilován. To zajišťuje vysokou rychlost výpočtu na stroji, kde je ATM nainstalován. Jazykově závislé uzly však komunikují se vzdáleným textminingovým serverem, 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 tetxminingový server posílají po několika nezávislých vláknech. Nemusí se tak čekat na dokončení zpracování předchozího dokumentu, jakmile některé z vláken skončí výpočet, začne zpracovávat další volný dokument. Pokud na textminingový server přichází více současných dotazů, automaticky se mu v cloudu alokují dodatečné hardwarové zdroje, a tak se zvyšuje jeho výkon.

Dávkové i paralelní zpracování je implementováno ve všech jazykově závislých uzlech. U dávkového zpracování uživatel volí velikost dávky, tj. počet dokumentů odeslaných na server v jednom dotazu. U paralelního zpracování uživatel řídí počet nezávislých vláken alokovaných pro výpočet. Dávkové a paralelní zpracování je možné spolu kombinovat, každé vlákno pak zasílá na server dokumenty po dávkách.

HW a SW požadavky

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