

SAS Text Analytics findet Zusammenhänge in Texten – Ergebnisse eines Selbstversuchs

Gerhard Svolba
SAS Austria
Mariahilfer Straße 116
A-1070 Wien
Sastools.by.gerhard@gmx.net

Zusammenfassung

SAS Text Analytics ist ein mächtiges Werkzeug für die Klassifikation von Dokumenten und die Extraktion von Wissen aus großen Dokumentensammlungen. Die Schritte Datenintegration, Text Parsing, Themenerkennung und Profiling der Themen z.B. mit Word Clouds sowie das automatische Clustering und die Kategorisierung von Texten sind in diesem Werkzeug abgedeckt. Um die Mächtigkeit von SAS Text Analytics zu untersuchen wurden 59 Kapitel von zwei SAS Press Büchern automatisch klassifiziert. Die gefundenen Gruppen wurden inhaltlich auf Sinnhaftigkeit untersucht. Es zeigt sich, dass SAS Text Analytics in der Lage ist automatisch und ohne jegliche Einwirkung von außen die Texte in sinnvolle Gruppen zusammenzufassen.

Schlüsselwörter: Text Mining, Text Analytics, SAS Contextual Analysis, SAS Press, Clustering von Dokumenten

1 Einleitung

Diese Text Mining Analyse wurde als sogenannter „Selbstversuch“ durchgeführt. Zwei vom Autor dieses Artikels verfasste Bücher wurden SAS Contextual Analysis nach Kapitel getrennt übergeben und ohne weitere manuelle Interaktion automatisch verarbeitet. Die automatische Zuteilung der Kapitel zu den einzelnen Clustern wurde vom Autor nach fachlichen Überlegungen geprüft und verifiziert.

2 Die Funktionen und Möglichkeiten von SAS Contextual Analysis

Blickt man in die Produktbeschreibung von SAS Contextual Analysis so gilt:

- Man kann damit große Sammlungen von Text-Dokumenten analysieren, Sentiments identifizieren und robuste Modelle zur Kategorisierung und Extraktion von Inhalten erstellen.
- Das Werkzeug erlaubt eine automatische Identifikation von Themen in Dokumentensammlungen und die Definition von Kategorien und der natürlich-sprachlichen Regeln für die Zuweisung zu diesen Kategorien.

- Methodisch liegt dem eine Kombination von automatischer Erkennung, Machine-Learning Methoden, Linguistischer Regeln und Experten-Input zur Entwicklungen eines Kategorisierungs/Extraktions-Modells zugrunde.
- SAS Contextual Analysis erlaubt ein interaktives Testen und visuelle Exploration über ein HTML5-Browser Interface mit Wizards und Context-sensitiver Hilfe.
- Das Werkzeug ist mit der SAS Plattform gut integriert und auch eine mögliche Ergänzung zum SAS Text Miner.
- Die individuelle Darstellung der Ergebnisse kann z.B. mit SAS Visual Analytics oder anderen in SAS Analytik Produkten erfolgen.

3 Der Selbstversuch – Text Analyse zweier Bücher im SAS Press Verlag

Um besser verstehen zu können, was bei der Text Analyse in SAS Contextual Analysis geschieht, wurde eine Dokumentensammlung verwendet, die dem Autor dieses Artikels sehr nahe ist und ihm inhaltlich sehr gut bekannt ist: Die 59 Kapitel der beiden Bücher SAS Press Bücher „Data Preparation for Analytics Using SAS“ und „Data Quality for Analytics Using SAS“.

Die geringe Anzahl von 59 Dokumenten ist zwar kein „Big Data Problem“ und der SAS High Performance Analytics Server bewältigt auch Dokumentensammlungen mit Millionen von Dokumenten. Zweck dieser Analyse ist aber zu sehen, ob SAS Contextual Analysis in den Kapiteln gemeinsame Themen finden kann und welche Kapitel zu einem thematischen Cluster zusammengefasst werden, ohne dass vorab Information beige-steuert wird oder Wissen des Autors in die Kategorisierung einfließt.

4 Die Text-Analytics Verarbeitung von SAS Contextual Analysis

Aus Data Mining Sicht handelt es sich hier um eine klassische „Un-Supervised Analysis“. Dem Analytik-Tool werden die Daten präsentiert, ohne dass es zusätzliche Informationen über eine mögliche Segmentzugehörigkeit oder das nächstfolgende Ereignis gibt. SAS Contextual Analysis läuft dabei automatisch die gesamte Prozesskette der Text-Analyse durch.

4.1 Einlesen der Dokumente

Als erster Schritt wurden die Dokumente als Wordfiles (eine Datei pro Kapitel) direkt aus dem Verzeichnis im Dateisystem gelesen. Die Kapitelnummer und die Zuordnung zu den beiden Büchern sind dabei zwar bekannt, fließt aber nicht direkt in die Analyse sondern erst in das Profiling ein.



















Name ^	Date modified	Type	Size
 AppA_new.docx	4/17/2012 11:00 AM	Microsoft Word Doc...	62 KB
 AppB_new.docx	4/17/2012 11:03 AM	Microsoft Word Doc...	62 KB
 AppC_new.docx	4/17/2012 11:04 AM	Microsoft Word Doc...	136 KB
 AppD_new.docx	4/17/2012 11:07 AM	Microsoft Word Doc...	73 KB
 AppE_new.docx	4/19/2012 3:04 PM	Microsoft Word Doc...	184 KB
 AppendixA.new.doc	1/8/2007 11:17 AM	Microsoft Word 97 -...	284 KB
 AppendixB.new.doc	1/8/2007 11:17 AM	Microsoft Word 97 -...	334 KB
 AppendixC.new.doc	1/8/2007 11:17 AM	Microsoft Word 97 -...	328 KB
 chap1.new.doc	1/8/2007 11:17 AM	Microsoft Word 97 -...	503 KB
 chap1_new.docx	4/13/2012 11:59 AM	Microsoft Word Doc...	141 KB
 chap2.new.doc	1/8/2007 11:17 AM	Microsoft Word 97 -...	294 KB
 chap2_new.docx	4/13/2012 12:01 PM	Microsoft Word Doc...	104 KB
 chap3.new.doc	1/8/2007 11:17 AM	Microsoft Word 97 -...	301 KB
 chap3_new.docx	4/13/2012 1:11 PM	Microsoft Word Doc...	99 KB
 chap4.new.doc	1/8/2007 11:17 AM	Microsoft Word 97 -...	287 KB
 chap4_new.docx	4/18/2012 9:51 AM	Microsoft Word Doc...	87 KB
 chap5.new.doc	1/8/2007 11:17 AM	Microsoft Word 97 -...	488 KB
 chap5_new.docx	4/13/2012 1:21 PM	Microsoft Word Doc...	105 KB

Abbildung 1: Einlesen der Word-Dokumente aus dem Dateisystem

4.2 Text Parsing

Im nächsten Schritt erfolgt ein Text Parsing.

- Dabei werden die Wörter in unterschiedliche Entitäten (Hauptwort, Zeitwort, ...) eingeteilt.
- Eine Synonym-Erkennung wird durchgeführt und Stop-Listen zur Entfernung von redundanten Wörtern, wie „der, die, das, und, von, mit,“ werden berücksichtigt.
- Es erfolgt eine Gewichtung der Wörter (Terme) und eine Identifikation von Termen, die sich zur Gruppierung von Dokumenten eignen.
- Auf Basis dieser Terme und deren Gewichtung erfolgt eine automatische Erkennung der zugrunde liegenden Themen in den Dokumenten

Terms and Synonyms	Number of Documents	Concept
▶ transactional	24	
▶ advantage	24	
▶ overview	24	
▶ standard	24	
▼ analysis subject	24	NOUN_GROUP
□ analysis subject	19	NOUN_GROUP
□ analysis subjects	13	NOUN_GROUP
□ analysis subjects	5	PROP_MISC
□ analysis subject	1	PROP_MISC
□ monthly	24	
▶ place	24	
▶ leave	23	
□ underlying	23	
□ yes	23	
▶ factor	23	
▶ purchase	23	
□ otherwise	23	
▶ simply	23	
□ common	23	

Abbildung 2: Ergebnisse des Text Parsings

4.3 Automatische Themen-Erkennung

SAS Contextual Analysis verwendet nun Machine Learning Algorithmen zur Analyse der Text Inhalte. Dabei werden Begriffe und Begriffskombinationen, die häufig gemeinsam auftreten, zu Themen zusammengefasst.

Diese automatische Erkennung der Themen auf Basis der Term-Weights in den Dokumenten ist ein wichtiger Schritt zur automatischen Extraktion von Zusammenhängen in den Dokumenten. Abbildung 3 zeigt die automatisch erkannten Themen, die wichtigsten beschreibenden Terme und die Anzahl der Dokumente pro Thema.

All Topics (59)	
+shop,+promotion,+label,+productgroup,+pg	3
detection,+outlier,+node,outlier detection,jmp	3
+simulation,+training,+training data,+response,+random	4
+record,correctness,+systematic,+bias,+database	4
+multiple observation,+analysis subject,+entity,+account,+measurement	4
+title,+profile,+var,+missing record,ts_profile_chain	3
+score,historic,+historic snapshot,people,+snapshot	6
mape,+history,+time history,mape,+disturbance	5
f,+transpose,+weight,data,+root	6
+access,+file,+text,+relational,+relational database	5
+boat,+sail,wind,+race,gps	1

Abbildung 3: Automatische Themenerkennung

In der Benutzeroberfläche von SAS Contextual Analysis besteht hier die Möglichkeit auf einzelne Themen zu klicken und in die Dokumentenansicht zu wechseln. Die ersten Zeilen jedes Dokuments werden wie in Abbildung 4 dargestellt, angezeigt. Zusätzlich wurden dieser Anzeige noch auf der linken Seite die Bücher als Farbcode hinzugefügt.

Es zeigt sich, dass die Kapitel die hier zum Thema „ACCESS/FILE/TEXT /RELATIONAL DATABASE“ zusammengefasst wurden, jene sind, die inhaltlich mit Datenstrukturen in den Datenbanken und dem Datenzugriff zu tun haben:

- Kapitel 5 (Origin of Data), 13 (Accessing Data) und 17 (Transformations for Categorical Data), sowie Appendix B im Data Preparation Buch
- Kapitel 3 „Data Availability“ im Data Quality Buch.

Diese Zuteilung macht aus inhaltlicher Sicht Sinn und zeigt, dass von Kontext her ähnliche Dokumente auch als solche gefunden werden.

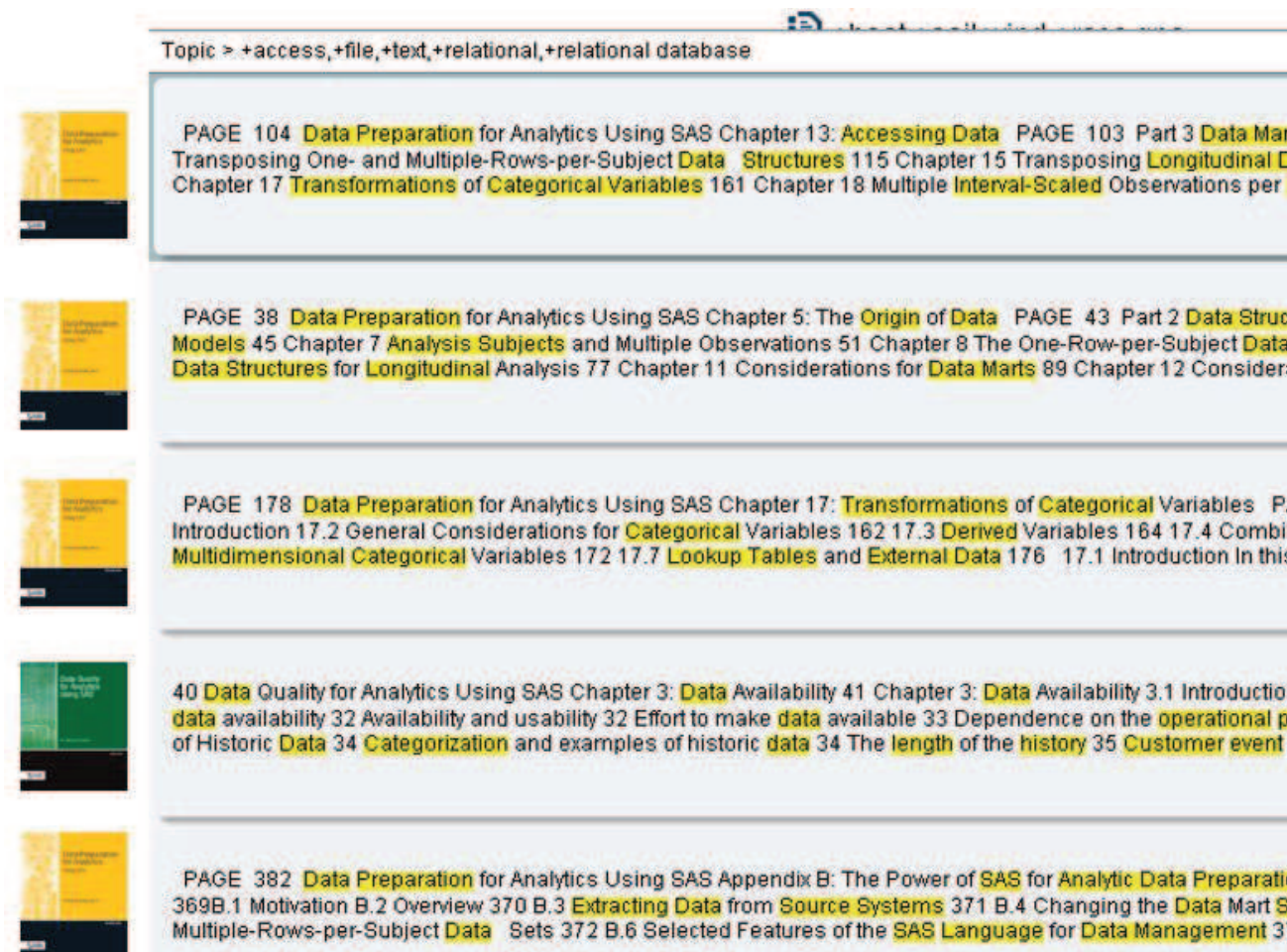


Abbildung 4: Profiling der Themen durch den Dokument-View

Zusätzlich können für das Profiling der Themen und der Thematik auch Word Clouds (s. Abb. 5) sowie Term Maps (s. Abb. 6) angezeigt werden.

5 k-means Clustering der Dokumente im SAS Enterprise Miner

Die obigen Ergebnisse in SAS Contextual Analysis können bereits als gute Evidenz gesehen werden, dass SAS Text Analytics in der Lage ist die richtigen Dokumente automatisch in Gruppen zusammenzufassen.

Zusätzlich wurde noch eine k-means Clusteranalyse auf Basis der Topic-Weights pro Dokument durchgeführt. Diese Analyse zeigt die enge Integration der einzelnen Analyserwerkzeuge in der SAS Analytic Plattform. Zusätzlich wird eine Möglichkeit gezeigt, im Text Mining die Dokumente eindeutig bestimmten Clustern zuzuordnen.

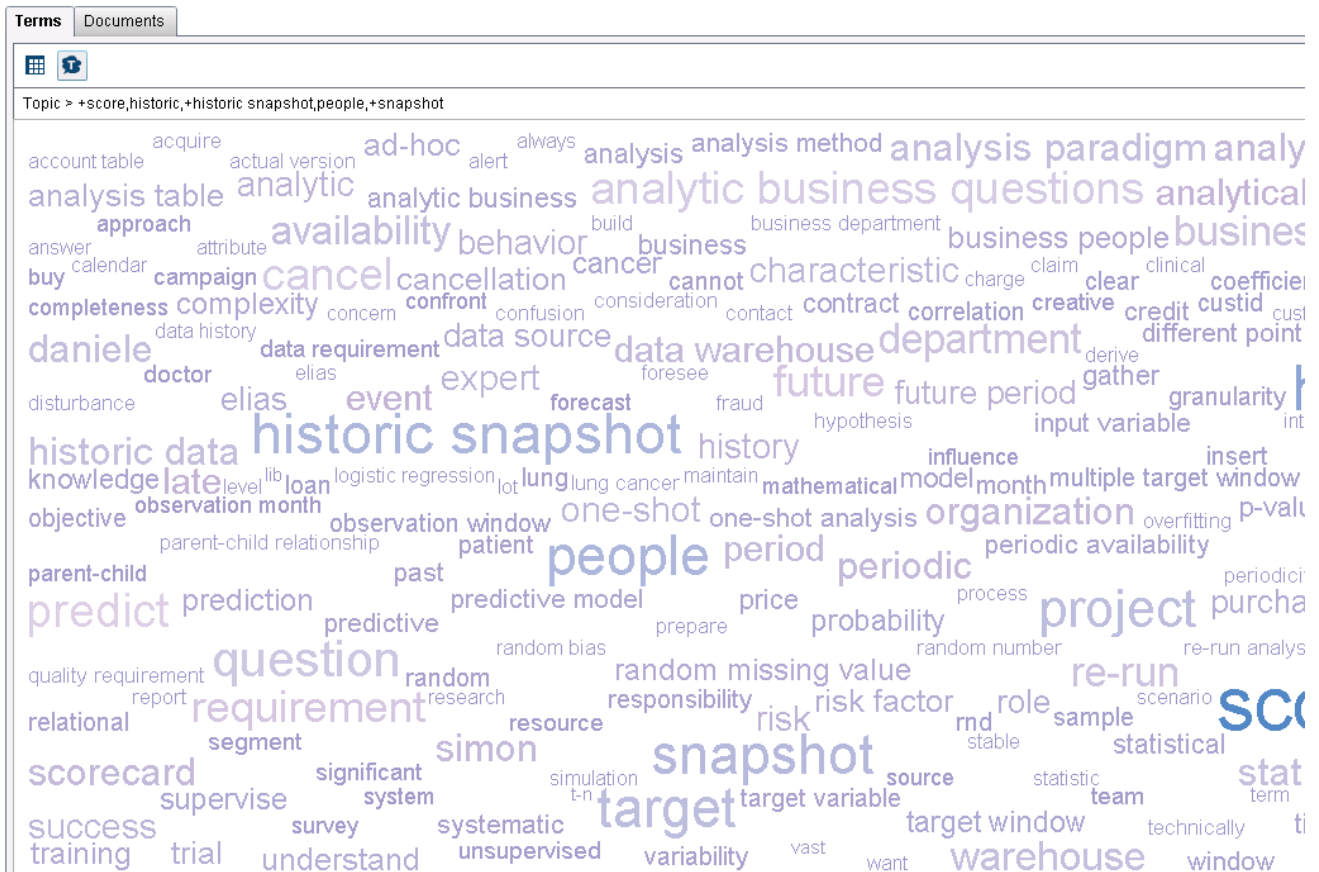


Abbildung 5: Word Cloud für das Thema „Score/Historic/Snapshot“

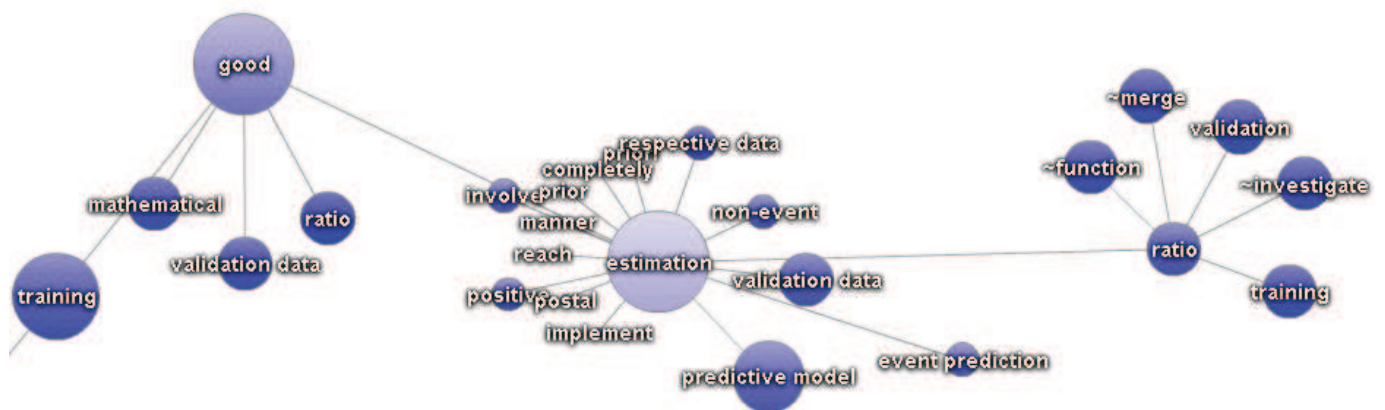


Abbildung 6: Term-Map für „Estimation“

5.1 Verfügbare Daten

Abbildung 7 zeigt die Topic Gewichte als Tabelle mit einer Zeile pro analysiertem Dokument. Diese Tabelle wird von SAS Contextual Analysis automatisch erstellt und steht in der SAS Analytic Plattform für weitere Analysen bereit. In der Abbildung 7 wurden ausgewählte Zellen hervorgehoben, die einen hohen Weight-Wert pro Dokument haben.

	topic_raw1	topic_raw2	topic_raw3	topic_raw4	topic_raw5	topic_raw6	topic_raw7	topic_raw8
1	0.047	-0.007	0.025	-0.002	0.018	0.755	-0.001	0.038
2	0.010	-0.064	0.043	-0.005	0.014	0.065	0.025	0.015
3	0.011	-0.050	0.276	-0.009	0.021	0.049	0.018	0.047
4	0.054	-0.026	0.014	-0.053	0.018	0.097	0.023	0.363
5	0.069	-0.112	0.024	0.000	0.075	0.085	0.047	0.026
6	0.048	-0.039	0.049	0.005	0.086	0.069	0.021	0.034
7	-0.002	-0.043	0.040	-0.032	0.091	0.086	0.003	0.023
8	0.031	-0.016	0.028	0.061	-0.736	0.018	0.053	0.019
9	0.032	-0.026	0.037	0.071	0.098	0.028	0.283	0.033
10	0.165	-0.029	0.009	0.022	0.190	0.032	0.058	0.031
11	0.022	-0.208	0.122	0.061	0.049	0.408	0.050	0.037
12	0.041	-0.053	0.045	0.023	0.155	0.022	0.075	0.008
13	-0.110	-0.090	-0.004	0.117	0.039	0.473	0.023	0.097
14	0.038	-0.023	0.121	0.025	0.055	0.015	-0.158	0.021
15	0.028	-0.035	0.030	0.173	0.134	0.064	-0.028	0.007
16	0.058	-0.024	0.026	0.020	0.103	0.029	0.008	0.015
17	0.007	-0.377	0.092	-0.110	0.075	0.064	0.082	0.072

Abbildung 7: Topic-Gewichte pro Dokument als Datentabelle (nur ausgewählte Spalten, eine Zeile pro Dokument)

5.2 Clusteranalyse

Die Tabelle aus Abbildung 7 wurde im SAS Enterprise Miner direkt weiterverwendet und mit dem Cluster Node ein k-means Clustering auf den Spalten TOPIC_RAW1 – TOPIC_RAW11 durchgeführt:

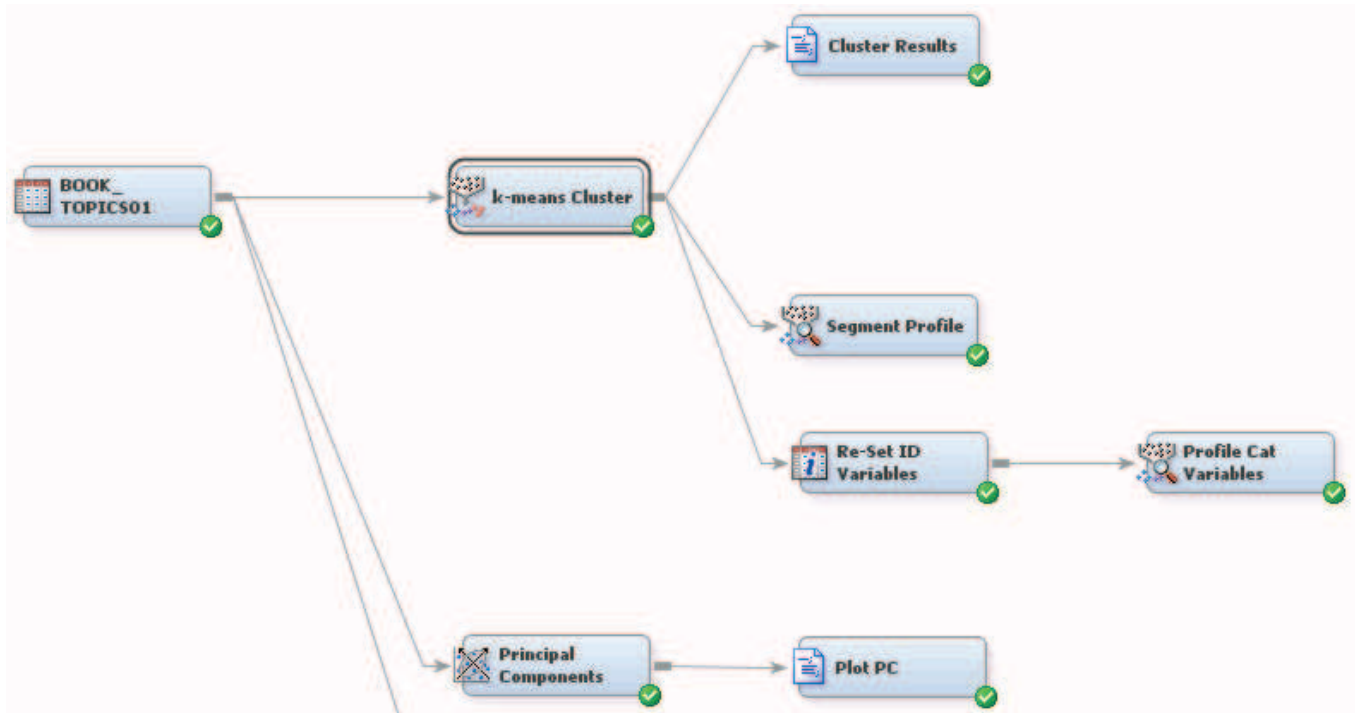


Abbildung 8: Prozessfluss für das Clustering im SAS Enterprise Miner

6 Ergebnisse

6.1 Acht sauber abgegrenzte Dokumenten-Cluster als Ergebnis

Bei der Clusteranalyse wurden 8 Cluster gefunden, die in Abbildung 9 dargestellt sind. Zur besseren Veranschaulichung sind die Kapitel des „Data Quality Buchs“ in grün (dunkel) und die Kapitel des „Data Preparation Buchs“ in gelb (hell) dargestellt.

1	Missing Values	10	11	A													
2	Erzeugen des Analytic-Marts	10	11	14	15	16	18	19	20	21	22	23	24	25	26	27	28
3	Data Origin und Data Management	5	13	17	B												
4	DQ Case Studies	1															
5	Fachliche Konzepte	1	2	3	4	12	2	3	4	5	6	7	8	9	12		
6	DQ mit Analytik und SAS	13	14														
7	Data Quality Simulationen	15	16	17	18	19	20	21	22	23	C	D					
8	Analytic Data Mart Structures	6	7	8	E												

Abbildung 9: Darstellung der Cluster Ergebnisse

Man sieht sehr eindrucksvoll, wie die Kapitel anhand ihrer Inhalte zu unterschiedlichen Clustern zusammengefunden werden. Manche Cluster enthalten nur Kapitel aus einem Buch:

- Im Cluster 1 finden sich all jene Kapitel die im Data Quality Buch das Thema der fehlenden Werte behandeln.
- Oder das Cluster 7, welches die Simulationsstudien umfasst, die in den Kapiteln 15-23 beschrieben sind.

In manchen Clustern sind Kapitel aus beiden Büchern enthalten.

- Cluster 8 enthält die Kapitel zu analytischen Datenstrukturen aus dem „Data Preparation Buch“, der Appendix E im „Data Quality Buch“ ist eine Zusammenfassung aus diesen Kapiteln.

Dies demonstriert eindrucksvoll, dass Inhalte, die quer über unterschiedliche Dokumente als „nahe“ oder „ähnlich“ erkannt werden sollen, tatsächlich auch als solche gefunden werden.

Die unterschiedliche Anzahl der Dokumente pro Cluster zeigt auch, dass hier nicht nach vorgegebenen Schemen vorgegangen wird, sondern dass die Kapitel ausschließlich anhand ihres Inhalts gruppiert werden. Cluster 4 enthält nur ein einzelnes Kapitel. Das Kapitel 1 im Data Quality Buch ist eine Sammlung von Fallbeispielen und vom Inhalt her nicht mit anderen Kapiteln vergleichbar.

6.2 Ergebnisse der Hauptkomponentenanalyse

Sie können die Variablen Spalten TOPIC_RAW1 – TOPIC_RAW11 aus der Term Weights by Document Tabelle auch für eine Hauptkomponentenanalyse verwenden. Abbildung 10 zeigt den Scatterplot für die 1. und 2. Hauptkomponente. Die einzelnen Kapitel sind farblich nach Buch und Buch-Sektion codiert.

- Wieder zeigt sich das isolierte Outlier Kapitel 1 aus dem Data Quality Buch rechts oben.
- Kapitel 5,7,13 und Anhang E aus dem Data Preparation Buch sowie Anhang E aus dem Data Quality Buch sind sich sehr ähnlich. Dies ist inhaltlich stimmig, da Anhang E eine Zusammenfassung der Inhalt dieser Kapitel aus dem Data Preparation Buch ist.

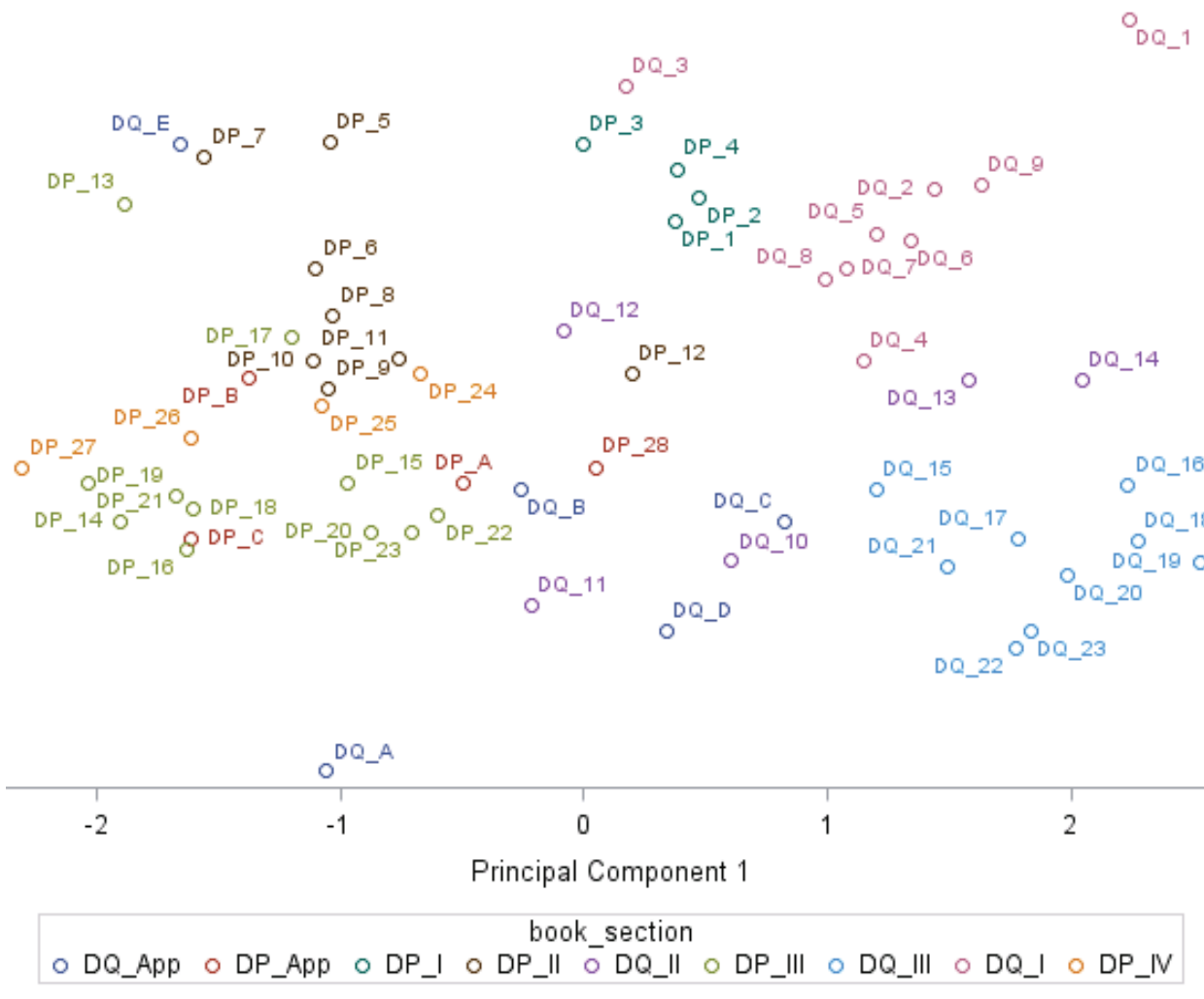


Abbildung 10: Scatterplot für die 1. und 2. Hauptkomponente

7 Schlussfolgerung

Mit Vertrauen gestärkt zu neuen Anwendungsfällen: Diese Ergebnisse bestärken das Vertrauen, dass mit SAS Contextual Analysis Einblick in Dokumentensammlungen gewonnen werden kann ohne das Zusammenhänge vorab definiert werden.

Diese Methoden erlauben es, dass vielfältige Dokumentensammlungen analysiert werden und so Einblick in die Inhalte und Zusammenhänge gewonnen werden können. Sie erfahren dadurch, welche Themen in Ihren Texten enthalten sind und wie Sie diese automatisch in Gruppen einteilen können, ohne jedes Dokument einzeln lesen zu müssen.

Die SAS Analytic Plattform erlaubt die Durchführung dieser Analysen und ermöglicht eine nahtlose Verbindung unterschiedlicher Werkzeuge.

Literatur

- [1] G. Svolba: Data Preparation for Analytics Using SAS: SAS Press 2006, Cary NC
- [2] G. Svolba: Data Quality for Analytics Using SAS: SAS Press 2012, Cary NC
- [3] G. Chakraborty et. al.: Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS®: SAS Press 2013, Cary NC