

Text Mining Application in Database Marketing



Lisa Chang
James Lally
Maria Skaletsky

Background

- Provider of information infrastructure; products include integrated hardware, software, and services
- Project sponsored by EMC's Database Marketing Organization
- Course Application: Text Mining



Business Challenge

- Data in EMC's customer database consists of information provided by customers during various marketing touch points
- Customers do not enter job title information (or enter false information i.e. "Jefe" or "God").
- Information is used to classify prospects for targeted marketing messaging and activities

Garbage In, Garbage Out

How much money is
your business losing
because of substandard
data quality?





Data

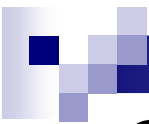
- Participant job information (Job Title, Key Player Description, Key Player Group)
- Objective: Classify “unclassified” and “unknown” into meaningful Key Player Group categories
- Approximately 1 million observations; sample size used includes 10,000 observations

Microsoft Excel - Copy of EMC small data set													
File Edit View Insert Format Tools Data Window Help													
Type a question for help													
M45													
Zoom													
	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Date Entered	Date Updated	Namefirst	Namelast	Job Title	Key Player Desc	Key Player Grp	Mailstate	Mailcountry	Contact Id	Country Desc		
45	5/11/2006 15:15	1/23/2007 10:48	Shelly	Wilson	Facilities Coordinator	General Operations Staff	LOB	ON	Canada	837007	Canada		
46	5/11/2006 15:15	1/23/2007 10:47	Garin	Thompson	Network Director	IT Architecture	IT	OK	United States	837008	United States of America		
47	5/11/2006 15:15	1/23/2007 10:47	Steve	Heinrichs	IT Dir	IT Operations	IT	AR	United States	837009	United States of America		
48	5/11/2006 15:15	1/23/2007 10:47	Karl	Jones	I T Director	IT Architecture	IT	NY	United States	837010	United States of America		
49	5/11/2006 15:15	1/23/2007 10:47	Joe	Walsh	MIS Director	IT Operations	IT	MA	United States	837011	United States of America		
50	5/11/2006 15:15	1/23/2007 10:49	Ray	Riddle		No Title Provided (null)	Unknown	OK	United States	837012	United States of America		
51	5/11/2006 15:15	1/23/2007 10:47	Tom	Blangas	Director of IT	IT Operations	IT	ON	Canada	837013	Canada		
52	5/11/2006 15:15	5/11/2006 15:16	Larry	Gardner	CIO	CIO	CxO	TX	United States	837014	United States of America		
53	5/11/2006 15:15	1/23/2007 10:47	Norm	Goldwyn	Dir of IT	IT Architecture	IT	CA	United States	837015	United States of America		
54	5/11/2006 15:15	1/23/2007 10:49	Jeff	Perry	MIS	No Title Provided (null)	Unknown	MA	United States	837016	United States of America		
55	5/11/2006 15:15	1/23/2007 10:40	Terry	Norton	Network Administrator	Networking Administrator	IT	TX	United States	837017	United States of America		
56	8/18/2006 10:02	8/18/2006 10:02	Eckhart	Beattie	Consultant	Consultants	IT	CA	United States	837018	United States of America		
57	5/11/2006 15:15	1/23/2007 10:47	Miroslav	Dragojevic	IT Manager	IT Operations	IT	CA	United States	837019	United States of America		
58	5/11/2006 15:15	1/23/2007 10:47	Jamie	Jelchuk	IT Manager	IT Operations	IT	ON	Canada	837020	Canada		
59	5/11/2006 15:15	1/23/2007 10:49	Eric	Weasner		No Title Provided (null)	Unknown	CA	United States	837021	United States of America		
60	5/11/2006 15:15	1/23/2007 10:49	Darryl	Martin		No Title Provided (null)	Unknown	IL	United States	837022	United States of America		
61	5/11/2006 15:15	1/23/2007 10:49	Doug	Ward		No Title Provided (null)	Unknown	IL	United States	837023	United States of America		
62	5/11/2006 15:15	1/23/2007 10:49	Danna	Tillamant		No Title Provided (null)	Unknown	IL	United States	837024	United States of America		
63	5/11/2006 15:15	1/23/2007 10:47	Emma	Kavanaugh	Manager of Business Systems	IT Architecture	IT		Canada	837025	Canada		
64	5/11/2006 15:15	1/23/2007 10:40	Scott	Waterhouse	MIS Manager	Systems Administrator	IT	MA	United States	837026	United States of America		
65	5/11/2006 15:15	1/23/2007 10:49	Jonathan	Coose		No Title Provided (null)	Unknown	IL	United States	837027	United States of America		
66	5/11/2006 15:15	1/23/2007 10:47	Ron	Stewart	Leader of IT	IT Architecture	IT	ON	Canada	837028	Canada		
67	5/11/2006 15:15	1/23/2007 10:49	Pedro	Contreras		No Title Provided (null)	Unknown	TX	United States	837029	United States of America		
68	5/11/2006 15:15	1/23/2007 10:49	Adam	Xxxxxxxx		No Title Provided (null)	Unknown	IL	United States	837030	United States of America		
69	5/11/2006 15:15	1/23/2007 10:47	William	Chu	Manager of IT Operations	IT Operations	IT		Canada	837031	Canada		
70	5/11/2006 15:15	11/9/2007 16:31	Hal	Hagan	Manager of ETS	Unclassified	Unclassified	ON	Canada	837033	Canada		
71	5/11/2006 15:15	1/23/2007 10:47	Jack	Chow	Manager IT	IT Operations	IT	ON	Canada	837034	Canada		
72	5/11/2006 15:15	1/23/2007 10:49	Sumitro	Chowdhury		No Title Provided (null)	Unknown	TX	United States	837035	United States of America		
73	5/11/2006 15:15	1/23/2007 10:49	Chase	Miller		No Title Provided (null)	Unknown	TX	United States	837036	United States of America		
74	5/11/2006 15:15	1/23/2007 10:49	Jerry	Rupp		No Title Provided (null)	Unknown	TX	United States	837037	United States of America		
75	5/11/2006 15:15	1/23/2007 10:40	Lynn	Maust	Sys Admin	Systems Administrator	IT	PA	United States	837038	United States of America		
76	5/11/2006 15:15	1/23/2007 10:49	Scott	Ellenberger		No Title Provided (null)	Unknown	CO	United States	837039	United States of America		
77	5/11/2006 15:15	1/23/2007 10:49	Marcia	Morrow		No Title Provided (null)	Unknown	TX	United States	837040	United States of America		
78	5/11/2006 15:15	1/23/2007 10:40	Bill	Covington	Network Engineer	Networking Administrator	IT	NC	United States	837041	United States of America		
79	5/11/2006 15:15	1/23/2007 10:47	Dean	Quiram	Director of IT Operations	IT Architecture	IT	CO	United States	837042	United States of America		
80	5/11/2006 15:15	1/23/2007 10:40	Brian	Asci	Systems Engineer	Systems Administrator	IT	PA	United States	837043	United States of America		
81	5/11/2006 15:15	1/23/2007 10:49	Ernest	Escobedo		No Title Provided (null)	Unknown	PA	United States	837044	United States of America		
82	5/11/2006 15:15	1/23/2007 10:49	Dean	Hubert		No Title Provided (null)	Unknown	AZ	United States	837045	United States of America		
83	5/11/2006 15:15	1/23/2007 10:49	Peter	Gustafson		No Title Provided (null)	Unknown	TX	United States	837046	United States of America		
84	5/11/2006 15:15	1/23/2007 10:49	Stephen	Robertson		No Title Provided (null)	Unknown	TX	United States	837048	United States of America		
85	5/11/2006 15:15	1/23/2007 10:40	mark	jeffers	LAN Administrator	Networking Administrator	IT	AZ	United States	837049	United States of America		
86	5/11/2006 15:15	1/23/2007 10:49	Stan	Sargent		No Title Provided (null)	Unknown	AZ	United States	837050	United States of America		
87	5/11/2006 15:15	1/23/2007 10:40	Ray	Wiley	Network	Networking Administrator	IT	LA	United States	837051	United States of America		
88	5/11/2006 15:15	1/23/2007 10:49	Steve	Eschenburg		No Title Provided (null)	Unknown	TX	United States	837055	United States of America		
89	5/11/2006 15:15	1/23/2007 10:52	Chris	Powell	Storage Administrator	Storage Administrator	IT	PA	United States	837056	United States of America		
90	5/11/2006 15:15	1/23/2007 10:49	Mike	Rock		No Title Provided (null)	Unknown	AZ	United States	837057	United States of America		
91	5/11/2006 15:15	11/9/2007 16:31	Alfonso	Linares	Manager of production syst	Unclassified	Unclassified	TX	United States	837059	United States of America		
92	5/11/2006 15:15	1/23/2007 10:47	Toby	Matula	IT Manager	IT Operations	IT	TX	United States	837060	United States of America		
93	5/11/2006 15:15	1/23/2007 10:47	Troy	Sanchez	LAN Admin	IT Operations	IT	NY	United States	837061	United States of America		
94	5/11/2006 15:15	1/23/2007 10:49	Hal	Porter		No Title Provided (null)	Unknown	TX	United States	837062	United States of America		
95	5/11/2006 15:15	1/23/2007 10:47	Kyle	Chapman	IT Director	IT Operations	IT	MO	United States	837063	United States of America		
96	5/11/2006 15:15	1/23/2007 10:49	Sean	Reed		No Title Provided (null)	Unknown	MS	United States	837064	United States of America		
97	5/11/2006 15:15	1/23/2007 10:49	Scott	Bell	Executive Director, Infrastru	Other Officers	CxO	CO	United States	837066	United States of America		
98	5/11/2006 15:15	1/23/2007 10:49	Adam	Sears	Project Leader	Other Officers	CxO	AZ	United States	837067	United States of America		
99	5/11/2006 15:15	1/23/2007 10:49	Randy	Holliday		No Title Provided (null)	Unknown	CA	United States	837068	United States of America		
100	5/11/2006 15:15	1/23/2007 10:47	PHIL	WALTHER	WAN ENGINEER	IT Architecture	IT	VA	United States	837069	United States of America		
101	5/11/2006 15:15	1/23/2007 10:49	Kenneth	Logsdon		No Title Provided (null)	Unknown	KY	United States	837070	United States of America		
102	5/11/2006 15:15	1/23/2007 10:48	Eric	Freed	SUPERVISOR OF NETWORK	General Operations Staff	LOB	AZ	United States	837071	United States of America		
103	5/11/2006 15:15	1/23/2007 10:52	Jay	Calbi	sr midrange systems	Storage Administrator	IT	NJ	United States	837072	United States of America		
104	5/11/2006 15:15	1/23/2007 10:49	David	Johnson		No Title Provided (null)	Unknown	TX	United States	837075	United States of America		
105	5/11/2006 15:15	11/9/2007 16:31	Patrick	Kelley	in between	Unclassified	Unclassified	CA	United States	837076	United States of America		



Approach

- Remove non-value-added data
- Create new variable: Results (concatenation of Job Title and Key Player Description)
- Perform cluster analysis using new variable:
 - many of the terms included in cluster terms were not informative i.e. 'Unclassified', 'None', 'No', 'Title', 'Provided'
 - all cases that had a Key Player Description 'Unclassified' were put into this cluster
- Set the status of those terms to inactive, which forced the cases that had 'Unclassified' Key Player descriptions to be assigned to clusters other than 'Unclassified' and received much more meaningful results



Cluster Analysis Results

■ Clusters created based on Job Title and Key Player Description

Descriptive Terms	Freq	Percentage	RMS Std.
lan + specialist, development, mis, + application, developer, webmaster, + engineer, analyst, no, dba, software, + programmer, administrator, networking	1341	13%	0.130250501
head, director, + partner, pharmacy, + officer, cno, president, other, vice, + program, regional, + facility, nursing, nursing, engineering	925	9%	0.077506336
owner, other, financier, broker, tc, principal, branch, ceo, treasurer, chairman, board, comptroller, cfo, directeur, direttore	767	8%	0.110861492
marketing, + sale, personnel, hr, purchasing, vice, office, advertising, publisher, + resource, staff, + market, human, marketing, exec	1400	14%	0.133306608
infrastructure, + engineer, architecture, responsible, ti, + network, architech, informatique, + service, gerente, vp, engineer, telecom, client, manager	544	5%	0.095722524
+ consultant, + technick analyst, manager, consultor, support, + programer, + operation, consultants, + telecommunication, + adviser, sr., operations, senior, business	325	3%	0.115322029
admin, + facility, general, administrative, city, receptionist, + admission head, staff, administrative, secretary, administration, management, assistant, managing	822	8%	0.104434225
senior, executive, purchasing, + service, staff, controller, director, + computer, chief, + purchase, vp + operation, supervisor, plant, coordinator	3876	39%	0.15130129

...continued

Crosstabs between Cluster memberships and Key Job Descriptions

Table of Key_Player_Desc by _CLUSTER_

Key_Player_Desc	_CLUSTER_(Cluster ID)				
Frequency	1,	2,	3,	4,	Total
Percent					
Row Pct					
Col Pct					
Application Administrator	29 ,	0 ,	0 ,	0 ,	30
	0.29 ,	0.00 ,	0.00 ,	0.00 ,	0.30
	96.67 ,	0.00 ,	0.00 ,	0.00 ,	
	2.17 ,	0.00 ,	0.00 ,	0.00 ,	
Application Development	98 ,	0 ,	0 ,	0 ,	105
	0.98 ,	0.00 ,	0.00 ,	0.00 ,	1.05
	93.33 ,	0.00 ,	0.00 ,	0.00 ,	
	7.32 ,	0.00 ,	0.00 ,	0.00 ,	
CEO	0 ,	0 ,	506 ,	0 ,	668
	0.00 ,	0.00 ,	5.08 ,	0.00 ,	6.70
	0.00 ,	0.00 ,	75.75 ,	0.00 ,	
	0.00 ,	0.00 ,	66.14 ,	0.00 ,	

- Crosstab created
 - all or most cases with the same Key Player description were placed into the same cluster



...continued

- Forced cases that had a Job Title but did not have a Key Player Description to be placed into these clusters
- After completing the cluster analysis, we saved a dataset with only cases that had the Key Player Description as 'Unclassified' and checked which clusters they were placed in.
- 166 cases out of 10,000 – this enabled us to check whether or not the membership was accurate
- Model limitation: Due to lack of time, we did not thoroughly review the extracted terms; also, there are some terms remaining that do not provide any real information

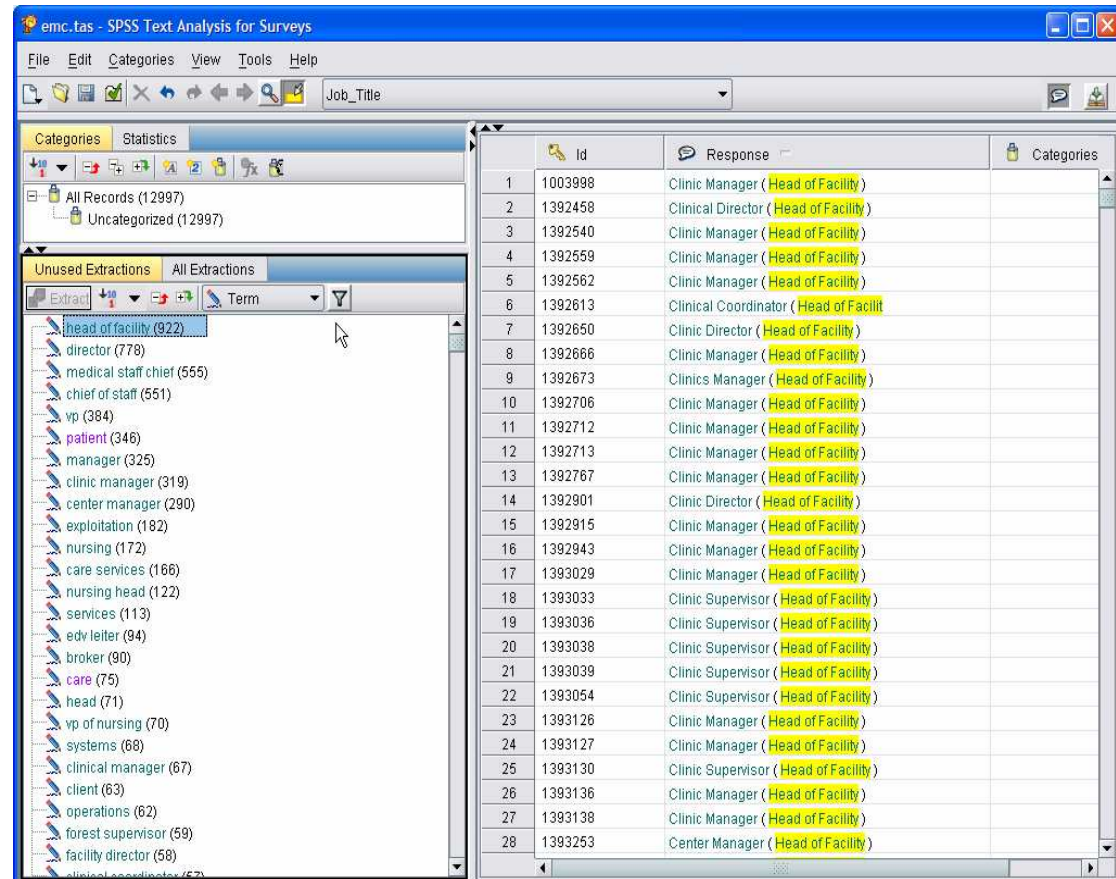


SPSS Text Analysis

- Using the originally cleaned dataset and saving only the cases where Key Player Description was 'Unclassified' (12,997 observations)
- Converted it into SPSS dataset and imported into SPSS Text Analysis
- Ran the analysis on Job Title variable. Multiple common terms were extracted that could be used to classify the unclassified Job Titles.

Example

- There are 922 cases with 'Head of Facility' included in the Job Title.
- We could create a category named 'Head of Facility' and move all 922 cases into that category. Based on the created categories, we could then assign Key Player Description values to these cases.
- Requires more manual work than SAS Text Miner clustering method, but it could produce more accurate results.



	Id	Response	Categories
1	1003998	Clinic Manager (Head of Facility)	
2	1392458	Clinical Director (Head of Facility)	
3	1392540	Clinic Manager (Head of Facility)	
4	1392559	Clinic Manager (Head of Facility)	
5	1392562	Clinic Manager (Head of Facility)	
6	1392613	Clinical Coordinator (Head of Facility)	
7	1392650	Clinic Director (Head of Facility)	
8	1392666	Clinic Manager (Head of Facility)	
9	1392673	Clinics Manager (Head of Facility)	
10	1392706	Clinic Manager (Head of Facility)	
11	1392712	Clinic Manager (Head of Facility)	
12	1392713	Clinic Manager (Head of Facility)	
13	1392767	Clinic Manager (Head of Facility)	
14	1392901	Clinic Director (Head of Facility)	
15	1392915	Clinic Manager (Head of Facility)	
16	1392943	Clinic Manager (Head of Facility)	
17	1393029	Clinic Manager (Head of Facility)	
18	1393033	Clinic Supervisor (Head of Facility)	
19	1393036	Clinic Supervisor (Head of Facility)	
20	1393038	Clinic Supervisor (Head of Facility)	
21	1393039	Clinic Supervisor (Head of Facility)	
22	1393054	Clinic Supervisor (Head of Facility)	
23	1393126	Clinic Manager (Head of Facility)	
24	1393127	Clinic Manager (Head of Facility)	
25	1393130	Clinic Supervisor (Head of Facility)	
26	1393136	Clinic Manager (Head of Facility)	
27	1393138	Clinic Manager (Head of Facility)	
28	1393253	Center Manager (Head of Facility)	

Analysis of Results

#	Descriptive Terms	Freq	Percentage	RMS Std.
1	lan, + specialist, development, mis, + application, developer, webmaster, + engineer, analyst, no, dba, software, + programmer, administrator, networking	1341	13%	0.1302505011
2	head, director, + partner, pharmacy, + officer, cno, president, other, vice, + program, regional, + facility, nursing, nursing, engineering	925	9%	0.0775063363
3	owner, other, financier, broker, tc, principal, branch, ceo, treasurer, chairman, board, comptroller, cfo, directeur, direttore	767	8%	0.1108614917
4	marketing, + sale, personnel, hr, purchasing, vice, office, advertising, publisher, + resource, staff, + market, human, marketing, exec	1400	14%	0.1333066082
5	infrastructure, + engineer, architecture, responsable, ti, + network, architect, informatique, + service, gerente, vp, engineer, telecom, client, manager	544	5%	0.0957225243
6	+ consultant, + technic, analyst, manager, consultor, support, + programmer, + operation, consultants, + telecommunication, + adviser, sr., operations, senior, business	325	3%	0.1153220287
7	admin, + facility, general, administrative, city, receptionist, + admission, head, staff, administrative, secretary, administration, management, assistant, managing	822	8%	0.1044342251
8	senior, executive, purchasing, + service, staff, controller, director, + computer, chief, + purchase, vp, + operation, supervisor, plant, coordinator	3876	39%	0.1513012903

Classifications:

1 – Administrator/Developer (App, DBA, Network) - IT Department

2 – Pharma, Life science, Healthcare

3 – Finance, CxO

4 – Functional area – Support

5 – Engineer/Architects, Tech Gurus - Create IT infrastructure

6 – Middle Management

7 – Administration

8 – Upper Management



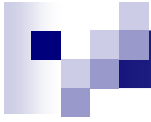
Analysis of Results cont.

- *Clusters map well to:*

- ☐ *“Top of the building” – CxO*
- ☐ *“Back of the building” – Developers and Administrators*
- ☐ *Support and Administrative*

- *Need to look into:*

- ☐ *CIO placement*
- ☐ *Refine Cluster 8*



Questions / Comments?