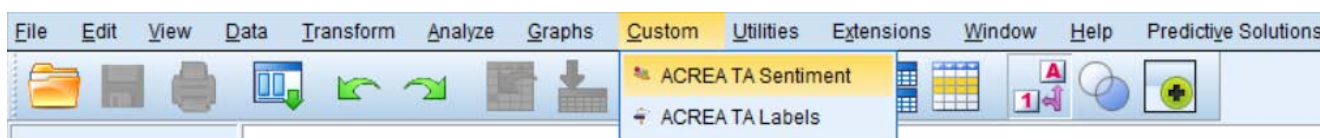


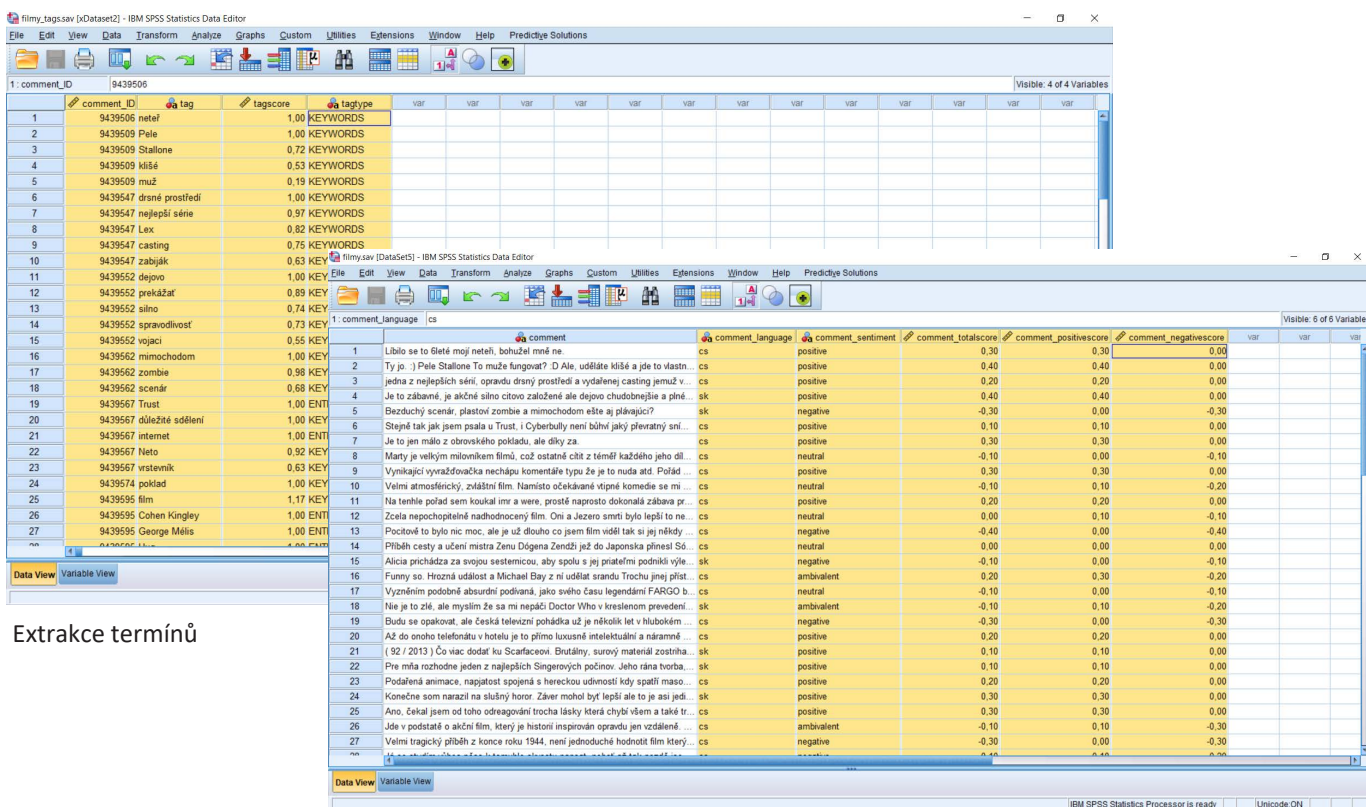
ACREA Text Analytics

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 předradit 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.



comment_ID	tag	tagtype	comment	comment_language	comment_sentiment	comment_totalscore	comment_positivescore	comment_negativescore
1	9439506 neteř	1.00 KEYWORDS	Liblo se te šleté moji neteři, bohužel mně ne.	cs	positive	0.30	0.30	0.00
2	9439509 Pele	1.00 KEYWORDS	Ty jo :) Pele Stallone To může fungovat? :D Ale, udiště klíšé a jde to vlastn...	cs	positive	0.40	0.40	0.00
3	9439509 Stallone	0.72 KEYWORDS	jedna z nejlepších sérií, opravdu drsný prostředí a vydálený casting jemuž v...	cs	positive	0.20	0.20	0.00
4	9439509 klíšé	0.53 KEYWORDS	Je to zábavné, je akčné silno citové založené ale dejvo chudobnejšie a plíné...	sk	positive	0.40	0.40	0.00
5	9439509 muž	0.19 KEYWORDS	Bezduchý scénár, plastoví zombie a mimochodom ešte aj plávajúci?	sk	negative	-0.30	0.00	-0.30
6	9439547 drsné prostředí	1.00 KEYWORDS	Stejně tak jak jsem psala u Trust, i Cyberbully není bůhví jaký převratný sni...	cs	positive	0.10	0.10	0.00
7	9439547 nejlepší série	0.97 KEYWORDS	Je to jen málo z obrovského pokladu, ale díky za...	cs	positive	0.30	0.30	0.00
8	9439547 Lex	0.62 KEYWORDS	Marty je velkým mlounkem filmů, což ostatně čít z téměř každého jeho díl...	cs	neutral	-0.10	0.00	-0.10
9	9439547 casting	0.76 KEYWORDS	Vynikající vyražďovačka nechápu komentáře typu že je to tuča atd. Pořád...	cs	positive	0.30	0.30	0.00
10	9439547 zabíjak	0.63 KEYWORDS	Veľmi atmosférický, zážitní film. Namisto očakávaného výpne komedie se mi...	cs	neutral	-0.10	0.10	-0.20
11	9439552 deňovo	1.00 KEYWORDS	Na tenhle pořad sem koukala mr a webu, prostě naprosto dokonalá zábava pr...	cs	positive	0.20	0.20	0.00
12	9439552 pokračat	0.89 KEYWORDS	Zcela nepochopitelné nadhodnocení film. Oni a bezesro smrti bylo lepší to ne...	cs	neutral	0.00	0.10	-0.10
13	9439552 silno	0.74 KEYWORDS	Pochová to bylo nic moc, ale je už dlouho co jsem film viděl tak si jej někdy...	cs	negative	-0.40	0.00	-0.40
14	9439552 spravedlivost	0.73 KEYWORDS	Příběh cesty a učení mistra Ženu Džena Zenzú jež do Japonska přinesl Sč...	cs	neutral	0.00	0.00	0.00
15	9439552 vojáci	0.56 KEYWORDS	Alicia přichádza za svojou sestenicou, aby spolu s jej priateľmi podnikli výle...	sk	negative	-0.10	0.00	-0.10
16	9439552 mimochodom	1.00 KEYWORDS	Funny so. Hrozná událost a Michael Bay z ní udělat srandu Trochu jinéj příst...	cs	ambivalent	0.20	0.30	-0.20
17	9439552 zombie	0.98 KEYWORDS	Vyzněním podobně absurdní podivná, jako svého času legendární FARGO b...	cs	neutral	-0.10	0.00	-0.10
18	9439552 scénár	0.68 KEYWORDS	Nie je to zlé, ale myslím že sa mi nepáči Doctor Who v kreslenom prevedení...	sk	ambivalent	-0.10	0.10	-0.20
19	9439557 Trust	1.00 ENT	Budu se opakovat, ale česká televizní pohádka už je několik let v hlubokém ...	cs	negative	-0.30	0.00	-0.30
20	9439567 důležitě sdělení	1.00 ENT	Až do onoho telefonátu v hotelu je to přímo luxusně intelektuální a náramně ...	cs	positive	0.20	0.20	0.00
21	9439567 intarnt	1.00 ENT	(92 / 2013) Čo viac dodať ku Scarfaceovi. Brutálny, surový materiál zostriha...	cs	positive	0.10	0.10	0.00
22	9439567 Neto	0.92 KEYWORDS	Pre mňa rozhodne jeden z najlepších Singeroových počínov. Jeho rána tvorba...	sk	positive	0.10	0.10	0.00
23	9439567 vrstevník	0.63 KEYWORDS	Podatáňa animace, napjatost spojená s hereckou úvodnosti kdy spafili maso...	cs	positive	0.20	0.20	0.00
24	9439574 poklad	1.00 KEYWORDS	Konečne som narazil na slušný horor. Záver mohol byť lepší ale to je asi jedi...	cs	positive	0.30	0.30	0.00
25	9439596 film	1.17 KEYWORDS	Ano, čekal jsem od toho odraďování trocha lásky která chybí všem a také tr...	cs	positive	0.30	0.30	0.00
26	9439596 Cohen Kingley	1.00 ENT	Lide v podobě o akční film, který je historii inspirovaná opravdu jen vzdáleně ...	cs	ambivalent	-0.10	0.10	-0.30
27	9439596 George Melis	1.00 ENT	Veľmi tragický príbeh z konca roku 1944, netí jednoduchú hodnotu film ktorý...	cs	negative	-0.30	0.00	-0.30

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 rozpoznáný 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 soustředíme 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 extrahována 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 nespolehat 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 rozpoznáný 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.