Using Galileo's Catpac and ThoughtView Software to Analyze Texts

The Galileo Company (www.galileoco.com), is the brainchild of Dr. Joseph Woelfel, Professor in the School of Informatics, State University of New York at Buffalo. Galileo software allows users to scan lengthy texts, to identify recurring phrases and concepts. It has been used in marketing research, to analyze transcripts from focus groups. Galileo has several potential uses in archives and libraries. For example, it could identify and rank frequently occurring phrases in a collection of digitized manuscripts, assisting in the construction of archival finding aids. As a test, we used Galileo to process a collection of about 50 advance fee fraud letters.

Galileo's Catpac program can analyze narratives and identify key concepts and the strengths of their relationships, by means of a neural network. The companion ThoughtView program can generate perceptual maps which correspond to SSM's "rich pictures" or entity relationship diagrams, but which are determined by precise and replicable mathematics, and derived from the data themselves, not from the analyst's interpretation of the data. Because the output of a Galileo study contains only concepts, and the strengths of their relationships, Galileo output may provide both more solid ground for the determination of T-W parings and more understandable graphics for presentation to clients. According to the Catpac manual:

> *Catpac* is a self-organizing artificial neural network that has been optimized for reading text. *Catpac* is able to identify the most important words in a text and determine patterns of similarity based on the way they're used in text. It does this by assigning a neuron

Library & Archival Security, Vol. 19(1) 2004
Available online at http://www.haworthpress.com/web/LAS

doi:10.1300/J114v19n01_09

> to each major word in the text. It then runs a scanning window through the text. The neuron representing a word becomes active when that word appears in the window, and remains active as long as the word remains in the window. Up to *n* words can be in the window at once, where *n* is a parameter set by the user.
>
> As in the human brain, the connections between neurons that are simultaneously active are strengthened following the law of classical conditioning. The pattern of weights or connections among neurons forms a representation within *Catpac* of the associations among the words in the text. This pattern of weights represents complete information about the similarities among all the words in the text. Technically, the pattern of connections among neurons is a complete paired comparison similarities matrix, and so lends itself to the most powerful and sophisticated of statistical analyses. (Woelfel)

As a test, we concatenated about fifty "419 letters" to determine Galileo's effectiveness in identifying key concepts, and to arrive at a profile of the language used in typical advance fee fraud schemes. Using common UNIX utilities such as the vi editor and the group search and replace program, we first pasted the letters into a single file. We then removed extraneous lines, such as those in the e-mail headers, and sorted the resulting lines of text, leaving a file containing only the content of the messages. This file was then processed by Catpac and ThoughtView.

The relationships among the first twenty-five significant words in the text are summarized in Figure 1. The frequency list and its associated alphabetically ordered list allow users to identify the most commonly occurring words in a body of text, in this case, approximately fifty "419 letters," within a few seconds. Galileo automatically produces bar-charts which demonstrate the frequencies in graphical form. The entire process took just a few seconds given an input file containing 3147 words (see Figure 2).

Next, Galileo's ThoughtView software was used to produce the perceptual map in Figure 3. This graph has been rotated about the X, Y, and Z axes, so that the labels would be more readable. This two-dimensional rendering of a three-dimensional display may appear a bit confusing at first. Several concepts are so closely associated that their labels overlap. However, the association of the concept of pleading with that of govern-

It then runs a scanning window
presenting a word becomes active
indow, and remains active as long
low. Up to *n* words can be in the
rameter set by the user.
connections between neurons that
rengthened following the law of
ttern of weights or connections
itation within *Catpac* of the asso-
ext. This pattern of weights repre-
ut the similarities among all the
he pattern of connections among
mparison similarities matrix, and
ful and sophisticated of statistical

ty "419 letters" to determine Gali-
concepts, and to arrive at a profile
e fee fraud schemes. Using com-
itor and the group search and re-
etters into a single file. We then
hose in the e-mail headers, and
ng a file containing only the con-
then processed by Catpac and

twenty-five significant words in
The frequency list and its associ-
users to identify the most com-
text, in this case, approximately
onds. Galileo automatically pro-
he frequencies in graphical form.
nds given an input file containing

vare was used to produce the per-
s been rotated about the X, Y, and
e readable. This two-dimensional
lay may appear a bit confusing at
ssociated that their labels overlap.
ot of pleading with that of govern-

FIGURE 1

TOTAL WORDS	3147	THRESHOLD	0.000
TOTAL UNIQUE WORDS	25	RESTORING FORCE	0.100
TOTAL EPISODES	3141	CYCLES	1
TOTAL LINES	3121	FUNCTION	Sigmoid (-1 - +1)
		CLAMPING	Yes

DESCENDING FREQUENCY LIST

WORD	FREQ	PCNT	CASE FREQ	CASE PCNT
I	775	24.6	2250	71.6
WILL	253	8.0	1170	37.2
COM	172	5.5	487	15.5
MONEY	165	5.2	722	23.0
ME	163	5.2	855	27.2
RECEIVED	159	5.1	278	8.9
ACCOUNT	126	4.0	677	21.6
US	116	3.7	638	20.3
AM	108	3.4	503	16.0
COMPANY	87	2.8	461	14.7
BANK	84	2.7	447	14.2
TRANSACTION	82	2.6	406	12.9
BUSINESS	78	2.5	391	12.4
GOD	78	2.5	329	10.5
FUND	75	2.4	443	14.1
MILLION	72	2.3	390	12.4
WANT	72	2.3	386	12.3
MR	70	2.2	313	10.0
PLEASE	67	2.1	295	9.4
COUNTRY	63	2.0	358	11.4
MESSAGE	59	1.9	156	5.0
CONTACT	58	1.8	320	10.2
ID	56	1.8	138	4.4
DOLLARS	55	1.7	258	8.2
GOVERNMENT	54	1.7	260	8.3

ALPHABETICALLY SORTED LIST

WORD	FREQ	PCNT	CASE FREQ	CASE PCNT
ACCOUNT	126	4.0	677	21.6
AM	108	3.4	503	16.0
BANK	84	2.7	447	14.2
BUSINESS	78	2.5	391	12.4
COM	172	5.5	487	15.5
COMPANY	87	2.8	461	14.7
CONTACT	58	1.8	320	10.2
COUNTRY	63	2.0	358	11.4
DOLLARS	55	1.7	258	8.2
FUND	75	2.4	443	14.1
GOD	78	2.5	329	10.5
GOVERNMENT	54	1.7	260	8.3
I	775	24.6	2250	71.6
ID	56	1.8	138	4.4
ME	163	5.2	855	27.2
MESSAGE	59	1.9	156	5.0
MILLION	72	2.3	390	12.4
MONEY	165	5.2	722	23.0
MR	70	2.2	313	10.0
PLEASE	67	2.1	295	9.4
RECEIVED	159	5.1	278	8.9
TRANSACTION	82	2.6	406	12.9
US	116	3.7	638	20.3
WANT	72	2.3	386	12.3
WILL	253	8.0	1170	37.2

ment is immediately apparent. Typical 419 letters appeal to the recipient's sense of sympathy for ills undergone at the hands of nefarious governments. The cluster of concepts to the left of the diagram appears to be associated with the method of concluding the proposed business transaction, a money-laundering scheme. The cluster to the right appears to be associated with the sender's claim of having come into possession of a certain sum of money through a business transaction.

This test is somewhat artificial, since the content of 419 letters is easily recognized without recourse to analysis programs. However, the Galileo software could be of considerable value when the contents of a collection of documents are not known beforehand. For instance, a digitized collection of private correspondence could be run through Galileo, in order to identify key concepts to be used in the construction of archival finding aids.

FIGURE 2

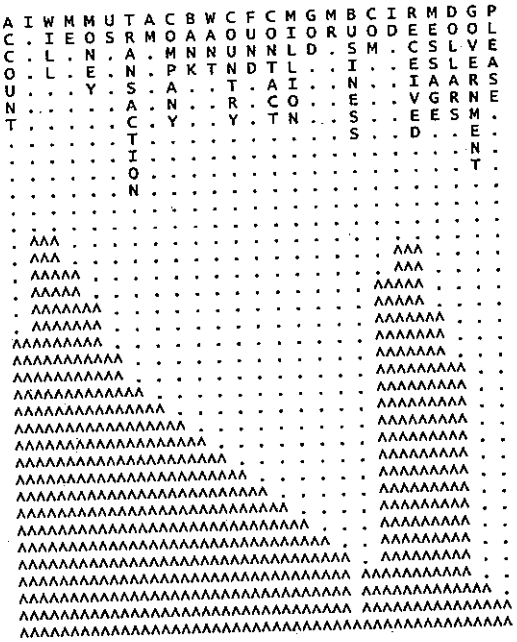
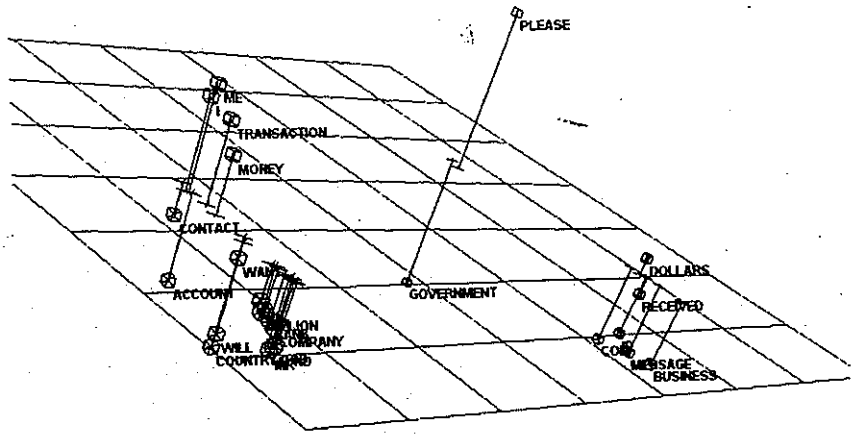


FIGURE 3

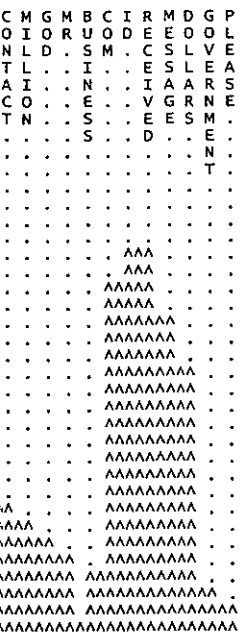


REFERENCE

Woelfel, J. K. (1998). *User's guide; Catpac II, Version 2.0.* Buffalo, The Galileo Company.

'AL SECURITY

E 2



E 3