Jest to tekst związany z odczytem wygłoszonym na XLIV Szkole Matematyki Poglądowej *Do czego to się przydaje?*, Sulejów, styczeń 2010.

Poszukiwanie struktury w danych na przykładzie analizy korespondencji

Przemysław BIECEK, Warszawa

Osoby pracujące przy analizie danych, czy to statystycy, czy informatycy, czy ekonomiści, czy lekarze, wszyscy oni mają do dyspozycji bardzo szeroki wachlarz technik i algorytmów. Część z tych technik służy do weryfikacji lub potwierdzenia przypuszczeń, tzw. hipotez badawczych. W medycynie czy gospodarce często mamy do czynienia z sytuacją, gdy badacz chce wykazać, że jego odżywka powoduje szybszy wzrost roślin, że opracowany lek powoduje poprawę jakości życia pacjentów, że wymyślony algorytm powoduje istotne przyspieszenie działania programu. W wymienionych przypadkach celem badacza jest weryfikacja pewnego przypuszczenia, ocena na ile dane uzyskane podczas przeprowadzania eksperymentów to przypuszczenie wspierają. Grupa metod statystycznych weryfikujących przypuszczenia badacza nazywana jest często procedurami weryfikacji hipotez statystycznych lub procedurami testowania statystycznego. Poza analizami weryfikującymi hipotezy badawcze, analitycy danych mają też całkiem potężny zestaw technik służących do analizy eksploracyjnej. Jest wiele sytuacji, w których do dyspozycji mamy olbrzymi zbiór danych opisujący wiele zmiennych i jesteśmy ciekawi, jakie zależności występują między zmiennymi w tym zbiorze danych. Możemy chcieć odkryć nowe zależności, które być może później będziemy weryfikować z użyciem testów statystycznych, ale często interesuje nas wyłącznie opis struktury zależności i/lub może ich graficzne przedstawienie. Tego typu analizy nazywa się często analizami eksploracyjnymi. Badacz, nie wiedząc, czego się w danych może spodziewać, odkrywa zależności, które w danych występują. Do takich eksploracyjnych badań świetnie nadają się różnorodne techniki z pogranicza statystyki i dziedziny nazywanej przez niektórych *data-mining* (jest wiele polskich tłumaczeń, w tym „kopanie danych", pozwolę sobie jednak używać angielskiej nazwy), przez innych *machine-learning* (tłumaczonej najczęściej jako „uczenie maszynowe", choć jest to raczej skrót od „uczenie się, gdy do dyspozycji mamy taką maszynę jak np. komputer, zdolną do bardzo szybkich obliczeń"), przez innych „statystyką stosowaną" (bo wszystko, co dotyczy analizy danych, to przecież statystyka).

Poniższy tekst poświęcę opisowi pewnej, bardzo użytecznej, techniki eksploracyjnej analizy danych. Zarysuję podstawy algebraiczne tej metody, czyli dekompozycję na wartości osobliwe oraz przedstawię przykład użycia w analizie tekstu ze stenogramów z polskiego Sejmu i Senatu. Technika, którą poniżej przedstawię nazywana jest analizą korespondencji.

Dekompozycja na wartości osobliwe (ang. Singular Value Decomposition, SVD) i analiza korespondencji

Niech X oznacza macierz o wymiarach $n \times m$. Macierz ta może być przedstawiona jako iloczyn trzech macierzy

$$X = UDV',$$

gdzie prim oznacza transpozycję macierzy, U to macierz o wymiarach $n \times n$, której kolumny stanowią wektory własne macierzy XX' (jest to więc baza ortonormalna w przestrzeni rozpiętej przez kolumny macierzy X), V to macierz o wymiarach $m \times m$, której kolumny stanowią wektory własne macierzy $X'X$ (jest więc to baza ortonormalna w przestrzeni wierszy macierzy X), D jest macierzą o wymiarach $n \times m$, diagonalną o nierosnących wartościach na przekątnej. Zauważmy, że macierz X jest rozłożona na $\min(n, m)$ iloczynów kolumn macierzy U i V zsumowanych z wagami określonymi przez przekątną macierzy D. Przypuśćmy, że chcemy przedstawić macierz X w postaci graficznej, ale liczba stopni swobody tej macierzy jest tak duża, że nie uda nam się wszystkich informacji przedstawić na dwuwymiarowym wykresie. Co zrobimy? Przedstawimy wiersze i kolumny macierzy X odpowiednio w bazach opisanych

przez pierwsze dwie kolumny macierzy U i V. W ten sposób na wspólnych osiach przedstawimy profile wszystkich wierszy i wszystkich kolumn macierzy X. Umożliwi nam to czytelną reprezentację graficzną profili wierszy i kolumn przy możliwie najmniejszej utracie informacji. Metoda dekompozycji SVD, jako sposób wizualizacji macierzy liczb, jest wykorzystywana w analizie składowych głównych (ang. Principal Component Analysis, PCA), w analizie korespondencji (ang. Correspondence Analysis) i np. w analizie kanonicznej (ang. Canonical Analysis). Dekompozycja SVD jest niezwykle użyteczna w wizualizacji struktury danych. Poniżej przedstawimy przykład zastosowania analizy korespondencji. Osoby głębiej zainteresowane tą metodą mogą znaleźć więcej szczegółów i bogatsze referencje np. w pozycji [1].

Analiza danych testowych – text mining

Przykład zastosowania analizy korespondencji przedstawimy z użyciem danych pochodzących z Korpusu Języka Polskiego IPI PAN (zobacz opis tego zbioru danych w [2]). Korpus to zbiór tekstów, tak tekstów literackich, poezji, prozy z różnych epok, jak i tekstów pochodzących z innych źródeł, np. stenogramów z posiedzeń Sejmu, Senatu czy komisji sejmowych. Aktualna wersja Korpusu IPI PAN zawiera ponad 250 mln segmentów, jest to dosyć masywna baza danych. Każde słowo z przetworzonych utworów literackich jest adnotowane morfosyntaktycznie, co oznacza, że mamy dla każdego słowa informację o rdzeniu słowa oraz formie, w której to słowo zostało użyte, tzw. fleksemie. Na tak bogatym zbiorze danych można wykonać wiele bardzo ciekawych analiz. Poniżej pokażemy przykład zastosowania analizy korespondencji, zanim jednak to zrobimy przedstawmy bliżej podzbiór danych, który będziemy analizować.

W analizach użyjemy podzbioru tekstów pochodzących ze stenogramów z posiedzeń Sejmu i Senatu. Każdemu słowu, które zostało zarejestrowane w stenogramach przypisano jeden fleksem morfosyntaktyczny określający formę i znaczenie danego słowa. Zamiast analizować oryginalne słowa będziemy analizować sekwencje fleksemów. Zobaczmy, jaka jest używalność różnych fleksemów w stenogramach z różnych kadencji Sejmu i Senatu. Informację o używalności przedstawia tabela na następnej stronie. Wartość na przecięciu i-tego wiersza i j-tej kolumny to liczba fleksemów typu i, która pojawiła się w stenogramach z posiedzeń kadencji opisanej przez kolumnę j. W wierszach tej tabeli przedstawiono wszystkie adnotowane leksemy, każdemu słowu przypisaliśmy jednoznacznie jeden fleksem. Pełną listę fleksemów wraz z ich opisem można znaleźć w [3] lub w [4].

Każdy fleksem odpowiada zbiorowi znaczeń. Przykładowo, fleksem ‚winien' określa czasownik w typu winien/powinien. Fleksem ‚subst' określa rzeczownik, fleksem ‚depr' opisuje formę deprecjatywną, fleksem ‚num' oznacza liczebnik, fleksem ‚bedzie' oznacza przyszłą formę czasownika ‚być' i tak dalej.

Zebrawszy informację o używalności fleksemów możemy zadać pytanie, czy używalność ta jest taka sama w różnych kadencjach Sejmu i Senatu, czy też są istotne różnice między Sejmem a Senatem, a może używalność fleksemów ewoluuje z czasem?

Jak analizować zależność tego typu? Możemy wykorzystać test niezależności, np. test chi kwadrat do badania istnienia jakiejkolwiek zależności między czynnikami opisanymi w kolumnach a czynnikami opisanymi w wierszach. Nawet jednak jeżeli wynik tego testu będzie wyraźnie wskazywał na brak niezależności (a dla tak dużej liczby słów z pewnością tak będzie), to wciąż nie wiemy jakie zależności występują w tej tabeli, wiemy tylko, że jakieś są. Aby przedstawić czytelnie strukturę zależności w tabeli 1 użyjemy analizy korespondencji. Ponieważ liczba słów w różnych kadencjach była różna, jak również ponieważ używalność fleksemów jest różna, analizie korrespondencji poddawana jest unormowana macierz liczebności. Dokładniej rzecz biorąc dekompozycja SVD zostanie zastosowana do tzw. unormowanych reszt Pearsonowskich, czyli do

	Sejm I	Sejm II	Sejm III	Sejm IV	Senat II	Senat III	Senat IV	Senat V
adj	1024585	2860670	3105710	2418792	168513	533630	1006341	514808
adja	6343	11840	9709	9736	557	1834	2634	1475
adjp	3629	7926	8694	6550	595	2089	3816	2178
adv	157809	440723	480886	365356	28312	93336	162407	86916
aglt	34185	92739	107769	76358	6980	22453	37352	18742
bedzie	29057	79289	94534	82294	4855	16677	30702	15630
conj	266817	748578	843214	636675	47561	153433	285330	141489
depr	15025	39543	42896	34505	2105	6468	10929	5987
fin	430583	1159721	1329316	1042782	84210	264547	489205	244348
ger	25693	73680	74717	59390	4757	19149	36195	14624
ign	161108	403124	491159	395459	53843	150962	266662	155049
imps	12604	34062	36535	27918	1856	5582	9637	4902
impt	25546	69211	78764	59248	4554	14709	24513	10895
inf	125174	338322	372703	288808	22824	65253	109382	56425
interp	114	363	426	307	16682	54769	37473	41271
num	28279	73630	82055	59864	5346	17146	42238	22308
pact	50779	157181	167006	133150	8645	25637	52477	28757
pant	321	874	1035	797	43	185	332	185
pcon	15249	41843	44303	34955	2353	6588	11911	6148
ppas	102684	305969	324264	256864	17487	59248	116860	61925
praet	170363	487832	562794	423838	32460	107166	204104	102894
pred	23772	60856	66753	49170	3940	11319	19492	9384
prep	103128	277207	300368	235394	17629	56216	106826	53229
qub	400309	1049368	1166979	903815	73178	230700	397547	197916
siebie	6399	15059	16918	13645	1185	3873	5921	2874
subst	2370523	6711350	7419553	5857298	392646	1242956	2341349	1236893
winien	11847	30263	29070	19225	2031	5646	8347	3897

Tabela. Liczba wystąpień określonego fleksemu w stenogramach z różnych kadencji Sejmu i Senatu. W wierszach przedstawiono używalności wszystkich adnotowanych fleksemów.

unormowanych różnic pomiędzy obserwowaną liczebnością, a liczebnością oczekiwaną przy pełnej niezależności wierszy i kolumn (ponownie, po bardziej formalny opis analizy korespondencji odsyłam do pozycji 1).

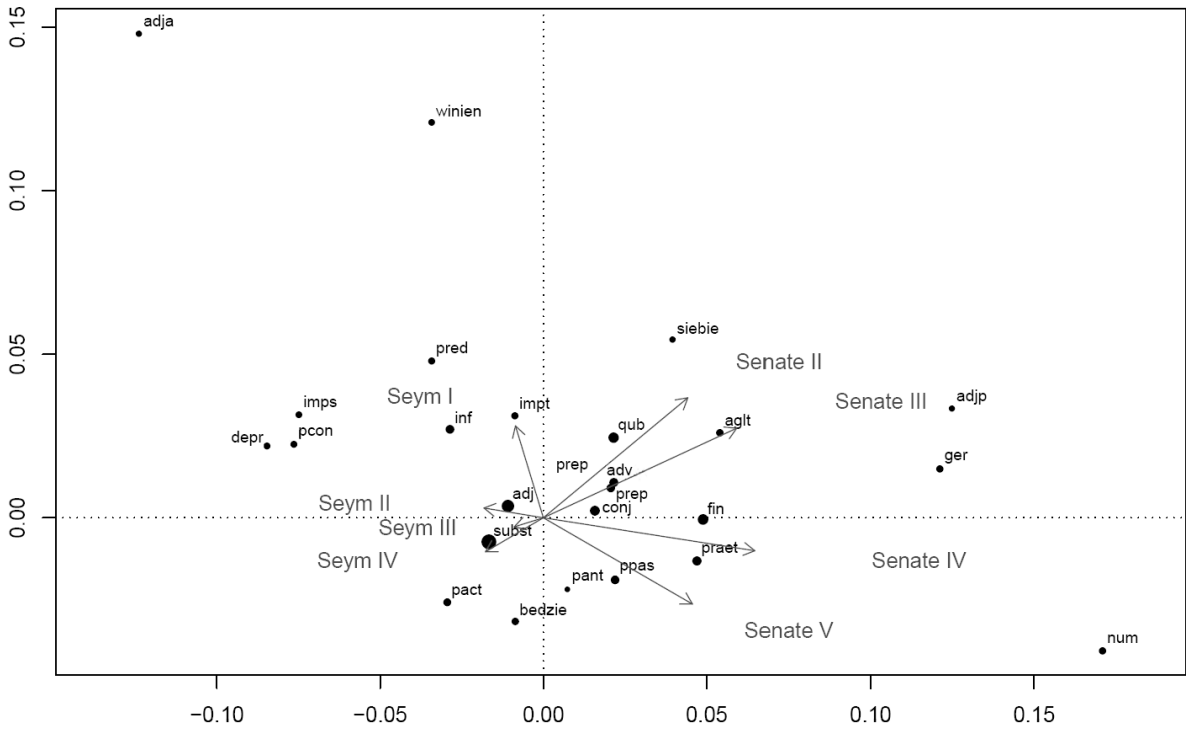
Używając dekompozycji SVD, możemy na wspólnym układzie współrzędnych zaprezentować profile wierszy i profile kolumn macierzy liczebności. Na dwuwymiarowym rysunku możemy zaznaczyć dwie współrzędne, oznacza to, że w bazie wyznaczonej w dekompozycji SVD wykorzystujemy tylko pierwsze dwie kolumny.

Zarówno profile wierszy, a więc profile używalności fleksemów w różnych kadencjach, jak i profile kolumn, przedstawione są na rysunku 1 we wspólnym układzie współrzędnych. Strzałki kończą się w punktach które reprezentują profile kolumn. Kropki odpowiadają profilom używalności fleksemów. Wielkość czarnej kropki odpowiada całkowitej liczbie wystąpień danego fleksemu, im większa kropka tym więcej wystąpień danego fleksemu (np. ‚subst').

Jak czytać tego rodzaju wizualizację tabeli 1? Otóż możemy odczytać z niej trzy rodzaje zależności.

Pierwsza grupa zależności to podobieństwa między profilami używalności różnych fleksemów. Jeżeli profile wierszowe dwóch fleksemów są podobne, to ich współrzędne w bazie pierwszych dwóch, najbardziej znaczących kolumn, również powinny być podobne, a tym samym punkty odpowiadające podobnym fleksemom powinny leżeć blisko na mapie przedstawionej na rysunku 1.

Druga grupa zależności to podobieństwa między profilami kolumn, a więc używalności fleksemów w różnych kadencjach Sejmu i Senatu. Używając tej samej argumentacji, co w przypadku profili wierszy, stwierdzimy, że kolumny



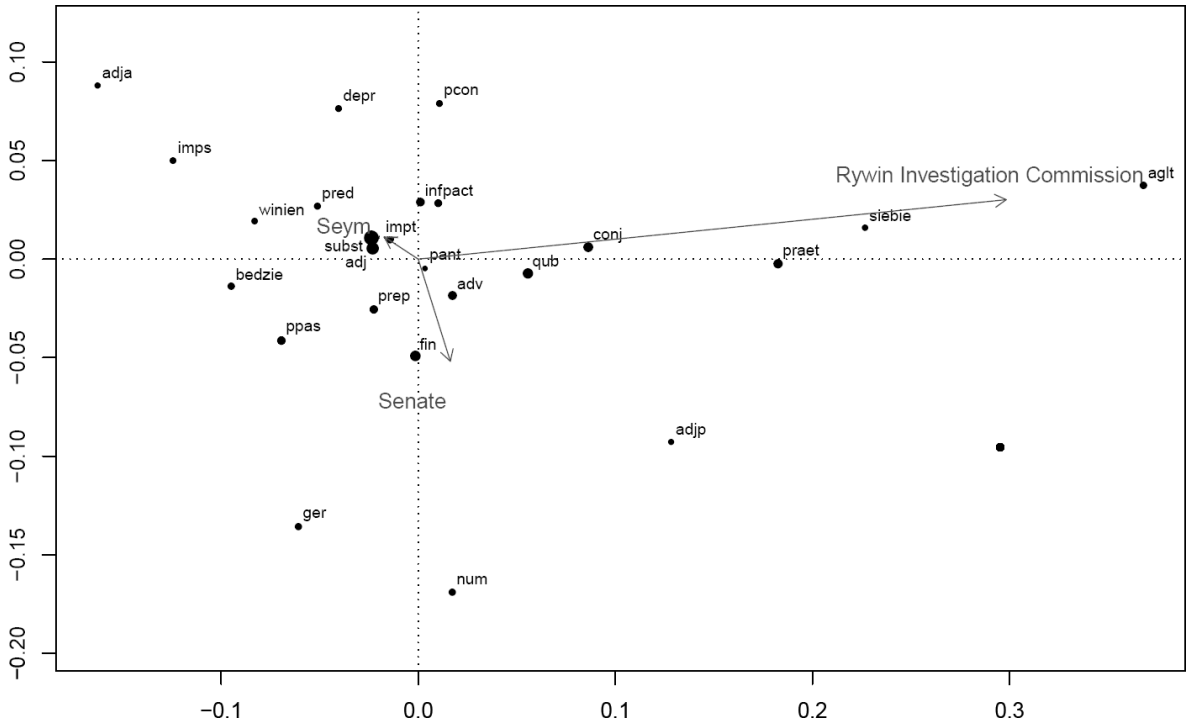
Rys. 1. Reprezentacja graficzna analizy korespondencji dla macierzy liczebności użyć fleksemów w stenogramach z kolejnych kadencji polskiego Sejmu i Senatu.

o podobnych profilach mają podobne współrzędne w bazie pierwszych dwóch składowych. Jeżeli końce wektorów odpowiadających dwóm kolumnom sąsiadują, to profile używalności fleksemów w obu źródłach danych też są podobne.

Trzecia grupa zależności to zależność między fleksemami a źródłami danych. Im zwrot strzałki dla kadencji jest bliższy zwrotowi do punktu opisującego fleksem, tym częściej dany fleksem występuje w danym źródle danych. W ten sposób możemy wskazywać fleksemy, które znacząco częściej występują w danym źródle danych niż w pozostałych źródłach. Podobnie fleksemy niedoreprezentowane w stenogramach z kadencji i to te, dla których zwrot kadencji i jest przeciwny do zwrotu fleksemu.

Wnioski z inspekcji graficznej analizy korespondencji

Biorąc pod uwagę wymienione trzy grupy zależności, które są przedstawiane przez analizę korespondencji, zastanówmy się, jakie ciekawe informacje możemy odczytać z rysunku 1. Pierwsza obserwacja będzie dotyczyła profili używalności fleksemów w Sejmie i Senacie. Widzimy, że pierwsza współrzędna (pierwsza jest najbardziej różnicująca) bardzo dobrze rozdziela profile sejmowe od senackich. Oznacza to, że używalność fleksemów w Sejmie różni się od używalności fleksemów w Senacie. Jak sprawdzić, które fleksemy najbardziej różnicują te dwa źródła profili? Jeżeli różnicuje je pierwsza współrzędna, to zobaczmy, które fleksemy mają duży udział w pierwszej współrzędnej. Takich fleksemów jest kilka, skupimy się na skomentowaniu dwóch, mianowicie fleksemu ‚depr', który ma dużą ujemną wartość pierwszej współrzędnej i fleksemu ‚num', który ma dużą wartość dodatnią na pierwszej współrzędnej. Przypomnijmy, że ‚depr' odpowiada za rzeczownik deprecjatywny, a więc rzeczownik o dużym negatywnym ładunku emocjonalnym. Fleksem ten ma zwrot zbliżony do zwrotu kadencji sejmowych, oznacza to, że najprawdopodobniej jego używalność w stenogramach sejmowych jest wyższa niż w stenogramach senackich. Jak widać debaty w sejmie są bogatsze w słowa obraźliwe. Fleksem ‚num' określa liczebniki. Na pierwszej współrzędnej ma on bardzo wysoką wartość dodatnią, podobnie jak profile kolumnowe odpowiadające stenogramom z Senatu. Odczytać to możemy w ten sposób, że liczebniki znacząco częściej pojawiały się w stenogramach senackich (oczywiście chodzi o względną częstość, unormowaną do liczebności słów w poszczególnych



Rys. 2. Reprezentacja graficzna analizy korespondencji dla macierzy liczebności użyć fleksemów w stenogramach z Sejmu, Senatu i Komisji śledczej do sprawy Rywina.

stenogramach). Czy oznacza to, że w Senacie częściej padają liczby i fakty, a w Sejmie dyskusja sprowadza się do obrażania kogoś lub czegoś? Nie chcemy niczego sugerować, przejdźmy więc do analizy osi pionowej tego wykresu.

Wzdłuż osi pionowej obserwujemy bardzo wyraźny efekt czasowy. Kolejne kadencje tak Sejmu jak i Senatu mają coraz to mniejsze wartości na współrzędnej pionowej. Widzimy, że używalność fleksemów ewoluuje z czasem, zobaczmy, w jaki sposób. Fleksemy, które mają skrajne wartości na drugiej współrzędnej a jednocześnie wartości bliskie zera na pierwszej współrzędnej, a więc te które w podobny sposób ewoluują i w Sejmie i w Senacie to fleksemy: ‚winien' o dużej współrzędnej dodatniej i ‚będzie' o dużej wartości ujemnej. Przypomnijmy, że fleksem ‚winien' odpowiada formom: ‚winien, powinien' a fleksem ‚będzie' odpowiada formom przyszłym czasownika ‚być'. W pierwszych kadencjach tak Sejmu jak i Senatu występowały częściej formy nakazujące, a w ostatnich kadencjach zaczynają dominować formy przyszłe czasownika być. Czy można z tego wyciągnąć wniosek, że starsze kadencje częściej nawoływały do zmian, mówiły jak być powinno, a z czasem trend zmienia się na życzeniowe opisy tego co będzie? Ponownie, wnioski z analiz zostawmy Czytelnikowi.

W zbiorze tekstów Korpusu IPI PAN, poza wieloma innymi źródłami tekstów znajdują się również stenogramy z posiedzeń komisji śledczej w sprawie afery Rywina. Prace tej komisji były emitowane w telewizji i cieszyły się relatywnie dużą oglądalnością. Dla części polityków była to możliwość dotarcia do szerszej publiczności, z pewnością więc podczas prac w tej komisji staranniej dobierali słowa. Również charakter prac komisji śledczej wymusza używanie innych zwrotów niż w przypadku prac w Sejmie i Senacie. Porównajmy profile używalności fleksemów z prac tej komisji z profilami używalności w Sejmie i Senacie. Ponownie wykonamy analizę korespondencji, której wyniki przedstawimy na rysunku 2.

Pierwsza obserwacja to, że używalność fleksemów w stenogramach z Komisji Rywina jest zupełnie inna niż tych, które obserwujemy w stenogramach z Sejmu i Senatu. Wciąż można zauważyć, że fleksemy ‚depr' i ‚num' są charakterystyczne dla odpowiednio Sejmu i Senatu, ale najbardziej w oczy rzuca się zupełnie inny profil używalności fleksemów w posiedzeniach komisji śledczej. Jakie fleksemy są znacznie częściej używane? Fleksemy ‚aglt' (aglutynant BYĆ), ‚siebie' (zaimek

SIEBIE) oraz ‚praet' (pseudoimiesłów) występowały podczas obrad komisji zdecydowanie częściej.

Podsumowanie

Jak widzimy analiza eksploracyjna pozwoliła nam spojrzeć na bazę danych, w której znajdowały się informację o ponad stu milionach słów, poprzez macierz liczebności zawierającej 189 liczb, znacząco mniejszą, ale wciąż trudną do ogarnięcia. Następnie tę macierz liczb przedstawiliśmy graficznie na jednym wykresie prezentującym zależności występujące pomiędzy czynnikami opisanymi przez wiersze, a czynnikami opisanymi przez kolumny macierzy współwystępowań. Dzięki analizie korespondencji udało nam się wychwycić zależności w danych, informacje o strukturze w danych, których nie sposób zauważyć patrząc na surowe wyniki liczbowe.

Analiza korespondencji to tylko jedna technika „kopania w danych". Osoby zainteresowane innymi technikami z pewnością znajdą wiele informacji w książce [5] lub w wiecznie niedokończonym dokumencie 6. Jest też wiele pozycji anglojęzycznych opisujących rozmaite techniki eksploracji danych, wystarczy poszukać takich haseł jak ‚analiza skupień' (ang. cluster analysis), ‚analiza kanoniczna' (ang. canonical analysis), ‚redukcja wymiaru' lub ‚konstrukcja nowych cech' (ang. dimensional scaling, feature extraction), wizualizacja danych (ang. data visualisation). Ostatnie z tych haseł obejmuje dziesiątki technik graficznej prezentacji danych, w taki sposób „by coś było widać". Wybór dobrej techniki zależy od charakteru analizowanych danych i liczby zmiennych oraz liczby obserwacji. Inaczej analizować będziemy dane ilościowe inaczej jakościowe, inaczej pomiary powtarzane w czasie, a inaczej dane o strukturze hierarchicznej. Omawianie tych technik zdecydowanie wykracza poza zakres tego krótkiego tekstu, warto jednak nadmienić, że w czasach, gdy zbiory danych rosną w niesamowicie szybkim tempie, analizowane są dziesiątki tysięcy cech, a zbiory danych zajmują już tera a nieraz i petabajty danych, techniki eksploracji danych zyskują na popularności i są coraz bardziej użyteczne.

Literatura

[1] *Correspondence Analysis in R, with Two- and Three-dimensional Graphics: The ca Package*, `http://www.jstatsoft.org/v20/a03/paper`.

[2] Korpus Języka Polskiego IPIPAN, www.korpus.pl.

[3] Projekt anotacji morfosyntaktycznej korpusu jezyka polskiego, `http://nlp.ipipan.waw.pl/ adamp/Papers/2001-tagset/ipi938.pdf`.

[4] System znaczników morfosyntaktycznych w korpusie IPI PAN, `http://nlp.ipipan.waw.pl/CORPUS/znakowanie.pdf`.

[5] *Statystyczne systemy uczące się*,. Jacek Koronacki, Jan Ćwik.

[6] *Na przełaj przez Data Mining*, Przemysław Biecek, `http://www.biecek.pl/R/naPrzelajPrzezDM.pdf`.