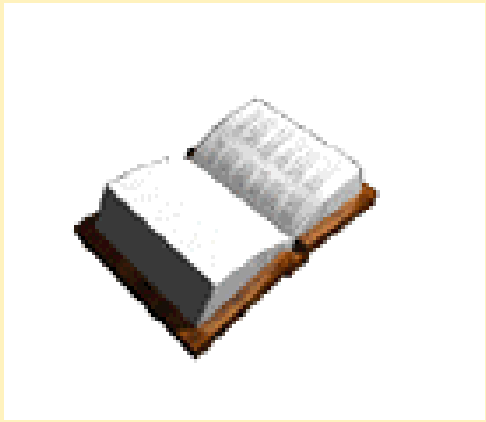
Informationsflut bewältigen - Textmining in der Praxis



Christiane Theusinger
Business Unit Data Mining & CRM Solutions
SAS Deutschland

Ulrich Reincke
Manager Business Data Mining Solutions
SAS Deutschland

The Power to Know.

Problemstellung

- Ist
  - Riesige Datenmengen in operativen Systemen
  - Datenhaltung in meist relationalen Strukturen
  - Erfolgreiche Analysen der Daten durch Data Mining auf der Basis von vorstrukturierten Informationen (Merkmale)
- Problem
  - Viele Informationen liegen in unstrukturierter Form vor, zum Beispiel in Texten: Dokumente, e-mails, Gebrauchsanweisungen, voice-mails etc.
  - Diese Informationen müssen für eine ganzheitliche Sicht aus diesen Daten gewonnen werden können!

Text Mining ist...

das Aufdecken und Verwenden des Wissens, dass in einer Sammlung von Dokumenten als Ganzes existiert. Dabei werden die Dokumente als Summe der in Ihnen vorkommenden Worte behandelt. Es handelt sich somit um eine semantische Dokumentenanalyse

Text Mining ist nicht...

- Suchen von Text-Strings
- Information Retrival
- Spracherkennung
- Sprachverarbeitung

The Power to Know.

Text Mining Anwendungen

- Automatische Klassifikation von Dokumenten
  - Automatisches Filtern von E-mails
  - Einstufung von Dokumenten in verschiedene Kategorien
- Clusterung
  - Intelligente Suche in großen Dokumente-Datenbanken
    - Patente
    - Abstracts aus Bibliografischen Katalogen, z.B. Medline
    - Patienten- und Behandlungsakten in Krankenhäusern
- Vorhersage
  - Kostenprognose basierend auf Call Center Log
  - Kategorisierung von Hotline-Anrufen

The Power to Know.

Schritte des Text Mining

- Texte einlesen
- Datenvorverarbeitung
- Dimensionsreduktion (Singulärwert Zerlegung)
- Text Mining
  - Einsatz verschiedener Verfahren (u.a. Clustering, Neuronale Netze, Regression)
  - Hinzufügen weiterer Variablen zu den Ergebnissen

The Power to Know.

Datenvorverarbeitung 1:

- Einlesen der Texte

- Adobe Portable Document Format (PDF) 1.1 to 4.0
- Applix Asterix Applix Asterix
- Applix Spread sheet 10
- Corel Presentations 7.0, 8.0
- Corel Quattro Pro for windows 7.0, 8.0
- Document Content Architecture (DCA)-RTF sc23-0758-1
- Framemaker Interchange Format (MIF) 5.5
- HTML All
- IBM DisplayWrite 1.0, 1.1
- Lotus 1-2-3 2, 3, 4, 96, 97, R9
- Lotus AMI pro 2.0, 3.0
- Lotus Freelance 96, 97, R9
- Lotus Word pro 96, 97, R9
- Microsoft Excel 3, 4, 5, 97, 98, 2000
- Microsoft PowerPoint 4.0, 95, 97
- Microsoft Rich Text Format All
- Microsoft Word 1.x, 2.0, 6.0, 7.0, 8.0, 95, 97, 2000
- Microsoft Word for DOS 2.2 to 5.0
- Microsoft Word for MAC 4.x, 5.x, 6.x, 98
- Microsoft Works 1.0, 2.0, 3.0, 4.0
- WordPerfect for DOS 5.0, 6.0
- WordPerfect for MAC 2.0, 3.0
- WordPerfect for Windows 7.0
- XYWrite 4.12

Datenvorverarbeitung 2:

- Abgleich mit bekannten Mustern und Synonymen
  - (z.B. Personennamen, Firmennamen wie „SAS Institute“, „SAS“, „SAS Institute Inc.“ und „SAS Institute GmbH“)
- Reduktion der Wortmorphologie
  - z.B „esse und „isst“
- Eliminierung von irrelevanten Wörtern
  - Durch Benutzergesteuerte „Stop Listen“
- Erstellen einer Häufigkeitstabelle der Ausdrücke

- **Am Ende wird ein Dokument repräsentiert als die Summe der darin enthaltenen Worte**

Business Case e-mail-Klassifikation

Ziel:

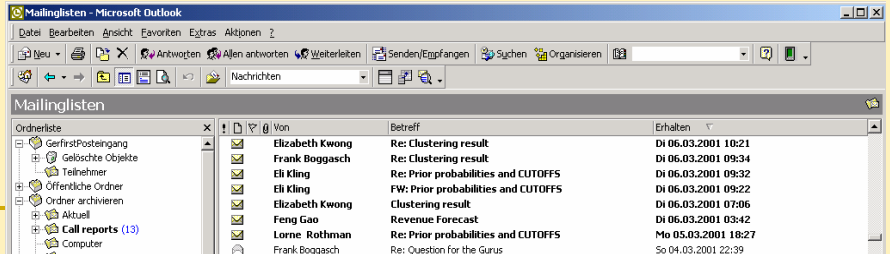
Automatische Klassifikation von e-mails, um eine automatische Vorverarbeitung zu erreichen

Benefits:

- Schnelle Reaktionszeiten
- Ressourceneinsparung
- Prozessautomatisierbarkeit

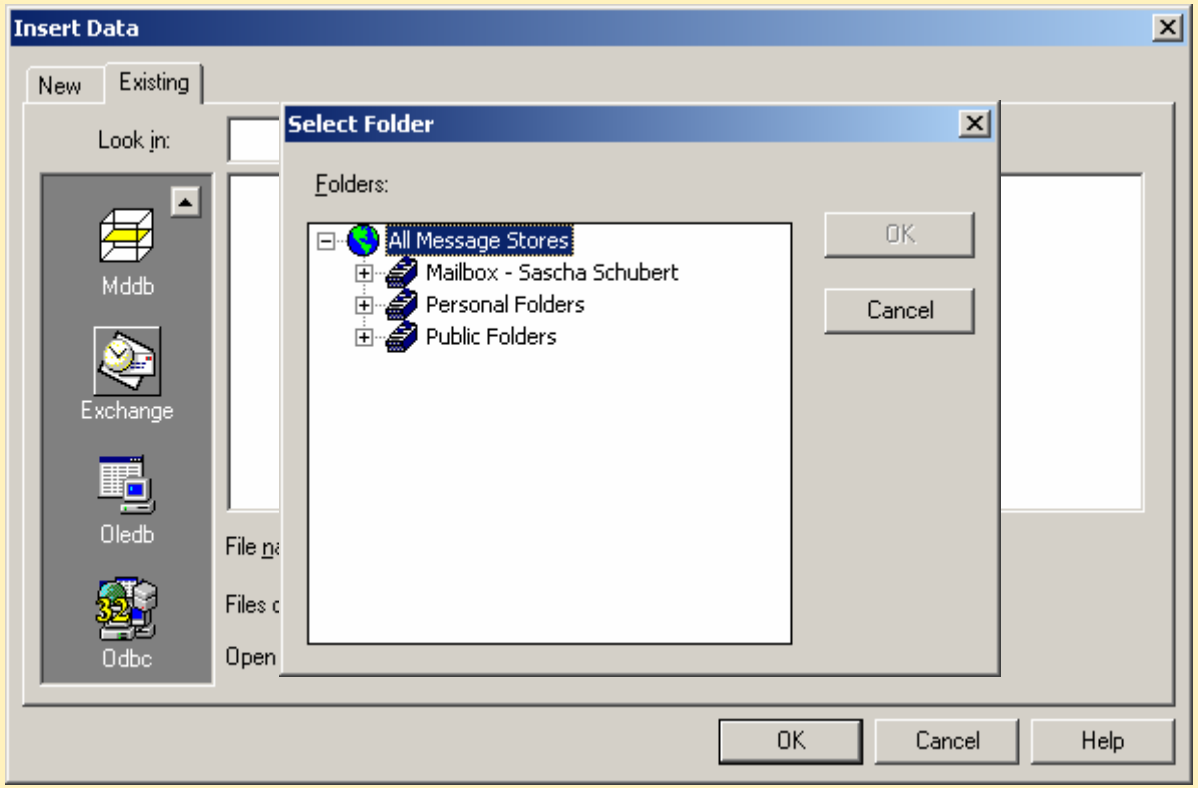
Umsetzung:

- Einlesen der e-mails
- Anwendung eines Modells, das aufgrund des Inhaltes der e-mail klassifiziert (nicht rein auf Schlüsselwörtern basierend)
- Weiterleiten der e-mail zur entsprechenden Stelle / Abteilung



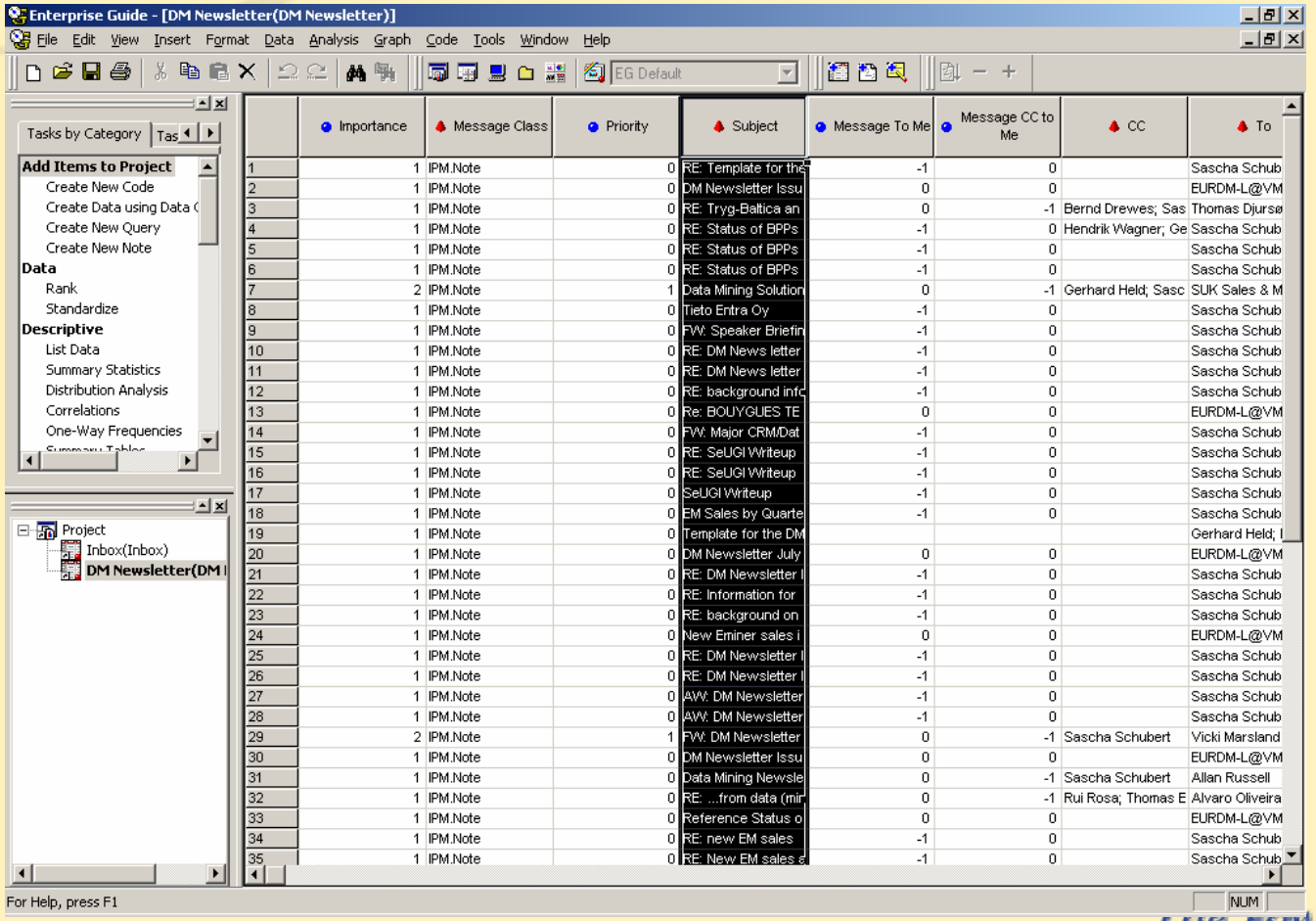
Einlesen der E-Mails in SAS

- Einfache Import Funktion mit dem Enterprise Guide
- Wähle Exchange server icon
- Wähle eine Mailbox



The Power to Know.

Generiert einen strukturierten SAS Datensatz

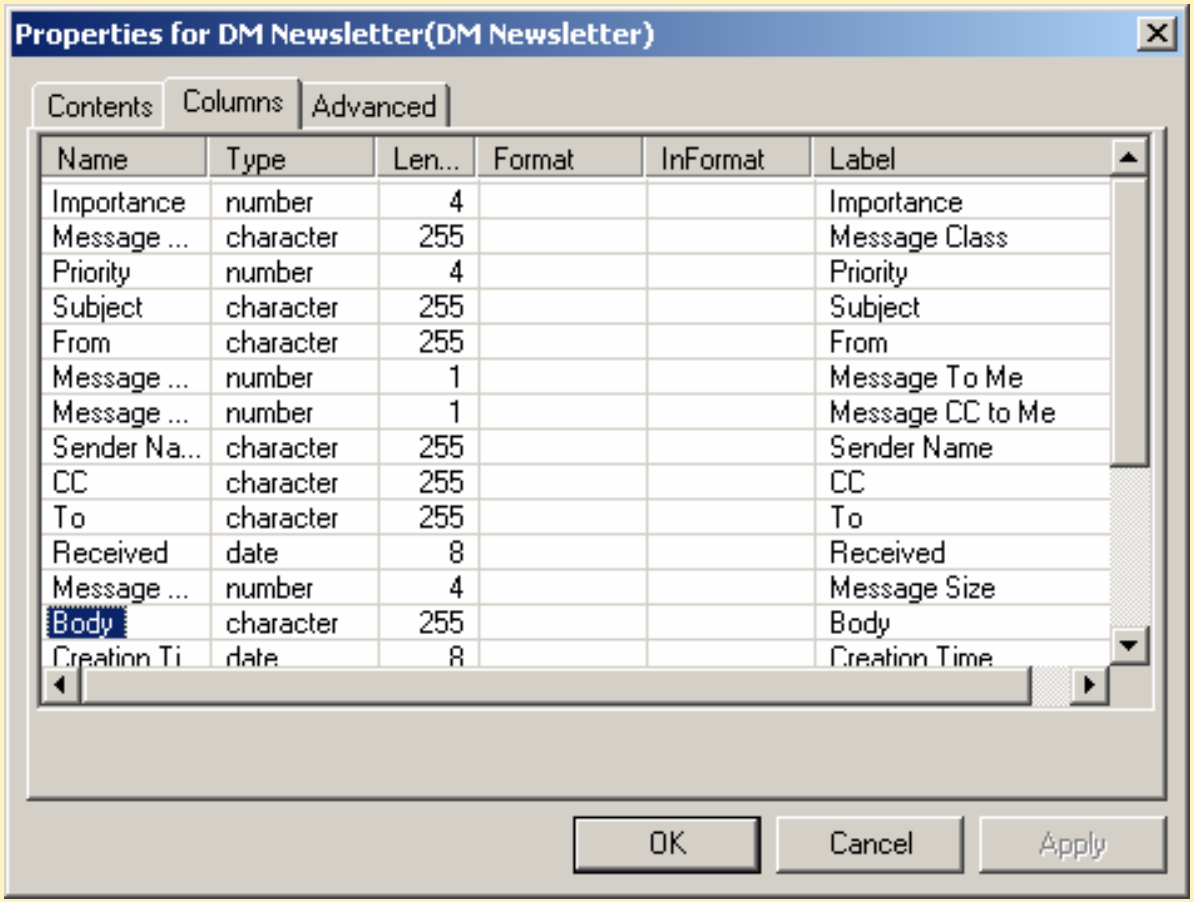


	Importance	Message Class	Priority	Subject	Message To Me	Message CC to Me	CC	To
1	1	IPM.Note	0	RE: Template for the	-1	0		Sascha Schub
2	1	IPM.Note	0	DM Newsletter Issu	0	0		EURDM-L@VM
3	1	IPM.Note	0	RE: Tryg-Baltica an	0	-1	Bernd Drewes; Sas	Thomas Djursø
4	1	IPM.Note	0	RE: Status of BPPs	-1	0	Hendrik Wagner; Ge	Sascha Schub
5	1	IPM.Note	0	RE: Status of BPPs	-1	0		Sascha Schub
6	1	IPM.Note	0	RE: Status of BPPs	-1	0		Sascha Schub
7	2	IPM.Note	1	Data Mining Solution	0	-1	Gerhard Held; Sasc	SUK Sales & M
8	1	IPM.Note	0	Tieto Entra Oy	-1	0		Sascha Schub
9	1	IPM.Note	0	FW: Speaker Briefin	-1	0		Sascha Schub
10	1	IPM.Note	0	RE: DM News letter	-1	0		Sascha Schub
11	1	IPM.Note	0	RE: DM News letter	-1	0		Sascha Schub
12	1	IPM.Note	0	RE: background info	-1	0		Sascha Schub
13	1	IPM.Note	0	Re: BOUYGUES TE	0	0		EURDM-L@VM
14	1	IPM.Note	0	FW: Major CRM/Dat	-1	0		Sascha Schub
15	1	IPM.Note	0	RE: SeUGI Writeup	-1	0		Sascha Schub
16	1	IPM.Note	0	RE: SeUGI Writeup	-1	0		Sascha Schub
17	1	IPM.Note	0	SeUGI Writeup	-1	0		Sascha Schub
18	1	IPM.Note	0	EM Sales by Quarte	-1	0		Sascha Schub
19	1	IPM.Note	0	Template for the DM				Gerhard Held;
20	1	IPM.Note	0	DM Newsletter July	0	0		EURDM-L@VM
21	1	IPM.Note	0	RE: DM Newsletter I	-1	0		Sascha Schub
22	1	IPM.Note	0	RE: Information for	-1	0		Sascha Schub
23	1	IPM.Note	0	RE: background on	-1	0		Sascha Schub
24	1	IPM.Note	0	New Eminer sales i	0	0		EURDM-L@VM
25	1	IPM.Note	0	RE: DM Newsletter I	-1	0		Sascha Schub
26	1	IPM.Note	0	RE: DM Newsletter I	-1	0		Sascha Schub
27	1	IPM.Note	0	AW: DM Newsletter	-1	0		Sascha Schub
28	1	IPM.Note	0	AW: DM Newsletter	-1	0		Sascha Schub
29	2	IPM.Note	1	FW: DM Newsletter	0	-1	Sascha Schubert	Vicki Marsland
30	1	IPM.Note	0	DM Newsletter Issu	0	0		EURDM-L@VM
31	1	IPM.Note	0	Data Mining Newsle	0	-1	Sascha Schubert	Allan Russell
32	1	IPM.Note	0	RE: ...from data (mir	0	-1	Rui Rosa; Thomas E	Alvaro Oliveira
33	1	IPM.Note	0	Reference Status o	0	0		EURDM-L@VM
34	1	IPM.Note	0	RE: new EM sales	-1	0		Sascha Schub
35	1	IPM.Note	0	RE: New EM sales a	-1	0		Sascha Schub

The Power to Know.

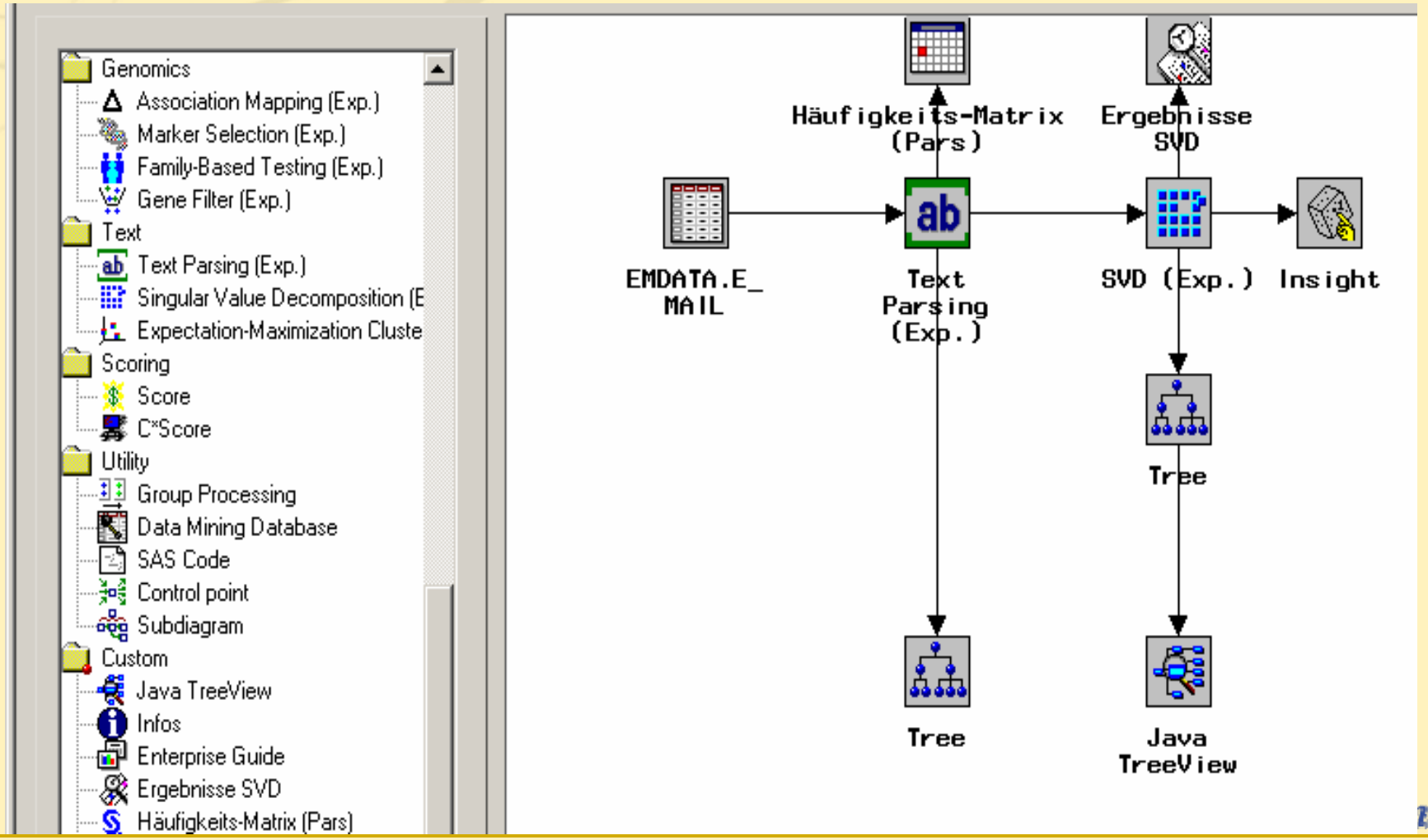
Attribute des E-Mail Data Set

- Sender und Empfänger Attirbute
- Zeit Attribute
- Attachments, Priorität etc.
- Body enthält den text



The Power to Know.

Prozessflussdiagramm



Ergebnisse Textparsing

Obs	TERM	KEY	FREQ	NUMDOCS	stopcnt
39	kidneys	33	1	1	1
40	lens	47	1	1	1
41	leukaemic	98	1	1	1
42	like	54	1	1	1
43	marrow	96	1	1	1
44	metabolic	18	1	1	1
45	mouse	32	1	1	1

Wort x Dokumente Frequenz-Matrix 10:01 Tuesday, February 26, 200

	Erstellungszeit			
	01JUN2001:15:38:28	01JUN2001:16:29:42	01JUN2001:18:01:29	01JUN2001:21:52:05
	N	N	N	N
TERM				
daten	.	.	.	.
de	.	.	.	.
dea	.	.	.	1
dear	.	.	.	.
denen	.	.	.	.
design	.	.	.	.
detailed	.	.	.	.
details	.	.	.	.
deutsche	.	.	1	.
deutschland	.	.	.	.

Verwendete Stop-Liste

Obs	term
1	Ihnen
2	Ihre
3	Ihrem
4	Ihren
5	Ihrer
6	Ihres
7	Sie
8	ab
9	aber
10	acht
11	achtzig
12	all
13	alle
14	allein
15	als
16	also
17	am
18	an
19	anderer
20	andererseits
21	anders
22	auch

Ergebnisse Entscheidungsbaum ohne Textinformationen

treeofzq - SAS/EM Tree

File Edit Model View Tree Window Help

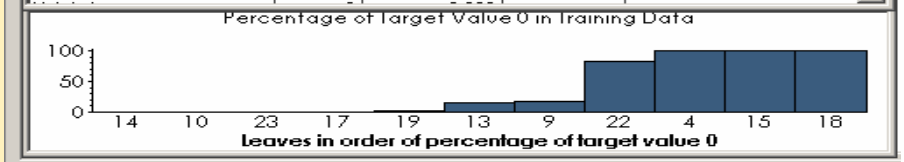
Leaf Statistics Table

		Leaf	N Cases	Predicted Target	Percent 0	Percent 1
▶	+	14	2	1	0	100
2	+	10	13	1	0	100
3	+	23	4	1	0	100
4	+	17	3	1	0	100
5	+	19	619	1	0.323102	99.6769
6	+	13	1088	1	15.8088	84.1912
7	+	9	22	1	18.1818	81.8182
8	+	22	43	0	83.7209	16.2791

Path Rule to Node 1

Variable	Values

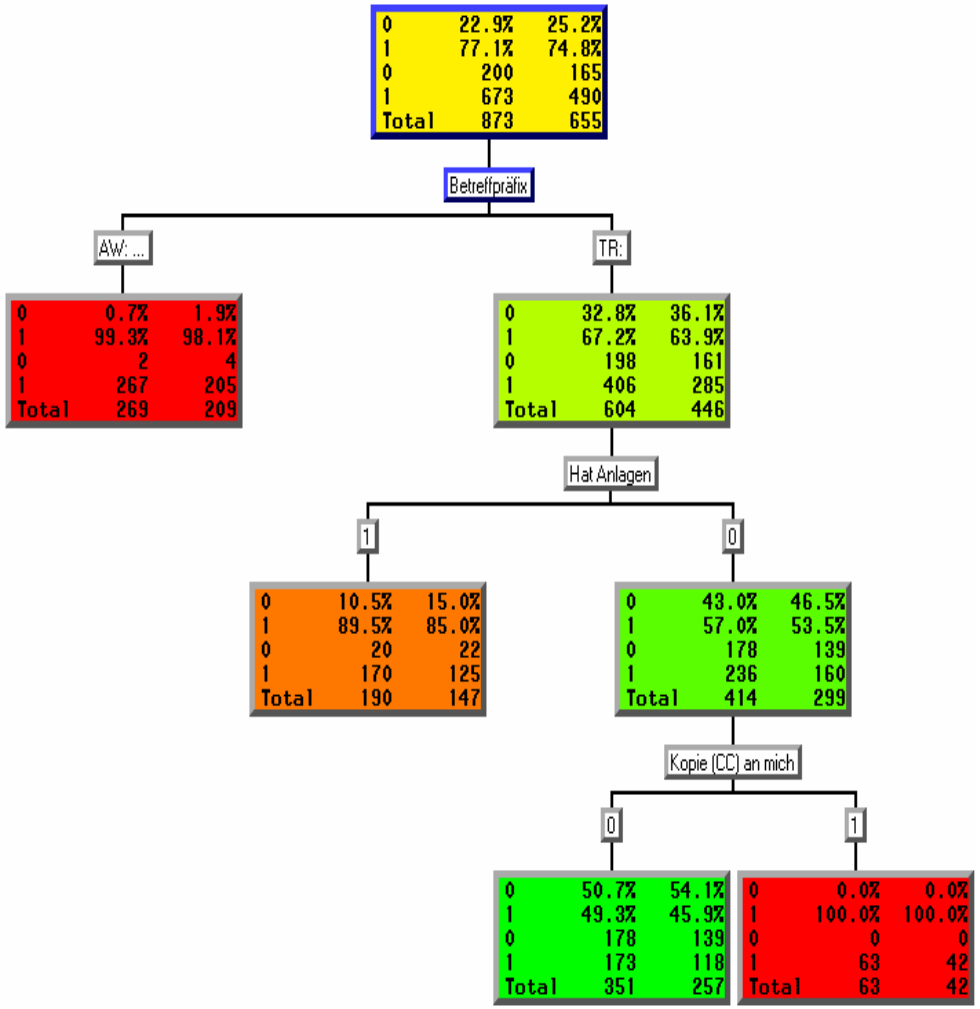
Variable	Nodes	Training
COL2	2	1.000
Betreffpräfix	2	0.261
COL3	1	0.249
COL4	2	0.166
Erhalten	1	0.109
Kopie (CC) an mich	1	0.088
COL1	1	0.068
Objekttyp	0	0.000
Dringlichkeit	0	0.000
Priorität	0	0.000



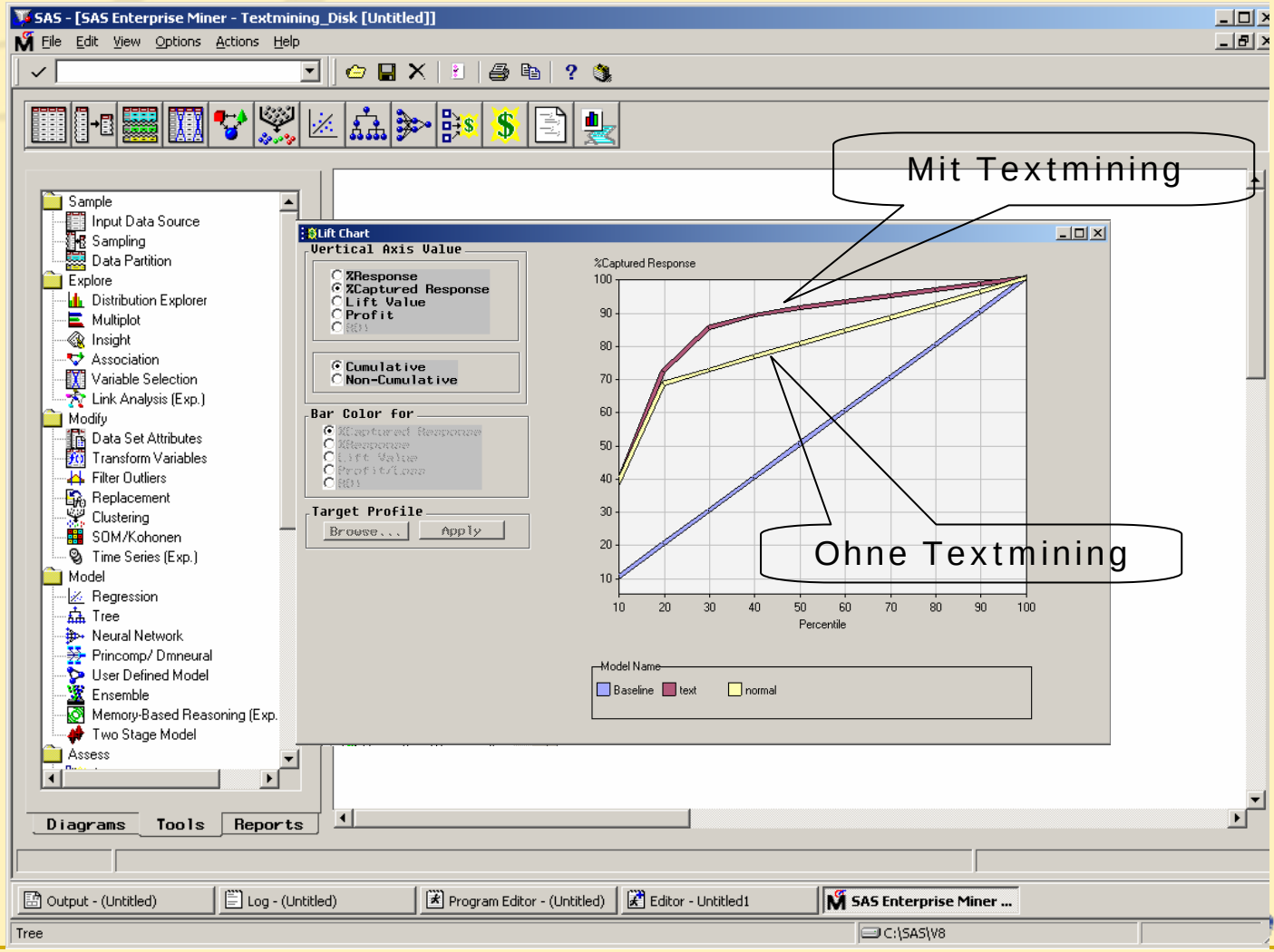
Path Rule to N | Variables | Leaf Statistics | Tree Map | Leaf Statistic B

For Help, press F1

Start | Posteing... | SAS® Hi...



Vergleich beider Ansätze



Zusammenfassung

- Wie beim „richtigen“ Data Mining“ benötigt man für ein Text Mining Projekt eine sehr aufwendige Datenvorverarbeitung (70-90% der Zeit)
- Mit der SAS Text Mining Lösung können Sie vom Einlesen über Reduktion bis hin zum Klassifizieren und Erstellen eines grafischen Output den ganzen Text Mining Prozess abbilden und mit der SAS Internet Technologie sogar weltweit zur Verfüung stellen
- Text Mining kann Ihnen das übergreifende und verdeckte Wissen erschließen, dass sich in großen Dokumente-Sammlungen verbirgt, die kein Mensch allein Lesen kann oder will.

The Power to Know.