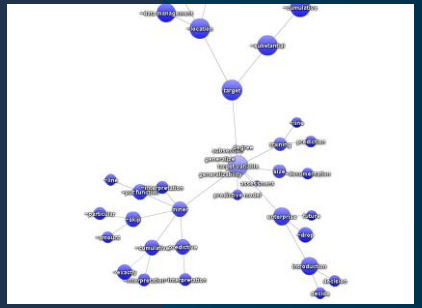
SAS Text Analytics findet Zusammenhänge in Texten – Ergebnisse eines Selbstversuchs



21. KSFE, Krefeld, 9.-10. März 2017

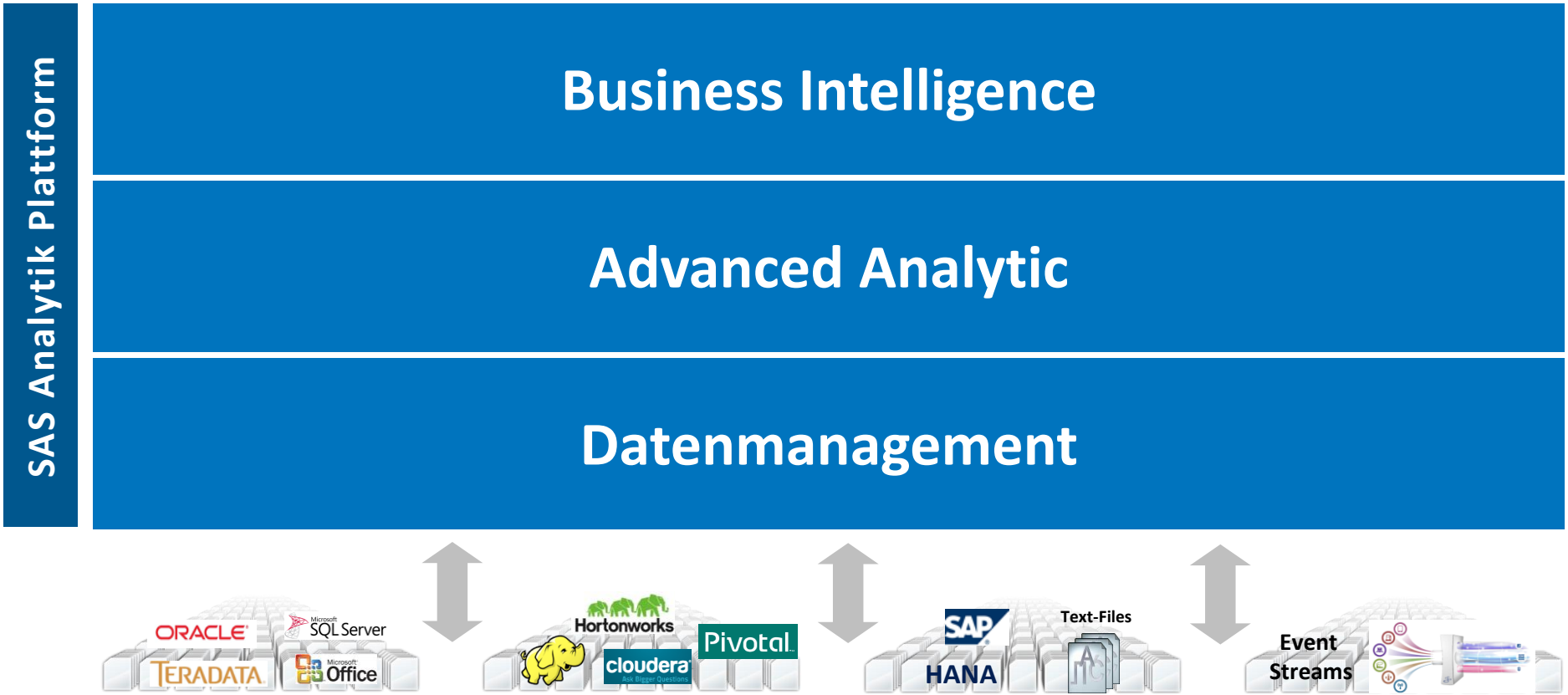
Gerhard Svolba

Die Vortragsfolien sind online → Google: Gerhard SAS Samples

THE POWER TO KNOW.

SAS Analytik Plattform

Unterschiedliche Layer aus konzeptioneller Sicht



SAS Analytik Plattform

Advanced Analytic Layer

SAS Analytik Plattform

Business Intelligence

Data Mining

Statistical Analysis

Forecasting

Text Analytics

Optimization & Simulation

Datenmanagement

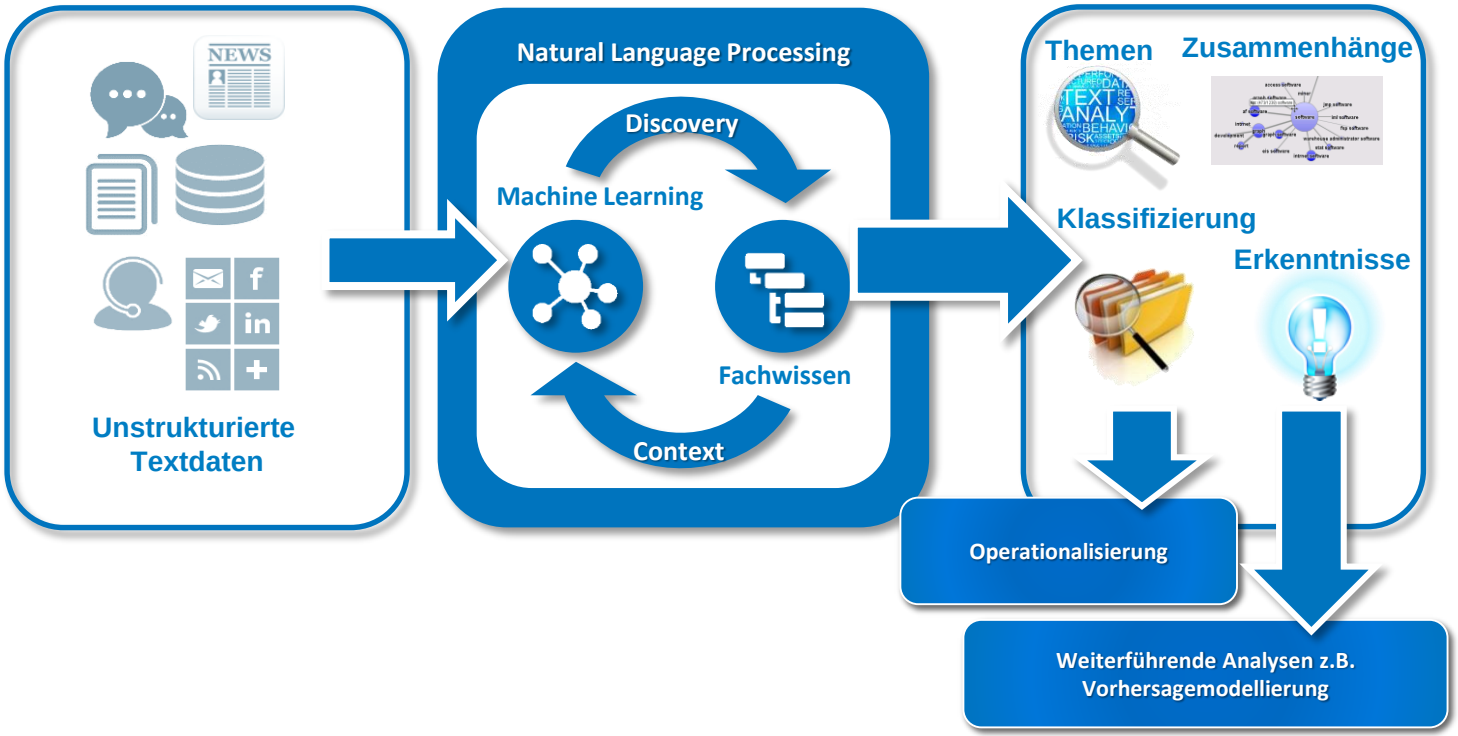
SAS® Contextual Analysis

EIN BLICK IN DIE LÖSUNGSBESCHREIBUNG

- Was macht die Lösung?
  - Es erlaubt große Sammlungen von Text-Dokumenten zu analysieren, Sentiments zu identifizieren und robuste Modelle zur Kategorisierung und Extraktion von Inhalten zu erstellen.
- Wie funktioniert das?
  - Kombination von automatischer Erkennung, Machine-Learning Methoden, Linguistischer Regeln und Experten-Input zur Entwicklungen eines Kategorisierungs/Extraktions-Modells
  - Automatische Identifikation von Themen in den Dokumenten, Definition von Kategorien und Überarbeitungen durch den Text-Analysten
  - Interaktives Testen und visuelle Exploration über ein HTML5-Browser Interface mit Wizards und Context-sensitiver Hilfe.
- Wie integriert sich SAS® Contextual Analysis in das SAS Portfolio?
  - Integrierter Teil der SAS Plattform (SAS Metadata Server, …)
  - (Mögliche) Ergänzung zum SAS Text Miner
  - Ergebnis-Darstellung mit SAS Visual Analytics, Weiterwendung in SAS Analytik Produkten

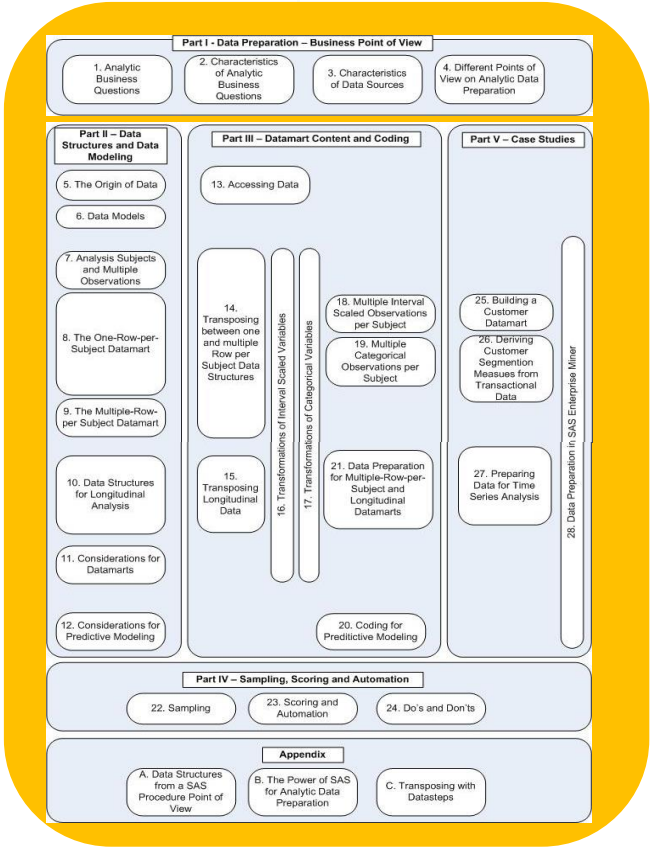
SAS® Contextual Analysis

PROZESSFLUSS UND ÖKO-SYSTEM



Der Selbstversuch

DIE AUSGANGSBASIS: 2 BÜCHER VON SAS-PRESS



Data Quality Defined

Case Studies – Definition- Availability – Quantity – Completeness – Correctness – Predictive Modeling – Analytics – Process Considerations

Profiling and Improvement

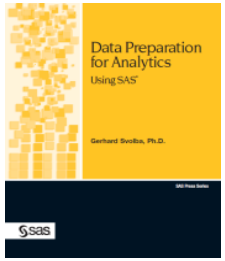
Missing Values – Time Series Data – Across Tables – Data Quality with Analytics – SAS Analytic Tools

Simulation Studies

Introduction – Predictive Modeling – Time Series Forecasting

Demo-Beispiel

THEMENSUCHE IN DOKUMENTEN UND CLUSTERING



28 Kapitel
3 Appendixe

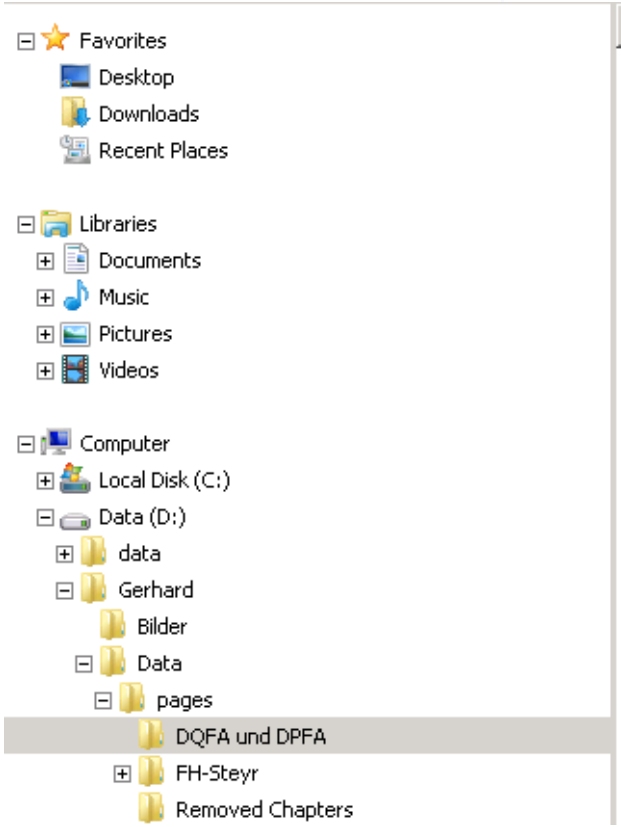
23 Kapitel
5 Appendixe

- 59 Word-Dokumente
- "Un-supervised" Analyse

- Welche Themen werden gefunden?
- In welche Cluster können die Dokumente eingeteilt werden?
- Wodurch beschreiben sich diese Cluster?

Datenintegration

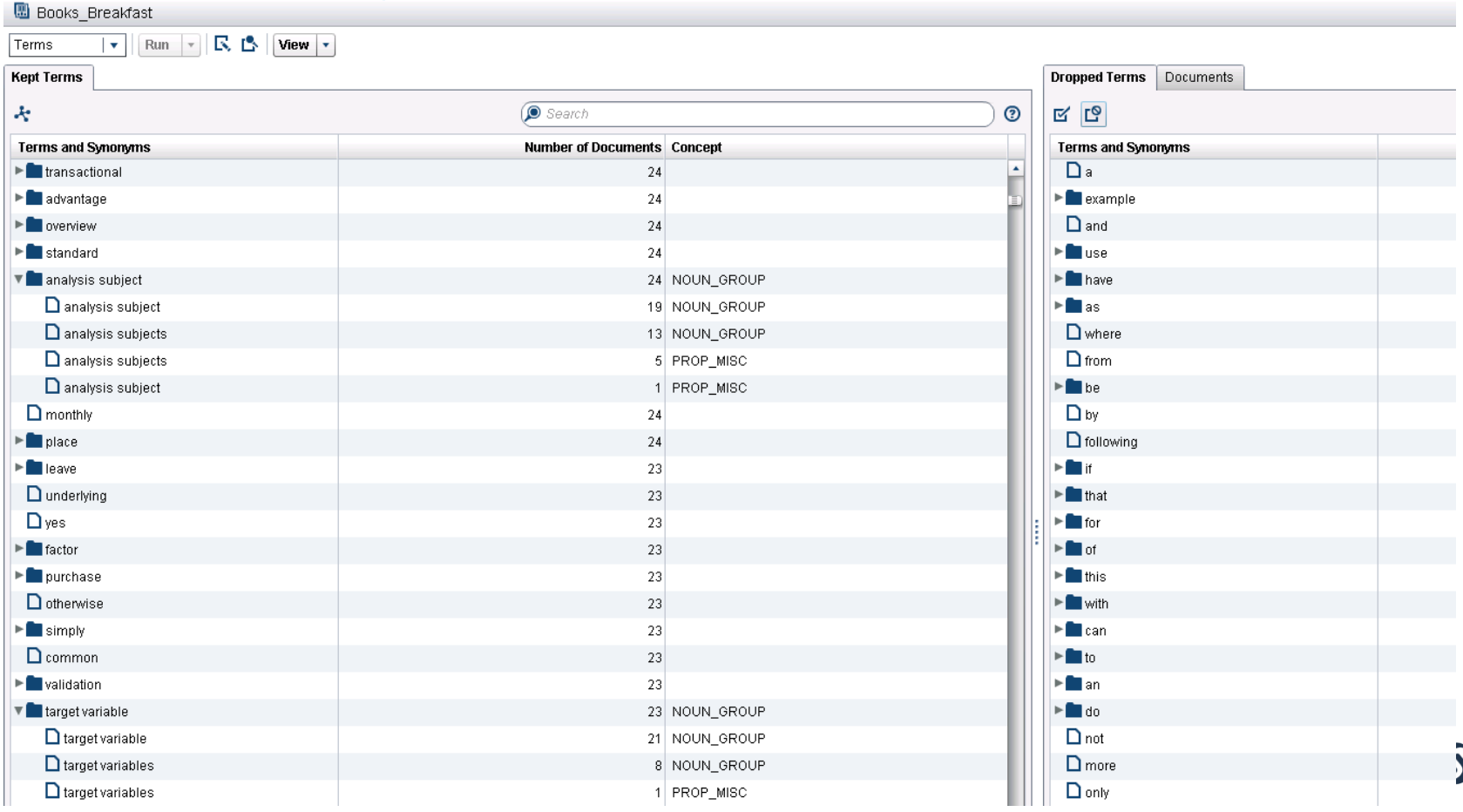
BEREITSTELLUNG DER QUELLDATEN ALS WORD-DOKUMENTE



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

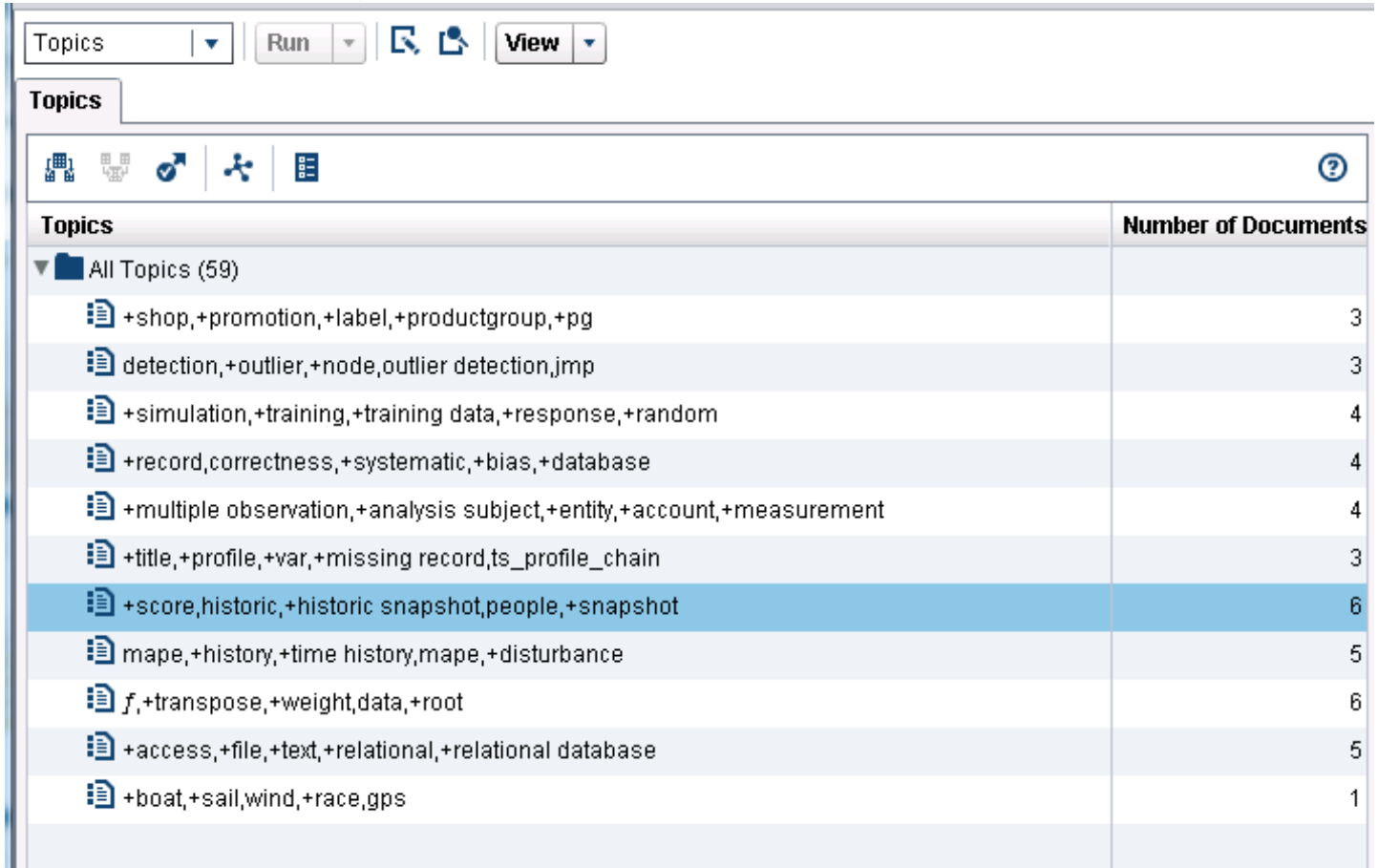
Text Parsing

AUTOMATISCHE SYNONYM-ERKENNUNG, STEMMING, STOP-LISTEN BERÜCKSICHTIGUNG



Erkennen von Themen

AUTOMATISCHE THEMEN-ERKENNUNG AUF BASIS ANALYTISCHER MODELLE



Detailanalyse der Themen

THEMA

detection,+outlier,+node,outlier detection,jmp

+simulation,+training,+training data,+response,+random

+record correctness +systematic +bias +database

Terms | **Documents**

Docume

Topic > +simulation,+training,+training data,+response,+random

254 Data Quality for Analytics Using SAS Chapter 19: Influence of Data Correctness on Model Quality in Predictive Modeling 253 Chapter 19: Influence of Data Correctness on Model Quality in Predictive Modeling 19.1 Introduction 243 General 243 Non-visible data quality problem 244 Random and systematic bias 244 Biased values in the scoring data partition 244 19.2 Simulation Methodology and Data Preparation 245 ...

236 Data Quality for Analytics Using SAS Chapter 18: Influence of Data Completeness on Model Quality in Predictive Modeling 237 Chapter 18: Influence of Data Completeness on Model Quality in Predictive Modeling 18.1 Introduction 231 General 231 Random and systematic missing values 232 Missing values in the scoring data partition 232 18.2 Simulation Methodology and Data Preparation 233 Inserting random missing values 233...

230 Data Quality for Analytics Using SAS Chapter 17: Influence of Data Quantity and Data Availability on Model Quality in Predictive Modeling 229 Chapter 17: Influence of Data Quantity and Data Availability on Model Quality in Predictive Modeling 17.1 Introduction 219 General 219 Data quantity 220 Data availability 220 17.2 Influence of the Number of Observations 220 Detailed functional question 220 Data preparation 220 Simulation ...

216 Data Quality for Analytics Using SAS Chapter 16: Simulating the Consequences of Poor Data Quality for Predictive Modeling 217 Chapter 16: Simulating the Consequences of Poor Data Quality for Predictive Modeling 16.1 Introduction 206 Importance of predictive modeling 206 Scope and generalizability of simulations for predictive modeling 206 Overview of the functional questions of the simulations 206 16.2 Base for the Business ...

Detailanalyse der Themen

THEMA

+access,+file,+text,+relational,+relational database

Topic > +access,+file,+text,+relational,+relational database

PAGE 104 Data Preparation for Analytics Using SAS Chapter 13: Accessing Data PAGE 103 Part 3 Data Mart Coding and Content Chapter 13 Acces Transposing One- and Multiple-Rows-per-Subject Data Structures 115 Chapter 15 Transposing Longitudinal Data 131 Chapter 16 Transformations of Chapter 17 Transformations of Categorical Variables 161 Chapter 18 Multiple Interval-Scaled Observations per Subject 179 Chapter 19 Multiple Catego

PAGE 38 Data Preparation for Analytics Using SAS Chapter 5: The Origin of Data PAGE 43 Part 2 Data Structures and Data Modeling Chapter 5 The Models 45 Chapter 7 Analysis Subjects and Multiple Observations 51 Chapter 8 The One-Row-per-Subject Data Mart 61 Chapter 9 The Multiple-Rows-p Data Structures for Longitudinal Analysis 77 Chapter 11 Considerations for Data Marts 89 Chapter 12 Considerations for Predictive Modeling 95 Introdu

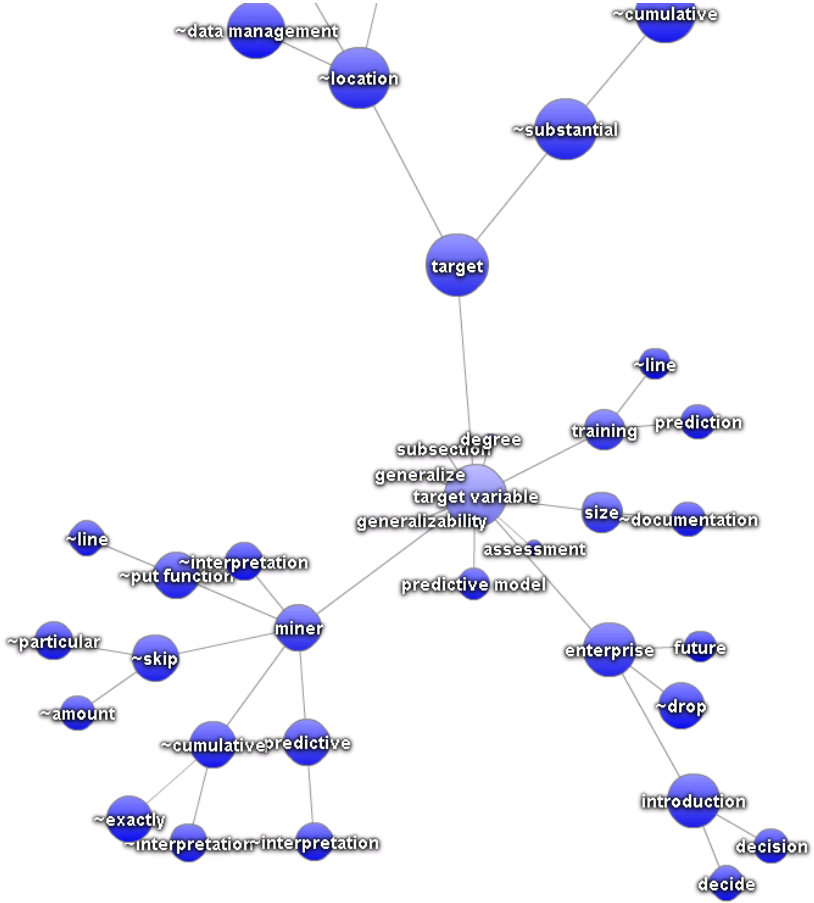
PAGE 178 Data Preparation for Analytics Using SAS Chapter 17: Transformations of Categorical Variables PAGE 177 Chapter 17 Transformations Introduction 17.2 General Considerations for Categorical Variables 162 17.3 Derived Variables 164 17.4 Combining Categories 166 17.5 Dummy Codi Multidimensional Categorical Variables 172 17.7 Lookup Tables and External Data 176 17.1 Introduction In this chapter we will deal with transformatio

40 Data Quality for Analytics Using SAS Chapter 3: Data Availability 41 Chapter 3: Data Availability 3.1 Introduction 32 3.2 General Considerations 32 Re data availability 32 Availability and usability 32 Effort to make data available 33 Dependence on the operational process 33 Availability and alignment in t of Historic Data 34 Categorization and examples of historic data 34 The length of the history 35 Customer event histories 35 Operational systems and a

PAGE 382 Data Preparation for Analytics Using SAS Appendix B: The Power of SAS for Analytic Data Preparation PAGE 381 Appendix B The Power o 369B.1 Motivation B.2 Overview 370 B.3 Extracting Data from Source Systems 371 B.4 Changing the Data Mart Structure: Transposing 371 B.5 Data Mar Multiple-Rows-per-Subject Data Sets 372 B.6 Selected Features of the SAS Language for Data Management 375 B.7 Benefits of the SAS Macro Langu

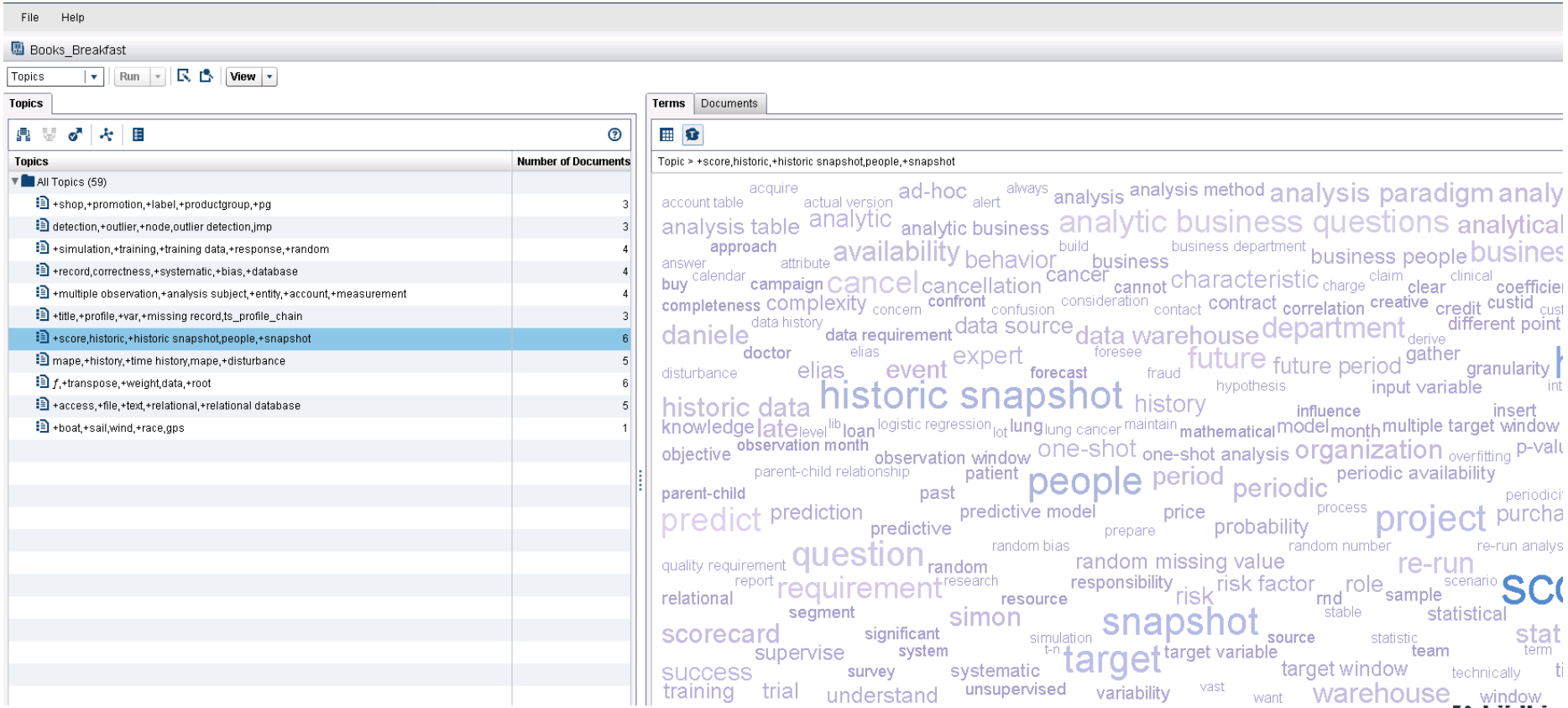
Darstellung von Term Maps

TERM-MAP FÜR DEN BEGRIFF TARGET VARIABLE



Darstellung von Word Clouds

WORD-CLOUD FÜR DIE DOKUMENTE DES THEMAS +score,historic,+historic snapshot,people,+snapshot



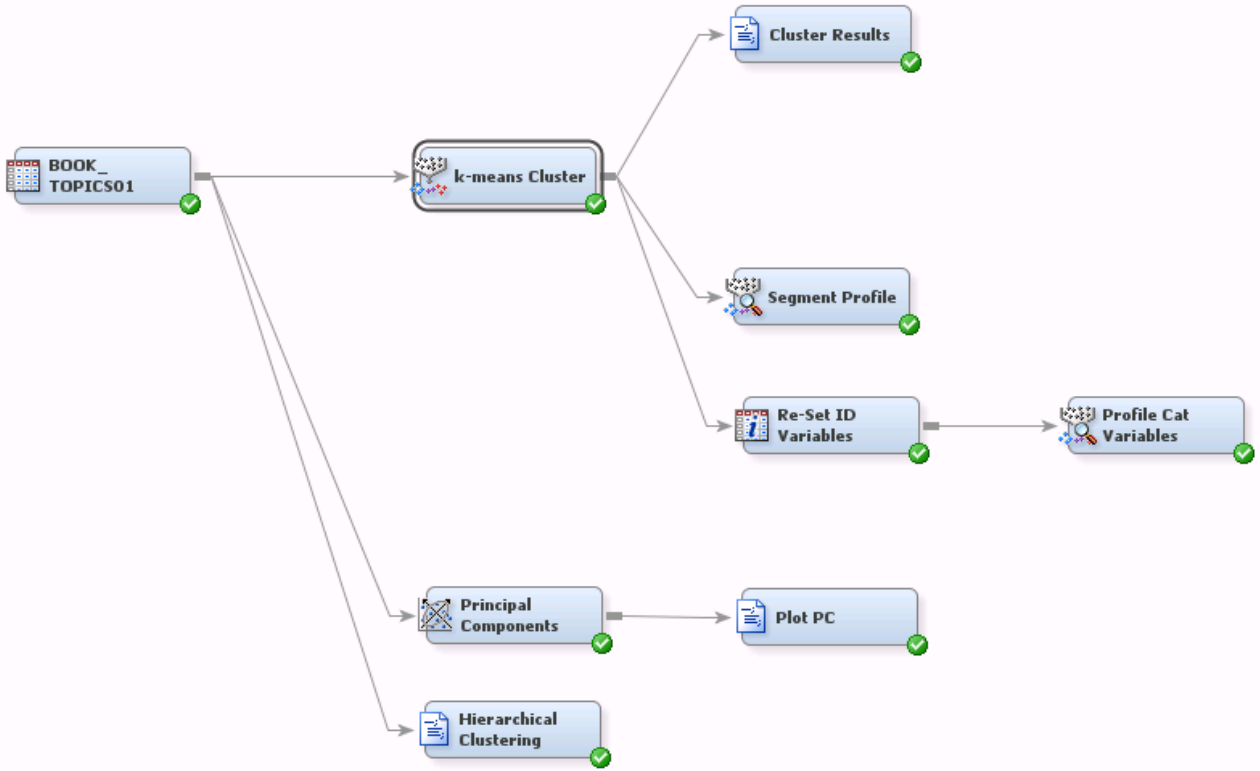
Topic-Weights pro Dokument

VERWENDUNG DER TOPIC-WEIGHTS FÜR EINE CLUSTERANALYSE

	topic_raw1	topic_raw2	topic_raw3	topic_raw4	topic_raw5	topic_raw6	topic_raw7	topic_raw8	topic_raw9	topic_raw10	topic_raw11	_DOCUMENT_	TEXT	URI	N
1	0.047	0.007	0.025	0.002	0.018	0.755	-.001	0.038	0.092	0.022	0.001	1	298 Data Quali...	file://D:\Gerha...	AppA_n
2	0.010	0.064	0.043	-.005	0.014	0.065	0.025	0.015	0.056	0.012	0.014	2	304 Data Quali...	file://D:\Gerha...	AppB_n
3	0.011	0.050	0.276	-.009	0.021	0.049	0.018	0.047	0.015	0.035	0.032	3	306 Data Quali...	file://D:\Gerha...	AppC_n
4	0.054	0.026	0.014	-.053	0.018	0.097	0.023	0.363	0.048	0.015	0.010	4	318 Data Quali...	file://D:\Gerha...	AppD_n
5	0.069	0.112	0.024	0.000	0.075	0.085	0.047	0.026	0.091	-.002	0.002	5	PAGE 368 D...	file://D:\Gerha...	Append
6	0.048	0.039	0.049	0.005	0.086	0.069	0.021	0.034	0.149	0.248	0.043	6	PAGE 382 D...	file://D:\Gerha...	Append
7	-.002	-.043	0.040	-.032	0.091	0.086	0.003	0.023	0.262	0.030	0.093	7	PAGE 390 D...	file://D:\Gerha...	Append
8	0.031	0.016	0.028	0.061	0.736	0.018	0.053	0.019	0.028	0.056	0.032	8	320 Data Quali...	file://D:\Gerha...	AppE_n
9	0.032	0.026	0.037	0.071	0.098	0.028	0.283	0.033	0.013	0.039	0.031	9	PAGE 2 Dat...	file://D:\Gerha...	chap1.n
10	0.165	0.029	0.009	0.022	0.190	0.032	0.058	0.031	0.040	0.058	0.018	10	PAGE 88 Da...	file://D:\Gerha...	chap10.
11	0.022	0.208	0.122	0.061	0.049	0.408	0.050	0.037	0.053	0.043	0.012	11	140 Data Quali...	file://D:\Gerha...	chap10_
12	0.041	0.053	0.045	0.023	0.155	0.022	0.076	0.008	0.083	0.082	0.002	12	PAGE 94 Da...	file://D:\Gerha...	chap11.
13	0.110	0.090	-.004	0.117	0.039	0.473	0.023	0.097	0.082	0.003	0.009	13	156 Data Quali...	file://D:\Gerha...	chap11_
14	0.038	0.023	0.121	0.025	0.055	0.015	0.159	0.021	0.033	0.024	-.009	14	PAGE 100 D...	file://D:\Gerha...	chap12.
15	0.028	0.035	0.030	0.173	0.134	0.064	-.028	0.007	0.003	0.077	0.033	15	164 Data Quali...	file://D:\Gerha...	chap12_
16	0.058	0.024	0.026	0.020	0.103	0.029	0.008	0.015	0.061	0.492	0.016	16	PAGE 104 D...	file://D:\Gerha...	chap13.
17	0.007	0.377	0.092	0.110	0.075	0.064	0.082	0.072	0.069	0.034	0.048	17	178 Data Quali...	file://D:\Gerha...	chap13_
18	0.007	0.004	0.020	-.014	0.150	0.071	0.013	0.028	0.342	0.036	0.091	18	PAGE 130 D...	file://D:\Gerha...	chap14.
19	0.033	0.483	0.124	0.061	0.049	0.115	0.074	0.072	0.035	0.072	0.150	19	198 Data Quali...	file://D:\Gerha...	chap14_
20	0.066	0.012	0.006	-.015	0.064	0.054	0.004	0.030	0.070	0.030	0.013	20	PAGE 138 D...	file://D:\Gerha...	chap15.
21	0.011	0.038	0.223	0.089	0.020	0.037	0.014	0.120	0.025	0.021	0.045	21	Part III: Conseq...	file://D:\Gerha...	chap15_
22	0.076	0.042	0.069	0.127	0.044	0.106	0.053	0.023	0.482	0.043	0.030	22	PAGE 146 D...	file://D:\Gerha...	chap16.
23	0.024	0.091	0.430	0.053	0.048	0.039	0.104	0.125	0.001	0.027	0.032	23	216 Data Quali...	file://D:\Gerha...	chap16_
24	0.077	0.136	0.031	0.010	0.070	0.035	0.036	0.016	0.107	0.284	-.027	24	PAGE 178 D...	file://D:\Gerha...	chap17.
25	0.027	0.041	0.458	0.037	0.025	0.022	0.049	0.068	0.026	0.009	0.020	25	230 Data Quali...	file://D:\Gerha...	chap17_

Cluster Analyse

CLUSTERING DER DOKUMENTE AUF BASIS DER TOPIC WEIGHTS IM SAS ENTERPRISE MINER

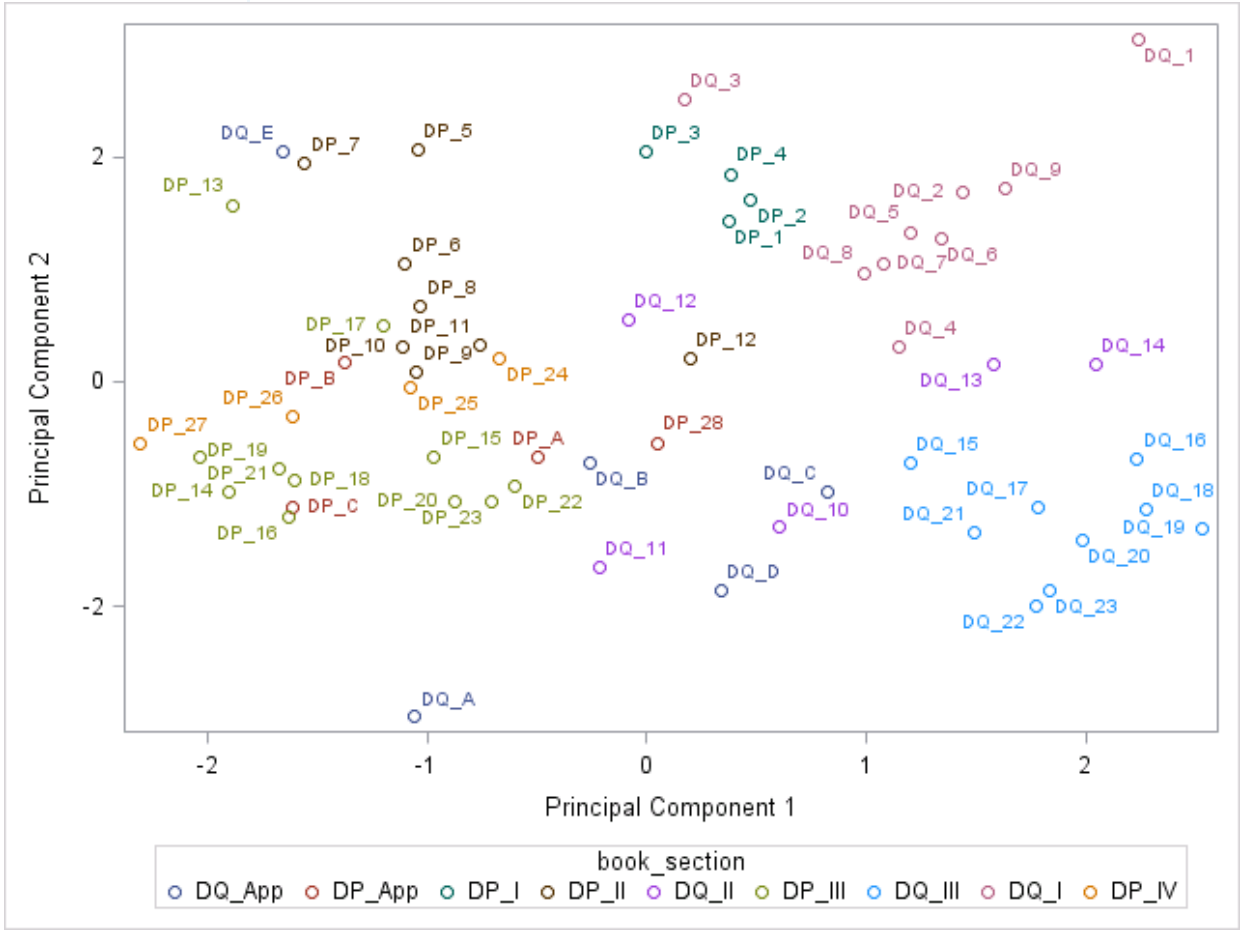


Cluster Analyse Ergebnisse

THEMATISCH ÄHNLICHE KAPITEL WERDEN AUTOMATISCH ZU CLUSTERN ZUSAMMENGEFASST

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	9	A	C	B
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																

SCATTER-PLOT DER KAPITEL BZGL. DER 1. UND 2. HAUPTKOMPONENTE



Text Mining

ZUSAMMENFASSUNG

- Kapitel der beiden Bücher wurden mit SAS® Contextual Analysis automatisch zu sinnvollen Clustern zusammengefasst
- Viele Einsatzgebiete für die automatische Gruppierung von Dokumenten in Unternehmen und Organisationen
- Analyse wurde mit minimaler Interaktion durch den Benutzer durchgeführt (hoher Interaktionsgrad ist möglich: SAS Code, Optionen, LITI-Rules, …)