SAS® GLOBAL FORUM 2017

April 2 – 5 | Orlando, FL

#SASGF

USERS PROGRAM

What's in *Your* Inbox?: An Introduction to Text Mining Using SAS®

Andrew Clapson, MD Financial Management

First, why bother?

- Potentially valuable analytical information stored in text.
- Corporate/agency knowledge ‘coded’ into words and phrases
- But…not easy!
- Our brains have evolved to understand and draw meaning from context and narrative
- Computers read instructions and execute them quite literally, so need a bit more help attributing meaning to the written word

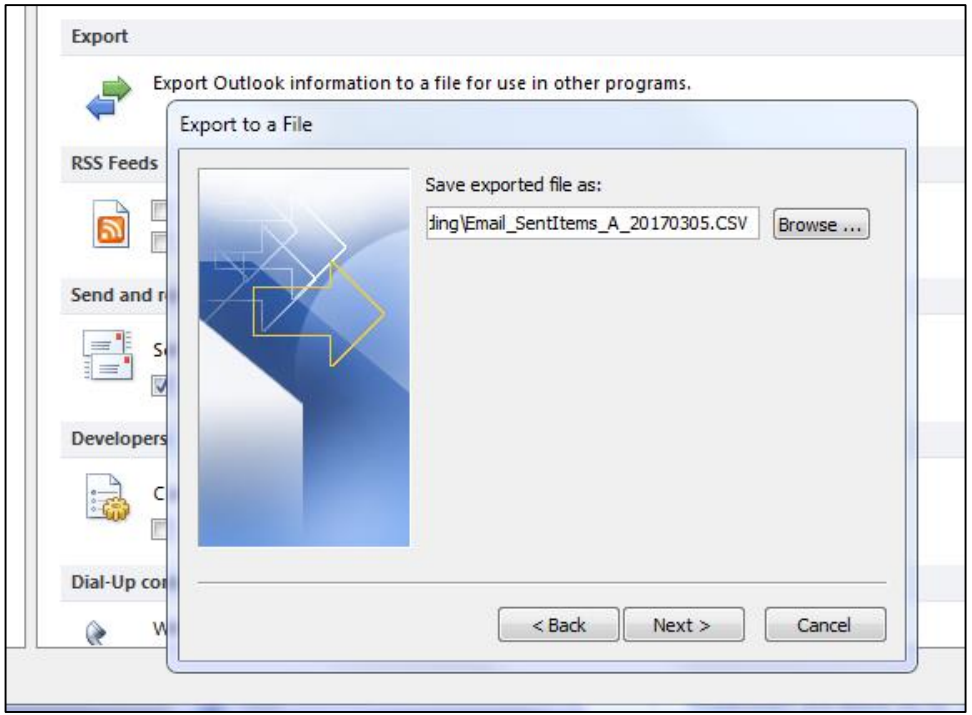
The process (and what I'll cover)

- Extract Outlook email data as a text file
- Bring the data into SAS
- Clean and process the data (ETL)
- Keep items of value (text), defining calculated fields, and just generally being clever
- Where next?

Blood from a stone

Extracting your data out of Outlook

- Microsoft Outlook 2012
- Options > Advanced > Export
- Export data – includes full text, you can pick the time period.
- Easy!
- …right?

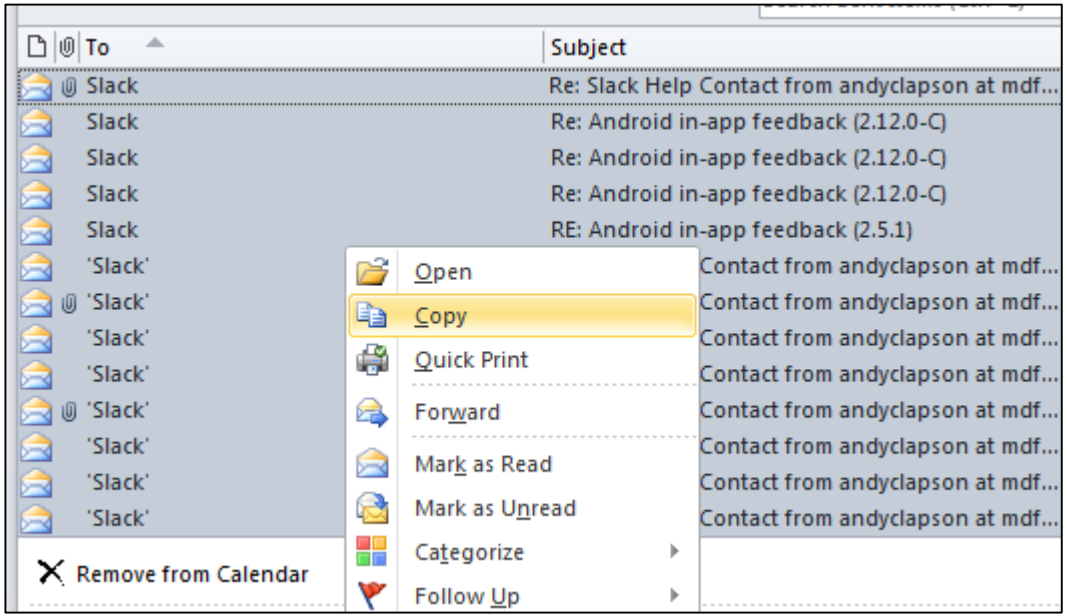


Oops….nope!


Windows
An error has occurred. To continue:
Press Enter to return to Windows, or
Press CTRL+ALT+DEL to restart your computer. If you do this,
you will lose any unsaved information in all open applications.
Error: 0E : 016F : BFF9B3D4
Press any key to continue _
FAILURE IS NOT AN OPTION
It comes bundled with every Microsoft product.
motifake.com

Extracting your data out of Outlook, round II

- Missing date and time of sending!
- Workaround: copy/paste from Preview pane
- Modify field names to include fields you want
  - (note you can include full text here!)
- *Calendar data export works great!



Email ID & Initial Data Cleanup

- Define message id (primary key)
- Then you can split the message text and email metadata
- Find ‘message type’ – new/RE:/FWD:
- Clean up Subject lines
- Ensure date/time in expected format, general integrity checks, size of email, etc.

```
ID = cats('MID_', put(_N_, Z8.));

----------------------------------

if substr(Subject,1,4)='RE: '
 then MessageType = 'Reply';

else if substr(Subject,1,4)='FW: '
 then MessageType = 'Forward';

else MessageType = 'Original';

----------------------------------

RegExId = prxparse('s/(RE|FW):
?//io');

call prxchange(RegExId, -1,
Subject, SubjectText);
```

Use Regular Expressions to format names/addresses

- We want a normalised data set!
- Generate a row for each addressee (To:/CC:/BCC:)

- More Regex! Yay!
  - Keep ‘full names’, if present (mostly internal people)
  - Flip around ‘last, first’ style names
  - Propcase()

```
i = 1; do while
((i <= (countw(To___Name_,';'))));
ToField =
dequote(scan((To___Name_),i,';');
output; i + 1; end;
 ---------------------------------
if find(toField, '@') AND
(find(toField, ',') OR
find(toField, '(')) then do;
RegExId = prxparse('s/\S*@\S*//');
call prxchange(RegExId, 1,
ToField, ToFullName);
end;
```

A whack at conversation ‘depth’

```
%*Now trying to define a 'conversation';
retain ConversationOrder InitialMessageID;
by SubjectText Sent MessageID;
if ((first.SubjectText = 1) OR (MessageType = 'Original')) AND (NOT
missing(SubjectText)) then do;
%*Populate the starting value depending on if I authored the Original;
    if MessageType = 'Original' then ConversationOrder = 0;
        else ConversationOrder = 1;
    InitialMessageID = MessageID;
end;
    else if NOT missing(SubjectText) then do;
        if first.Sent then ConversationOrder + 1;
    end;
        else ConversationOrder = 999;
ConversationID = catx('_', InitialMessageID, ConversationOrder);
```

Data prep is done!

- Final result is the data! Clean, formatted, normalized data with a lot of added value for analysis

	A	B	C	D	E	F	G	H	I
1	Sent	MessageID	MessageType	SubjectText	SizeInKb	ToFullName	ConversationOrder	InitialMessageID	ConversationID
078	2015-05-08	MID_005990	Original	OASUS	4	JOEL.ORR@STATCAN.GC.CA	0	MID_005990	MID_005990_0
079	2015-05-08	MID_000320	Forward	OASUS	3000	ANDY.CLAPSON@GMAIL.COM	1	MID_005990	MID_005990_1
080	2015-05-08	MID_005989	Reply	OASUS	1000	JOEL.ORR@STATCAN.GC.CA	1	MID_005990	MID_005990_1
081	2015-05-08	MID_005988	Reply	OASUS	14	JOEL.ORR@STATCAN.GC.CA	2	MID_005990	MID_005990_2
082	2015-08-31	MID_004737	Reply	OASUS	20	GREGORY.LUDWINSKI@STATCAN.GC	3	MID_005990	MID_005990_3
083	2015-09-29	MID_004736	Reply	OASUS Meeting Minutes	6	GREGORY.LUDWINSKI@STATCAN.GC	1	MID_004736	MID_004736_1
084	2015-10-22	MID_006931	Reply	OASUS Meeting minutes,	7	Gregory (Statcan/Statcan) Ludwinsk	1	MID_006931	MID_006931_1
085	2015-10-22	MID_006925	Reply	OASUS Meeting minutes,	8	Gregory (Statcan/Statcan) Ludwinsk	2	MID_006931	MID_006931_2
086	2015-10-22	MID_006924	Reply	OASUS Meeting minutes,	11	Gregory (Statcan/Statcan) Ludwinsk	3	MID_006931	MID_006931_3
087	2016-01-06	MID_006929	Reply	OASUS Presentation for t	11	Gregory (Statcan/Statcan) Ludwinsk	1	MID_006929	MID_006929_1
088	2016-01-07	MID_006928	Reply	OASUS Presentation for t	14	Gregory (Statcan/Statcan) Ludwinsk	2	MID_006929	MID_006929_2
089	2016-02-11	MID_006927	Reply	OASUS Presentation for t	20	Gregory (Statcan/Statcan) Ludwinsk	3	MID_006929	MID_006929_3
090	2016-02-22	MID_006926	Reply	OASUS Presentation for t	24	Gregory (Statcan/Statcan) Ludwinsk	4	MID_006929	MID_006929_4
091	2015-09-09	MID_004740	Reply	OASUS Presentation?	17	GREGORY.LUDWINSKI@STATCAN.GC	1	MID_004740	MID_004740_1

So now what?

- A script to generate pretty nifty ‘email analytics’ dataset
- But…more metadata than text analytics, right?
- The EMAILS themselves! Heavier duty text processing (not to mention the ‘reply problem’)
- Using message ID, we can concord the text to its parent email
- The holy grail? Both sides of the email chain – network/cluster diagrams