Text Mining
Finding Themes in Text Data

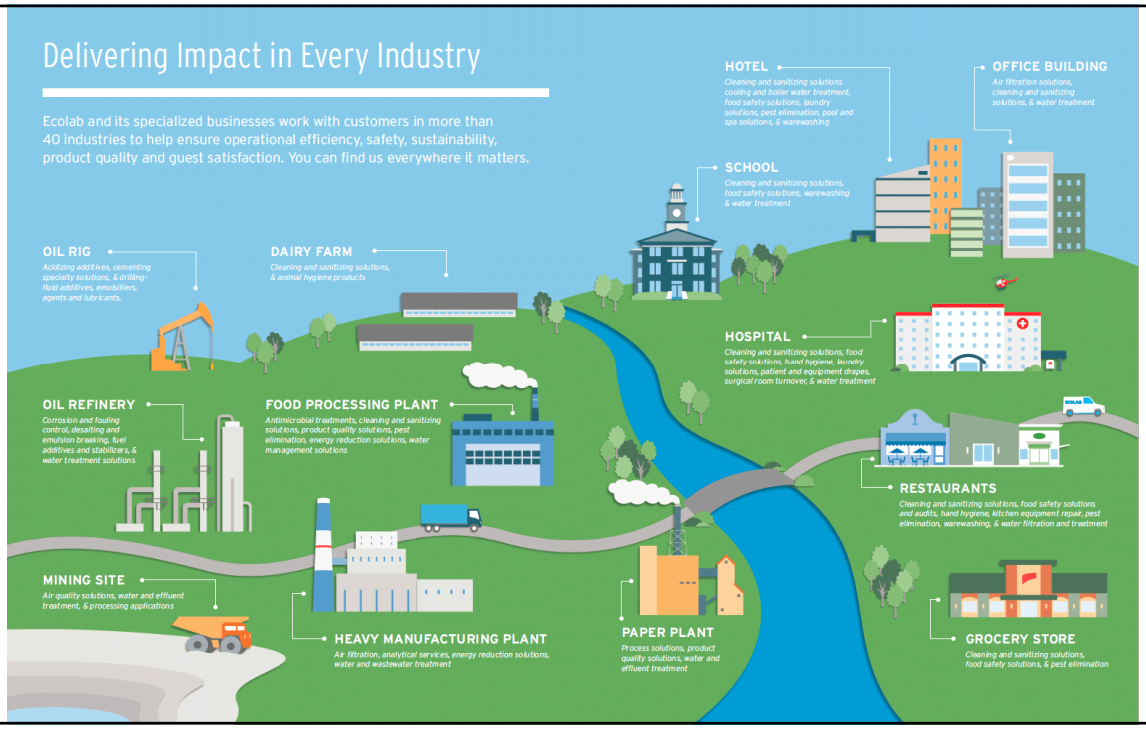
Paul Prew, Ecolab Statistician

ECOLAB

Ecolab

CLEAN WATER
SAFE FOOD
ABUNDANT ENERGY
HEALTHY ENVIRONMENTS

- By 2030, it is estimated that the world will need
  - 30% MORE ENERGY
  - 40% MORE WATER
  - 35% MORE FOOD
- Ecolab, Inc. – St. Paul, MN
  - Chemicals manufacturer
  - $15 billion sales 2018
  - 48,000 associates globally – 26,000 field service
  - R&D: 1,600 scientists, engineers and technical specialists
  - 19 Technology Centers – Eagan, MN primary campus



Feedback from the field

- Service Requests - field associates log their actions.
  - Each record has a section for 'Action_Comments'.
- These logs are stored electronically
  - 'free form' text data could not be analyzed systematically

13,000 service records for dishwasher dispensers

1	ACTION_COMMENT
2	the chemical dispensing pumps that come with the Glastender machine are poor at best and in this case stopped working- I replaced the defective pumps with Ecolab dispensing equipment for better chemical dosing
3	I replaced the injection fitting for detergent.
4	Limescale build up. Customer to order limeaway. I will return to delime.
5	Copper fitting were loose and water was leaking from water T. Tightened up comp fittings no further issues.
6	The detergent was off in its titration. I changed the detergent cell and it is functioning properly
7	manual detergent
8	Lowered Rinse Aid
9	replaced cover and hose
10	sanitizer dispenser water line had a small hole, fixed. manual pot and pan detergent was leaking from fitting, replaced.
11	The detergent injection fitting snapped off inside of the machine. I put a new injection fitting on the machine and it is as good

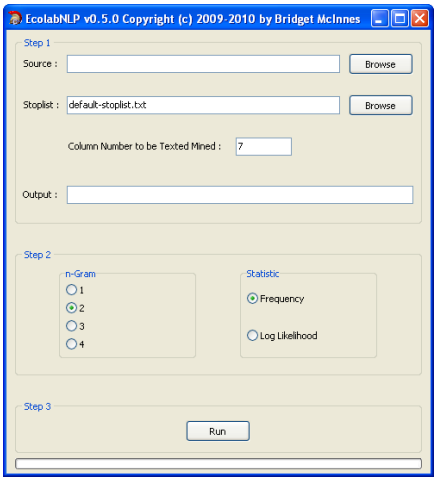
- What's in there?

Text mining

- Discovering useful patterns from collections of text records
- Big data – machine learning; algorithms
- Unsupervised learning – descriptive
  - Supervised learning - predictive: Y = f(X)
  - Unsupervised learning -

N-grams

- Dr. Bridget Thomson McInnes, formerly University of Minnesota, is the author of Ecolab NLP.
- The software performs Natural Language Processing (NLP)
- More specifically, N-gram Language Modeling
  - "An n-gram is a sub-sequence of n items from a given sequence. The items in question can be … letters, **words** ..."
    n.wikipedia.org/wiki/N-gram
  - Sequence: sentence
  - Sub-sequence: e.g. 2-gram = 2 adjacent words
- Ecolab NLP - display the most frequent n-grams



"Oasis 299 giving problems to several customers this past month. "

- Nine 2-grams
  - "**Oasis 299** giving problems to several customers this past month. "
  - "Oasis **299 giving** problems to several customers this past month. "
  - "Oasis 299 **giving problems** to several customers this past month. "
  - **…**
  - "Oasis 299 giving problems to several customers this **past month**. "
- Eight 3-grams
  - "**Oasis 299 giving** problems to several customers this past month. "
  - "Oasis **299 giving problems** to several customers this past month. "
  - **…**
  - "Oasis 299 giving problems to several customers **this past month**. "
- Seven 4-grams
  - "**Oasis 299 giving problems** to several customers this past month. "
  - **…**
  - "Oasis 299 giving problems to several **customers this past month**. "

Ignoring non-informative words "Stop-Word Lists"

- Some words are predictably uninteresting
  - "the dishmachine" vs. "dishmachine"
  - "the" - no additional information
  - Words that appear frequently in any context.
    - *Is, The,* and *A* are such words.
- How to keep them from dominating the counts?
  - Remove them via **stop-word list**
  - Sample from Ecolab NLP stop-word list

… their theirs them themselves then there the third thirteen this those though three…

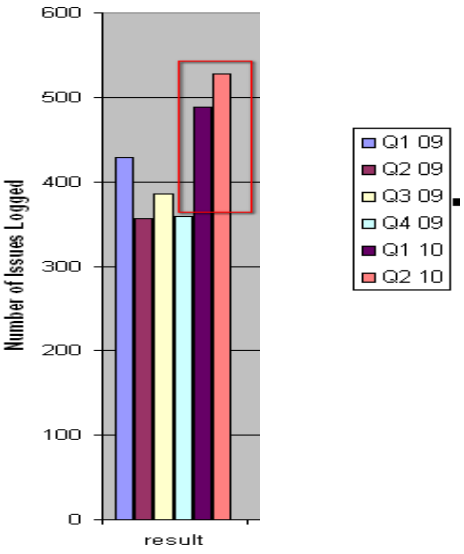
Using stop-word lists

- Stop-word list
  - giving
  - to
  - this
- Before: “Oasis 299 giving problems to several customers this past month. “
  - Nine 2-grams
- After: Oasis 299 problems several customers past month
  - Six 2-grams

  Oasis 299
  299 problems
  problems several
  several customers
  customers past
  past month

Case Study
Hotel Laundry service calls

- Q1 2010
  - Spike in ‘Results’ category
  - ‘Results’ is a catchall
- Spike is because…?
  - Technical Services investigated trends using Excel
  - Tried likely ‘smoking gun’ keywords; ineffective



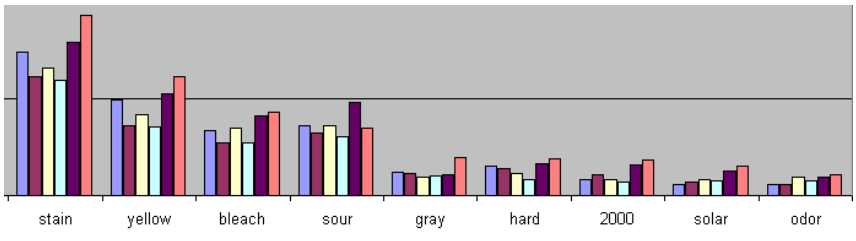
Laundry CATSWeb

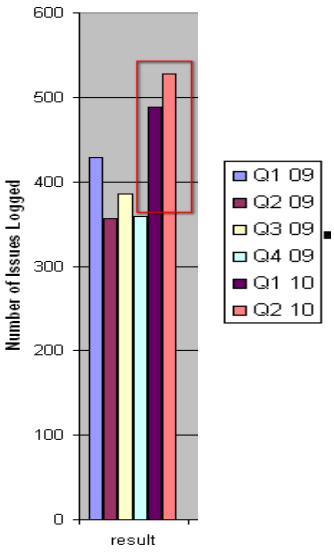
- 2,678 calls for 'Results' Jan 2009 - June 2010
- "Describe the Issue" free text
- 97,800 words 109 pages of text

Issue #	Employee or Customer #	Hire Date	Market Area	Describe the Issue	Creation Date
269988	27570	2/24/2003	Laundry - OPL - HHC	Large commercial laundry in No Miami. Soot on white linen from hot water supply. Situation permanent.	1/5/09 9:03 AM
270036	18994	4/1/1993	Laundry - OPL - HHC	Kitchen rags issues, formula questions	1/5/09 10:12 AM
270153	12949	10/31/1994	Laundry - OPL - HHC	Grease spots in linen, possibly from the machine, how best to remove. Stain Blaster S, have the customer get that washer looked at right away also.	1/5/09 12:24 PM
270581	38577	2/25/2008	Laundry - TCD	is surface rust on edge of dryer basket - he's using Injection Sour (unknown final ph) on new Milnor CBW & dryers check steam condensor, may also need to use conditioner to tie up iron (can't use iron controlling sour like Turbo Lizer in CBW)	1/6/09 9:43 AM
270593	34953	8/7/2006	Laundry - OPL - HHC	X-Static blocks are crumbling found that block was not installed on bottom of fin, temp is unknown to TM	1/6/09 9:53 AM
270602	30542	11/10/2005	Laundry - TCD	Bar towel formula products and formula setup	1/6/09 10:07 AM

Ecolab NLP 1-gram search

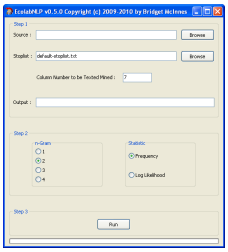
- "stain", "yellow", "bleach"





Most Frequent 2-grams

Power Pak	336
Stain Blaster	176
overall discoloration	157
Clearly Soft	156
2000 XP	124
Solar Brite	117
table linen	90



The software also collects the individual records

Issue #	**Power Pak**
270604	OVERALL DISCOLORATION. Formula 1 operation temperature? Installed F1 a mc
270700	Getting yellow robes - brand new robes. TM not at account. Advised most likely
271273	STAINS. Customer who does plastic surgery has greasy stains from silicone impl
271292	yellow spots on linen and Power Pak 1 is putting holes in linen - fixed timer - mi
271495	Getting yellow sheets. Sheets are relatively new (< three weeks), showing a crea
271651	STAINS. Account accidentally switched Builder C and Detergent I leaving deterg

Most Frequent 3-grams

Power Pak reclaim	54
2000 XP Destainer	40
Solar Brite Destainer	35
Frequent caller interruptions	35
Destainer Clearly Soft	28
pH iron chlorine	23
Bad cell connection	22
Reclaim Power Pak	18
reclaim Power Pak	18

Higher n-grams ➔ fewer records

N-gram = 1	Freq
linen	551
not	516
water	479
stains	435
account	393
bleach	377
yellow	363

N-gram = 2	Freq
Power Pak	186
overall discoloration	157
Clearly Soft	130
Stain Blaster	106
2000 XP	96
Solar Brite	88
yellow stains	80

N-gram = 3	Freq
2000 XP Destainer	36
Solar Brite Destainer	31
Destainer Clearly Soft	24
Frequent caller interruptions	20
Power Pak reclaim	19
Sour Clearly Soft	16
Discussed possible issues	15

N-gram = 4	Freq
Brite Destainer Clearly Soft	10
Solar Brite Destainer Clearly	10
Destainer Neutralizer Clearly Soft	10
Solar Brite Oxy Brite	10
home style top loading	9
XP Destainer Clearly Soft	8
2000 XP Destainer Clearly	8

Use software iteratively

- Subset of 157 records with 2-gram “Overall Discoloration”
- Find 2-grams within that subset

Power Pak	44
Solar Brite	40
Clearly Soft	29
2000 XP	18
No odor	16
not new	16
water hardness	16
iron chlorine	14

Examine the individual records

Issue #	iron chlorine
278650	OVERALL DISCOLORATION. Towels turned overall yellow over the last two days. Not new towe
281854	OVERALL DISCOLORATION. Industrial laundry doing white shirts. Shirts stay white until leavi
283886	OVERALL DISCOLORATION. Account switched from L-2000 XP to Solar Brite NP. Customer sinc
291137	OVERALL DISCOLORATION. Renaissance Hotel washed 65C/35P robes that turned yellow upo
293490	OVERALL DISCOLORATION. White bath towels developing an overall yellow discoloration. Gc

Larger example – Restaurant Kitchens

- 43,000 records from servicing restaurant dish machines
  - Chemistry dispensers
- JMP software
- N-gram = 'Phrase'
- Pre-processing: stemming
  - permutations of the same word
  - strip back to their 'stems'.
  - place a dot at the end of the stem.

Stem	Terms
machin·	machine,machine's,machined,machines

ACTION_COMMENT
Replaced squeeze tube in sanitizer dispenser. ...
fixed detergent and sanitizer leak
Unclogged metering tip on Quat 146 dispenser.
found pump bad on sanitizer and replaced. ...
Presoak dispenser leaking from the pex line. Cut ...
Unclog Apex detergent dispenser
quat dispenser broken order a new dispenser

Term and Phrase Lists - ACTION_COMMENT for Kitchen/Product Dispenser

Term	Count
dispens·	26291
replac·	22722
deterg·	10234
line·	8599
leak·	7972
sanit·	7847
rins·	7362
water·	6104
machin·	5768
tube·	5031
new·	4993

Phrase	Count	N
detergent dispenser	2576	2
squeeze tube	1835	2
water line	1724	2
rinse dispenser	1602	2
metering tip	1452	2
working properly	1412	2
dish machine	1234	2
wash max	1120	2
sanitizer dispenser	983	2
replaced detergent	878	2
quat sanitizer	828	2

Term-Document Matrix

ACTION_COMMENT

Doc_1: *Replaced cracked flush manifold*

Doc_2: *Replaced the 2026 board*

Doc_3: *Replaced fill coil and cleaned inside of solenoid valve*

Doc_4: *Chlorine Sanitizer too strong going into the dishmachine*

*stop words removed

Term-Document Matrix

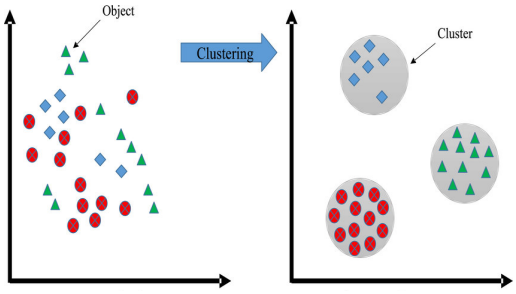
➔ Enables document clustering

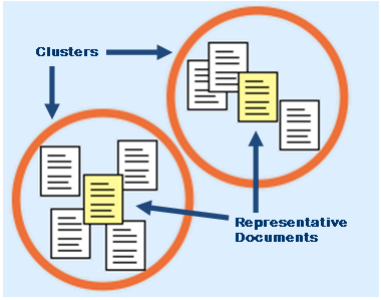
➔ Enables theme discovery

Term	Doc_1	Doc_2	Doc_3	Doc_4
replaced	1	1	1	0
2026	0	1	0	0
board	0	1	0	0
chlorine	0	0	0	1
cleaned	0	0	1	0
coil	0	0	1	0
cracked	1	0	0	0
dishmachine	0	0	0	1
fill	0	0	1	0
flush	1	0	0	0
going	0	0	0	1
inside	0	0	1	0
manifold	1	0	0	0
sanitizer	0	0	0	1
solenoid	0	0	1	0
strong	0	0	0	1
valve	0	0	1	0

Clustering

- Most-used unsupervised learning algorithm
- Assumption -- *themes* drive the creation of comments
- Cluster documents share-
  - the most words with documents in same cluster
  - the fewest words with documents in other clusters
- Shared words – interpret the theme





Themes from Clusters

- Cluster_1 = 'junk' cluster – affinity scores are weak.
  - Every doc forced into a cluster.
  - Contains 1/3 of the doc's - about 15,000 of the documents didn't have a lot of words in common.
- Cluster_2 has a theme about replacing lines, fixing leaks -- leaking lines?
  - Cluster_2 has ¼ of the doc's: almost 11,000.
  - Not all, or even most, will discuss a 'line' or a 'leak'; there are other connections in Cluster_2's aggregate bag of words that draw in 11,000 doc's.

Cluster1		Cluster2		Cluster3		Cluster4		Cluster5	
Term	Score	Term	Score	Term	Score	Term	Score	Term	Score
warewash·	0.0257	replac·	6.5618	dispens·	9.6018	deterg·	5.0923	dispens·	6.6348
details	0.0251	line·	4.4366	replac·	7.392	adjust·	2.7005	machin·	4.8767
comment·	0.0093	leak·	3.9688	tip·	2.1318	board·	2.0146	water·	4.0093

Proportion of Docs

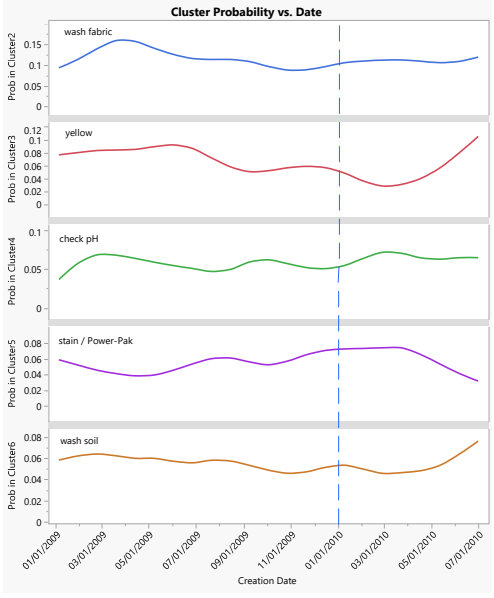
Cluster1	0.32
Cluster2	0.25
Cluster3	0.19
Cluster4	0.12
Cluster5	0.11

Clustering – Finding themes

- How the clustering algorithm assigns every doc to 1 of 5 clusters
  - Randomly assign an equal number of the 43,000 doc's to each cluster. So 5 clusters, each with 8,600 random doc's.
  - The comments are treated as a "bag of words". Ignore word sequence.
  - doc_A randomly assigned to Cluster_1.
    - Affinity score from shared words
  - doc_A gives consideration to Cluster_2
    - Affinity score from shared words
  - doc_A gives consideration to Clusters_3,4,5
    - Affinity score from shared words
  - Highest affinity score claims doc_A.
    - Repeat for every doc in every cluster.
  - Doc/cluster swapping ends when every doc is in the Cluster that maximizes its affinity score

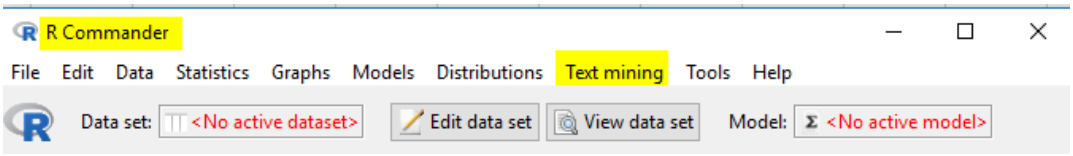
Time Trends in Clusters

- Re-examine Laundry Service Calls
- Create text clusters
- Plot cluster trends by month
- No big shift around Q1 2010
  - Original 1-gram analysis was more helpful than clustering



Open Source Text Mining Tools

- R Commander – GUI interface for R statistical language
- Package "RcmdrPlugin.temis"



Conclusion

- **Text mining**
  - Analyze large groups of documents
  - Discover themes, relationships
  - Unsupervised learning