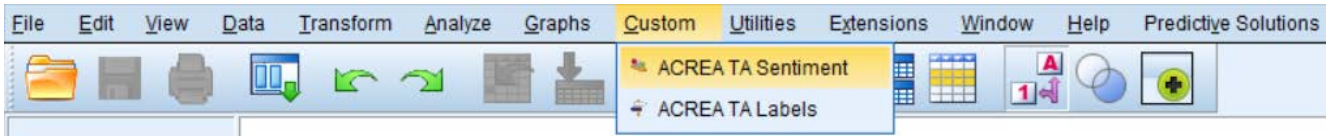
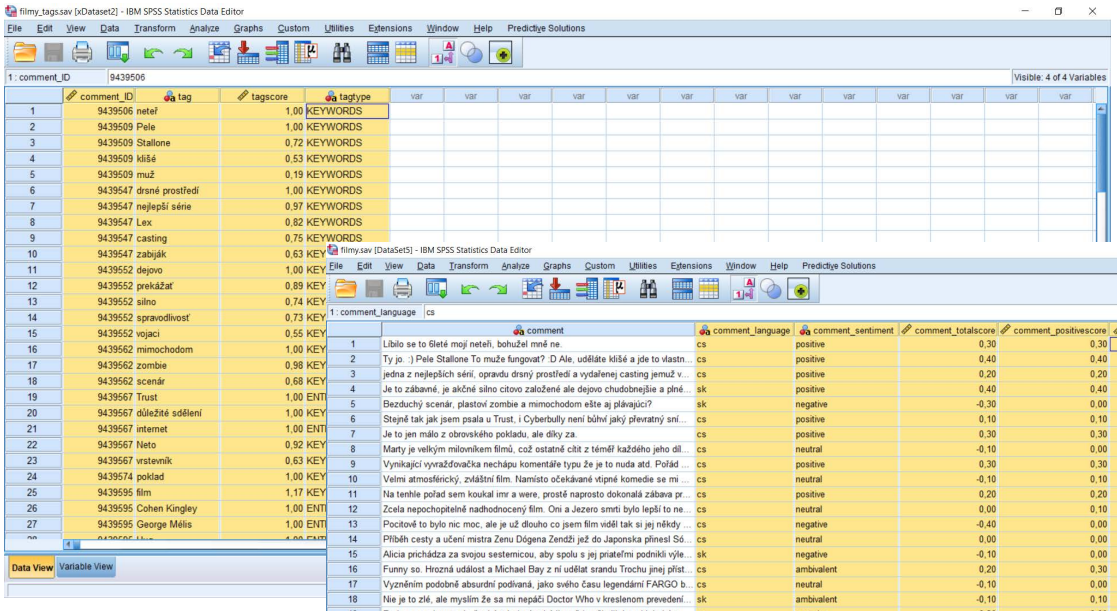
ACREA Text Analytics

Analytický nástroj **Acrea Text Analytics (ATA)** je modulem řešení PS IMAGO PRO, jehož součástí je populární statistický software IBM SPSS Statistics. 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 především z českých a slovenských textů a klasifikaci dokumentů podle jejich sentimentu. Procedury Analýza sentimentu a Extrakce termínů (Labels) jsou integrovány do prostředí IBM SPSS Statistics jako položky v menu Custom.

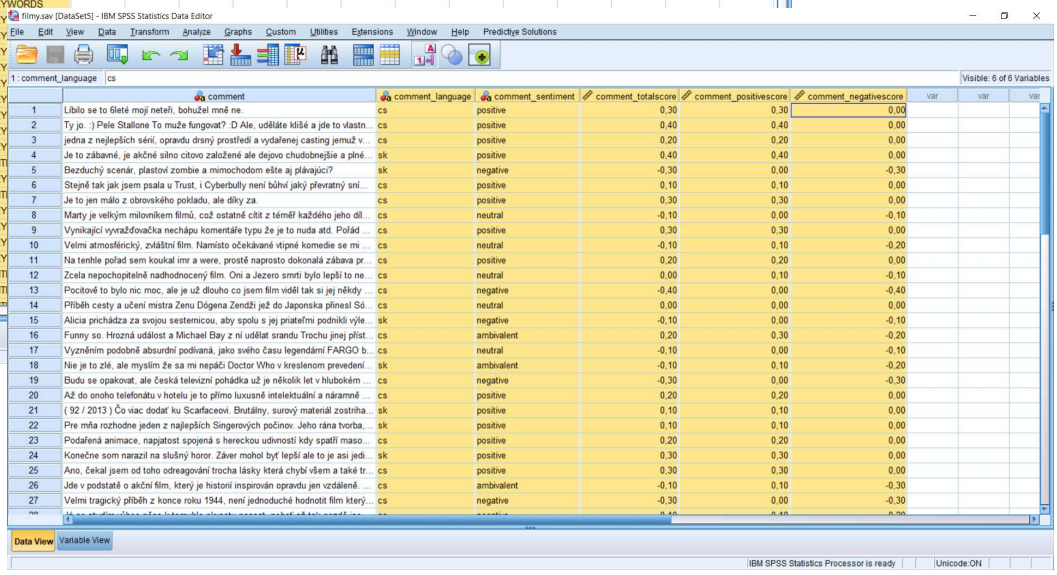


Volné texty v přirozeném jazyce mohou reprezentovat **emaily, přepisy hovorů z call center, záznamy o interakci se zákazníky, názory respondentů, žádosti** a mnoho dalších. Často se však stává, že v takových textech chybí diakritika. Před extrakci strukturovaných atributů či klasifikaci podle sentimentu je možné předřadit proceduru pro automatické obnovení diakritiky.

Vlastní zpracování textu se neprovádí na serveru či klientské stanici, kde je nainstalován IBM SPSS Statistics, 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.



Extrakce termínů



Analýza sentimentu

ATA je vyvíjen především pro zpracování českých a slovenských textů, avšak podporuje i práci s dokumenty v jiných jazycích. Pokud potřebujete zahrnout do svého analytického 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.

ATA je schopen zpracovávat dokumenty z různých domén jako jsou finance, medicína, telekomunikace, výroba, média apod. Při zpracování volného textu s názory zákazníků a uživatelů je možné využít specifické lingvistické zdroje a extrahovat z textu jiné typy hesel, a dokonce i identifikovat vztahy mezi nalezenými termíny. Další specificky zpracovávanou doménu představují média, kdy se při extrakci soustřeďujeme především na pojmenované entity, jakými jsou osoby a firmy.

Implementované funkce

Extrakce termínů (Labels)

Volný text zapsaný v přirozeném jazyce ukrývá množství informace. Aby tato informace mohla být vytěžena pomocí běžného strojového učení, je třeba dokumenty popsat sadou strukturovaných atributů. Z každého dokumentu jsou extrahovány hesla vypovídající o jeho obsahu. Počet extrahovaných hesel závisí na délce a variabilitě každého dokumentu. Hesla je možné využít jako strukturovanou reprezentaci textových dokumentů v úlohách na zpracování samotných dokumentů, jako jsou klasifikace či klastrování dokumentů.

Extrahovaná hesla zahrnují klíčová slova uvedená v základním tvaru. Obsahují jak jednoslovné termíny, tak sousloví. Díky specifickým lingvistickým zdrojům se nemusí jednat o přesné termíny z textu, ale do jednoho hesla mohou být shrnuta synonyma nebo například heslo může vyjadřovat plné znění zkratky vyskytující se v textu. Kromě klíčových slov v heslech najdeme jména osob, firem či míst.

Analýza sentimentu

Určení sentimentu textového dokumentu je jednou z klasifikačních úloh, kdy dokumenty rozřazujeme do kategorií pozitivního a negativního sentimentu. Rozpoznání sentimentu vyžaduje specifické lingvistické zdroje, 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 třeba 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 sentiment a skóre pro negativní sentiment, můžeme identifikovat 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.

Kategorie sentimentu může být velmi pozitivní, pozitivní, neutrální, negativní a velmi negativní. Celkové skóre se pohybuje na škále od mínus jedné do plus jedné. Pozitivní hodnoty indikují pozitivní náboj, negativní hodnoty jsou vyhrazeny pro negativní náboj. Celkové skóre vzniká jako součet pozitivního a negativního skóre. Obě komponenty jsou také k dispozici, jejich nenulové hodnoty indukují ambivalentní dokumenty.

Při současném využití detekce sentimentu a extrakce hesel z dokumentů je možné dokumenty podrobněji klasifikovat do specifických pozitivních a negativních kategorií. Buď se nejprve dokumenty 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.

HW a SW požadavky

- 64bitový operační systém Windows,
- software IBM SPSS Statistics nebo PS Imago PRO,
- trvalé připojení k internetu.