Automated text mining, cluster analysis, and multi-dimensional scaling: Which is the best?

Chong Ho Yu, Ph.D., D. Phil.

(2020, May 23)

International Data Engineering and Science Association

What is text mining?

- Also known as text analytic.
- A process of extracting useful information from document collections through the identification and exploration of interesting patterns (Feldman & Sanger, 2007).
- The ideal tool for tapping into under-utilized, unstructured data.

The forerunners of TM

0011

- TM is not entirely new.
- Qualitative researchers have been doing content analysis and grounded theory by hand.
- Yu, C. H. & Marcus-Mendoza, S. (1993). Attitudes of correctional staff. In B. R. Fletcher, L. D. Shaver, & D. G. Moon (Eds.), *Women prisoners: A forgotten population* (pp.111-118). Westport, Connecticut: Praeger.
- Yu, C. H., Jannasch-Pennell, A., & DiGangi, S. (2011). Compatibility between text mining and qualitative research in the perspectives of grounded theory, content analysis, and reliability. *Qualitative Report*, *16*, 730-744. [link]

Hand-coding

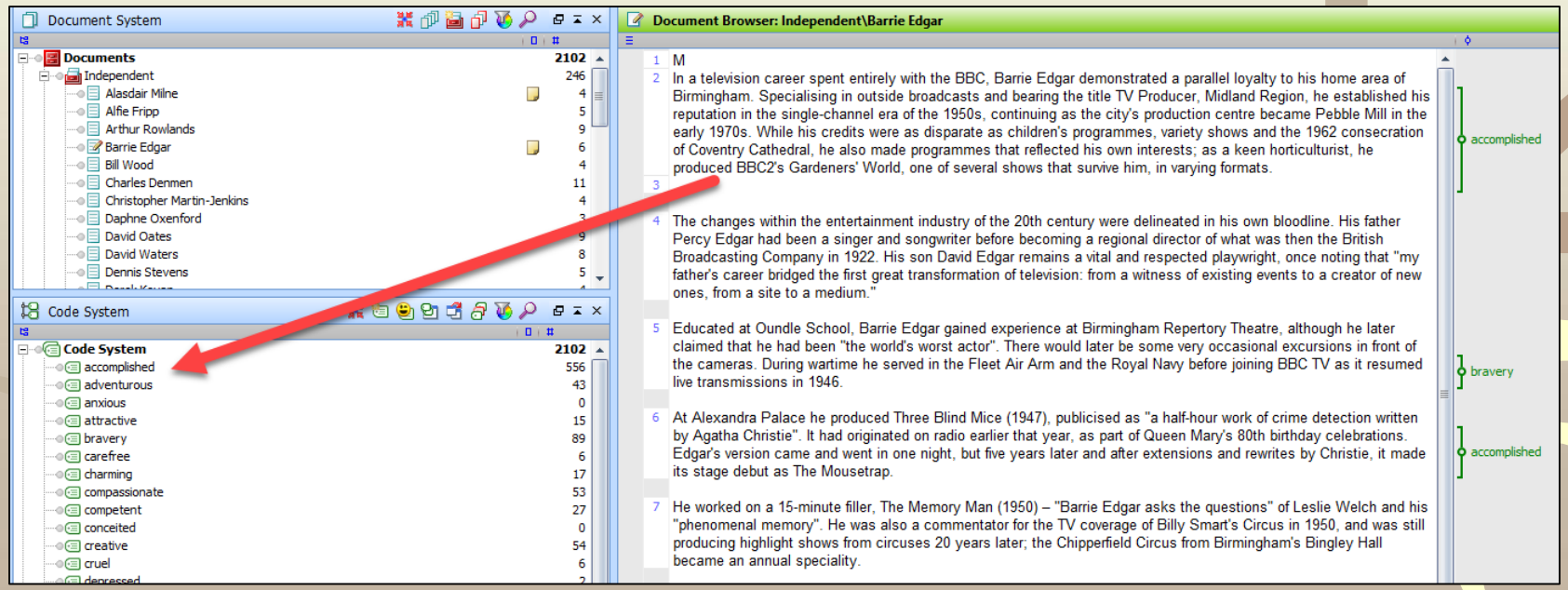
0011

Achievement-relatedness Social skills Subjective well-being Kindness/morality

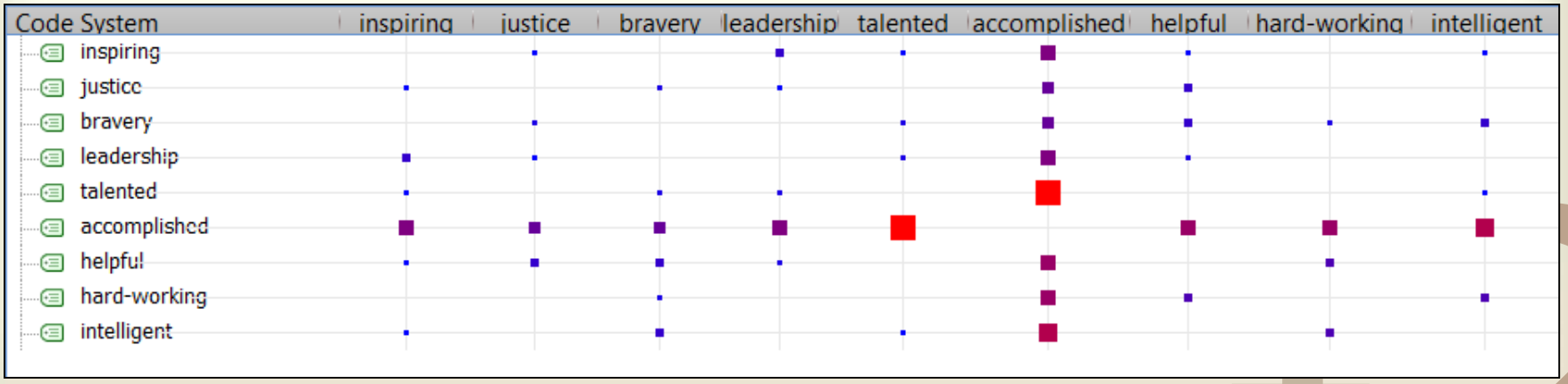
1	Charming	20	Hard-working
2	Sorrow	21	Happy with their lives
3	Kind	22	Good-looking
4	Creative	23	Ethical
5	Good sense of humor	24	Competent
6	Hypocritical	25	Efficient
7	Friendly	26	Conceited
8	Trustworthy	27	Moody
9	Wise	28	Knowledgeable
10	Phony	29	Dishonest
11	Sad	30	Likable
12	Attractive	31	Depressed
13	Shy	32	helpful
14	Anxious	33	Easy to get along with
15	Cruel	34	Selfish
16	Fun to work with	35	Loving
17	Intelligent	36	Accomplished
18	Happy with themselves	37	Psychologically healthy
19	Snobby	38	Talented

MAXQDA

- Human coders:
  - More accurate than machine coding
  - Subject to fatigue and bias
  - Take forever with big data
- Classify the passage by dragging and dropping



Code relation chart in MAXQDA



- Symmetrical data matrix; R1=C1, R2 = C2
- The frequency of co-occurrence is depicted by the size and the color of the square.

The forerunners of TM

- Hypothesis generation by Swanson process.
- Based on the idea of concept linking, Swanson (1986) manually scrutinized the medical literature and identified relationships between some apparently unrelated events, namely, consumption of fish oils, reduction in blood viscosity, and Raynaud's disease.

Hypothesis generation

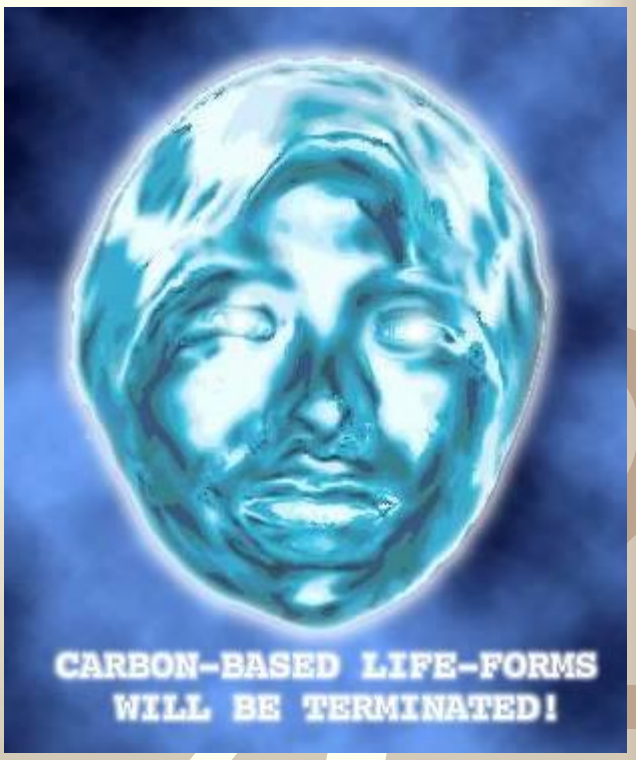
- His hypothesis that there was a connection between the consumption of fish oils and the effects of Raynaud's syndrome was eventually validated by experimental studies (DiGiacomo., Kremer, & Shah, 1989).
- Using the same methodology, the links between stress, migraines, and magnesium were also postulated and verified.

Don R. Swanson

- Swanson had no formal training in the biomedical field!
- He received a BS in Physics at Caltech, a Ph.D. in Theoretical Physics from UC Berkeley.
- Worked as a physicist at various lab.
- Become a professor and Dean of Graduate School of Library Science at the U. of Chicago.
- The key of his success is concept linking.

Artificial intelligence

- TM enhances grounded theory and concept linking by automation.
- TM utilizes the technology of natural language processing, a subfield of artificial intelligence (AI) & computational linguistics.
- Why do we need natural language processing in data mining?
- The software app must be smart enough to understand the context.



Challenges to NLP

- In some tricky situations the AI system could be fooled.
  - *Incorrect stop-word removal*
  - *Sarcastic expression*
  - *Negation*
  - *Valence-shifting*
  - *Irrealis*

Incorrect stop-word removal

- Remove trivial words, such as “a,” “an,” “the,” “is,” “am,” “are,” “however,” “although,” “but,” etc.,
- This process necessitates AI, otherwise, on some occasions important information might be lost.
  - “To be or not to be”
  - “If it is to be, it is up to me.”
  - “Brazil traveler to USA need a visa.”

Solution: BERT

0011

- BERT (Bidirectional Encoder Representations from Transformers) was invented by Google researchers of AI Language in 2018-2019.
- The traditional AI system “read” the sentence sequentially (e.g. from left to right).
- In BERT the reading is bi-directional (read the whole sentence at once) in order to determine the role of surrounding words (left and right), including common words, in the sentence.
- Transformer, developed by Google in 2017, is a mechanism that learns the contextual relations between words.

BERT

- In the sentence “2019 Brazil traveler to USA need a visa,” how the word “to” and other words related is important for understanding the meaning.
- Older NLP misinterpreted the first sentence as the visa requirement of US citizens traveling to Brazil.

As of June 16, **2019**, U.S. **citizens** do not **need a visa** if they are **traveling** to **Brazil** for tourism, business, transit, artistic or sport activities, with no intention of establishing residence. Jan 14, 2020

travel.state.gov › content › international-travel › Brazil

Brazil International Travel Information

About Featured Snippets Feedback

BERT

- With BERT Google NLP is able to learn that the very common word “to” matters a lot.

br.usembassy.gov › Visas

Visa Waiver Program | U.S. Embassy & Consulates in Brazil

... President Jair Bolsonaro of **Brazil** to the White House on Tuesday, March 19, **2019**. ... You plan to stay in the **United States** for 90 days or less. You are ... Your **passport** complies with all **Visa** Waiver Program requirements. ... If you meet all the requirements for the **Visa** Waiver Program, you do NOT **need** to apply for a **visa**.

People also search for

visa fees brazil	non immigrant visa brazil
visa costs for u.s. citizens	brazil tourist visa application
tourism visa usa	visa application brazil

nathanieltower.com › 2019-brazil-traveler-to-usa-need-...

2019 Brazil traveler to USA need a visa - Nathaniel Tower

In most cases, a **2019 Brazil traveler** heading to the **USA** will **need** to have a **visa**. According to the U.S. Embassy & Consulates in **Brazil** website: “In general, **tourists traveling** to the **United States require** valid B-2 **visas**.

Sarcasm

0011

- When the writer made sarcastic expressions, it could fool a regular text mining software package
- Consider this passage: “The professor is very great! I didn’t study at all. I closed my eyes throughout the whole semester and still got an A!”

MOST HELPFUL RATING

Jan 6th, 2017

Second easiest class I've ever taken in my life - only second to Kindergarten coloring.

4 0

Sarcastic or serious?

0011

“And then I see the disinfectant, where it knocks it out in a minute. One minute. And is there a way we can do something like that, by injection inside or almost a cleaning. Because you see it gets in the lungs and it does a tremendous number on the lungs. So it would be interesting to check that. So, that, you're going to have to use medical doctors with. But it sounds -- it sounds interesting to me.”

Solution: CNN

- Convolutional Neural Networks (CNN)
- Model local features to learn global features
- Detecting contradictions
  - "I love the pain of breakup."
  - "I work so hard to be so poor."
  - "Truly, you have a dizzying intellect. (Wesley, The Princess Bride)

Negation

0011

- The positive polarity is reversed by a negative word.
- "No one thinks it is good." In this case, although "good" is a positive word, the phrase "no one" alters its connotation.
- "Parking on a hill with no curb."

Solution: BERT

0011

- In the past Google ignore “no” but now…

For either **uphill** or downhill **parking**, if there is **no curb**, turn the wheels toward the side of the road so the car will roll away from the center of the road if the brakes fail. When you park on a sloping driveway, turn the wheels so that the car will not roll into the street if the brakes fail.

pages.cs.wisc.edu › ~gdguo › driving › HillParking

Parking on a Hill

About Featured Snippets Feedback

Value-shifting

- In some cases a single word in a sentence shifts the meaning.
- This is a *missed* opportunity”
- “The medicine *kills* cancer cells”

Irrealis

0011

- When a conditional sentence implies a counterfactual scenario, it might be difficult to tell whether it is positive or negative
- “It would be better if the Wi-Fi network is faster.” The student might be satisfied with the existing connection speed but he is looking for more bandwidth. It might also be the case that he could not stand the slowness of the current Wi-Fi network.

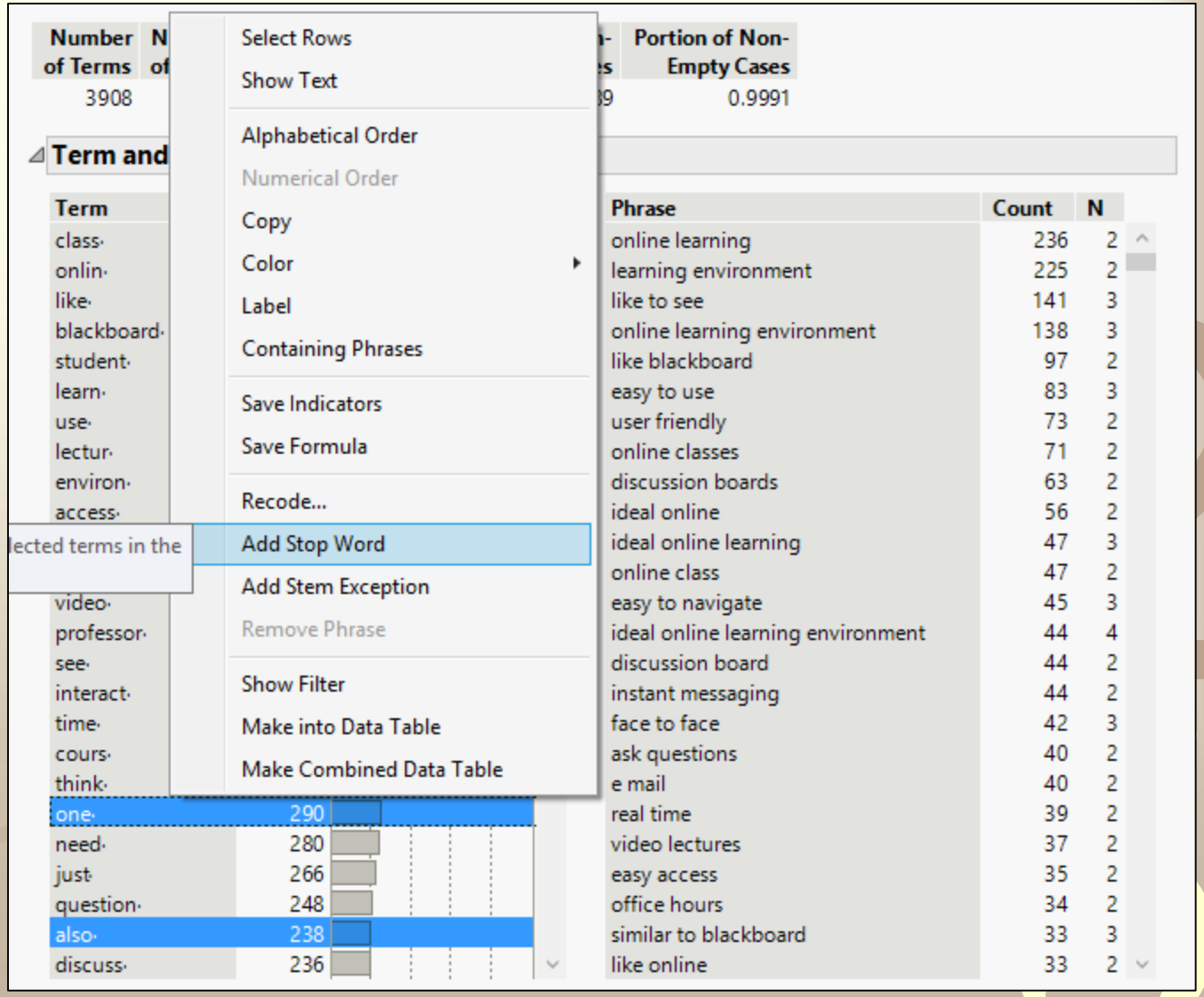
Example 1

- The data source, which encompasses responses to an open-ended survey item collected from a US Southwestern university, was used for extracting common threads.
- “If you had the ability to design your ideal online learning environment--What would you like to see? How would it look and feel? What features would it have?”
- Effective sample size: 3,193

Text Explorer in SAS/JMP

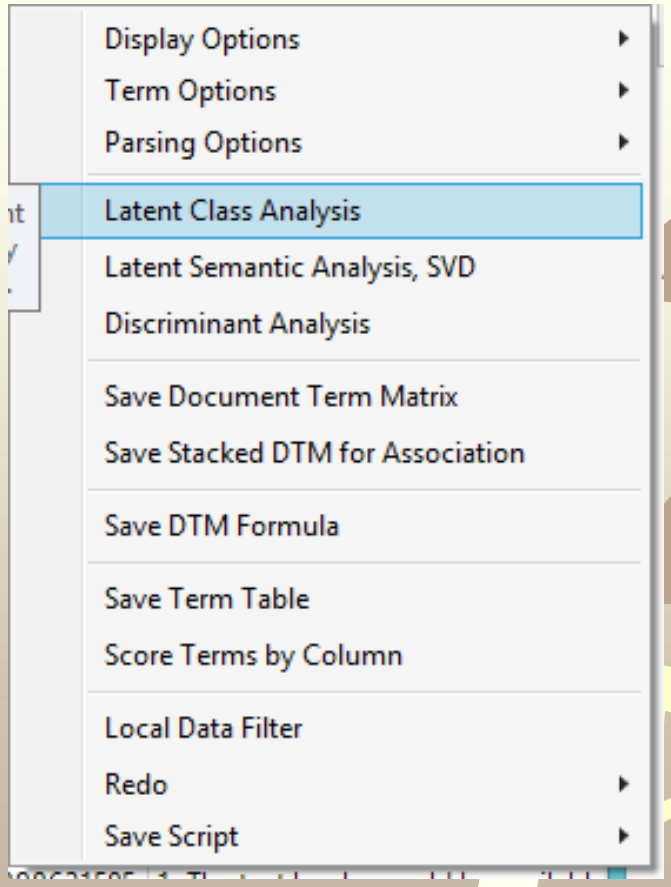
0011

- Phrases are good.
- Terms: Stop-word removal isn't good.



Latent class analysis

- LCA is a form of cluster analysis for textual data.
- The goal is to group similar concepts (or terms) together.



Latent class analysis

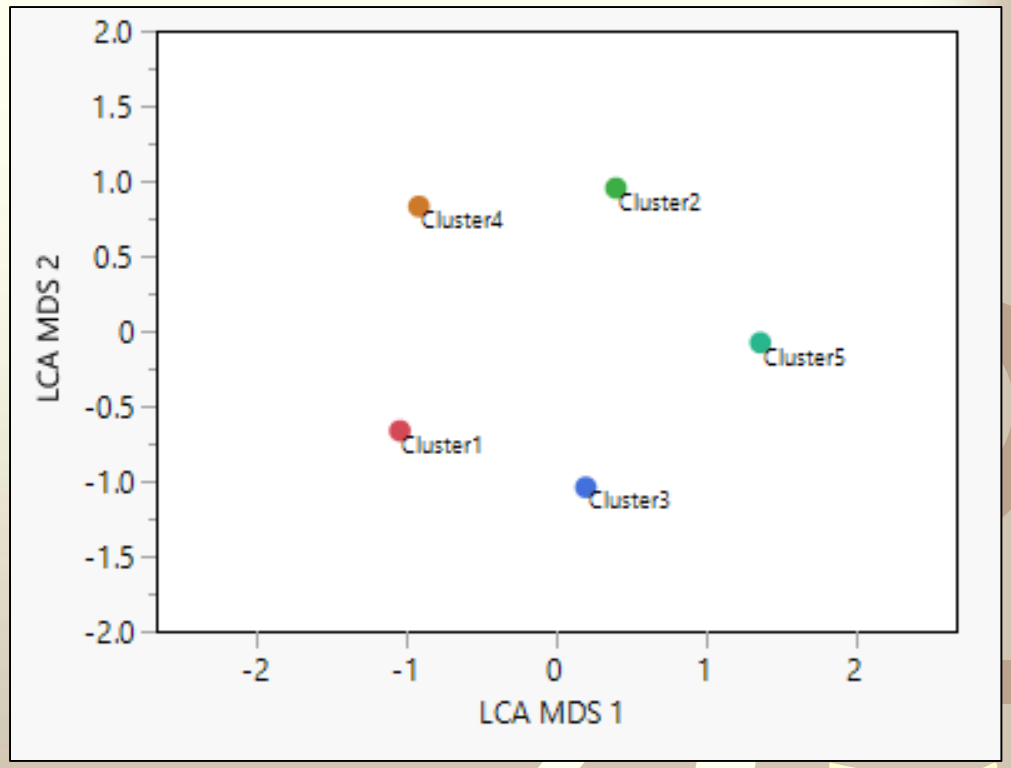
0011

Top Terms by Cluster

Cluster1		Cluster2		Cluster3		Cluster4		Cluster5	
Term	Score	Term	Score	Term	Score	Term	Score	Term	Score
fine	0.4915	video	4.1253	onlin	6.2192	easi	3.0887	class	8.7006
sure	0.4654	lectur	4.0287	learn	5.7406	blackboard	2.6889	student	7.9773
now	0.4382	professor	1.4118	class	4.6687	link	2.3457	use	7.3401
satisfi	0.2772	chat	1.3878	environ	3.1706	like	2.1312	onlin	6.6414
pretti	0.2359	note	1.2676	student	3.0904	page	2.0355	access	5.2855
happi	0.1944	audio	1.2519	interact	2.6731	look	1.6829	like	5.1466
sorri	0.1233	studi	0.6579	classroom	2.4225	simpl	1.4309	blackboard	4.9729
fan	0.119	board	0.6575	think	2.3127	system	1.3484	also	4.6245
idea	0.091	live	0.6573	like	2.2624	someth	1.1743	cours	4.5975
webwork	0.0783	communic	0.6504	comput	1.7612	email	1.117	assign	4.412

LCA and MDS

- Cluster analysis and Multi-dimensional scaling (MDS) can work hand in hand.
- The graph shows five distinct clusters of terms.



Latent Class Analysis

Using 715 terms across 3192 do

BIC 325222.5 Show Text

Cluster Mixture Proba

Cluster	Mixture Probability
Cluster1	0.36944
Cluster2	0.21141
Cluster3	0.17376
Cluster4	0.16942
Cluster5	0.07597

Term Probabilities by

Term	Cluster Most Characteristic
class	Cluster5
onlin	Cluster4
like	Cluster5
blackboard	Cluster5
student	Cluster5
learn	Cluster4
use	Cluster5
lectur	Cluster5
environ	Cluster5
access	Cluster5
teacher	Cluster5
easi	Cluster3
video	Cluster2
professor	Cluster5
see	Cluster5
interact	Cluster4
time	Cluster5
cours	Cluster5
think	Cluster5
one	Cluster5
need	Cluster5
just	Cluster5
question	Cluster5
also	Cluster5
discuss	Cluster5

Top Terms by Cluster

Cluster1	Cluster

Context for Selected Rows - JMP

File Edit Tables DOE Analyze Graph Six Sigma Tools Tools Add-Ins View Window Help

ideal

I'm not sold on online learning. Perhaps I'm too much of an old codger. [4]

Don't know. [7]

- user friendly - uniform from class to class [12]

A BBS that everyone in ASU can use it [30]

a better GUI for links in each subject on that subjects page. [32]

A better use of technology overall. Wi-fi, internet classes, etc. [35]

A choice between a degree program offered fully on-line, or classroom. [38]

A completly 3D desktop and learning institution. One that you could move around in and interact with the
objects in the world. [42]

A laptop environment would be nice [45]

A little more visually stimulating than blackboard is. Blackboard is very asthetically mundane. [49]

a lot of resources, acess to instructor, virtual classroom,user-friendly layout. [53]

A more intuitive GUI. [55]

A more organized version of blackboard. [56]

A nice start is what I said in the prior question. [58]

A system that could continually be upgraded. Don't settle on one system that is limited, have them create
a system that can continually be upgraded and adapt to newer technology. [72]

A University-wide free research system like Questia would be great. [78]

a user freindly one [79]

A very straightforward system would be good that is very easily accessable. [81]

Latent Class Analysis

Using 715 terms across 3192 d

BIC 325222.5 Show Text

Cluster Mixture Proba

Cluster	Mixture Probability
Cluster1	0.36944
Cluster2	0.21141
Cluster3	0.17376
Cluster4	0.16942
Cluster5	0.07597

Term Probabilities by

Term	Cluster Most Characteristi
class	Cluster5
onlin	Cluster4
like	Cluster5
blackboard	Cluster5
student	Cluster5
learn	Cluster4
use	Cluster5
lectur	Cluster5
environ	Cluster5
access	Cluster5
teacher	Cluster5
easi	Cluster3
video	Cluster2
professor	Cluster5
see	Cluster5
interact	Cluster4
time	Cluster5
cours	Cluster5
think	Cluster5
one	Cluster5
need	Cluster5
just	Cluster5
question	Cluster5
also	Cluster5
discuss	Cluster5

Top Terms by Cluster

Cluster1	Cluste

Context for Selected Rows 2 - JMP

File Edit Tables DOE Analyze Graph Six Sigma Tools Tools Add-Ins View Window Help

ideal

A combination of secondlife and wiki style information sharing and gathering. Ability to access transcripts
or recordings of lectures would be most desirable. [40]

A comfortable spot with laptops to do studying. Not like the computer lab or libray so where more cozy. [41]

a forum on the class site to talk about lecture and assignments [44]

A list of assignments including the assignment requirements as well as due dates; class lists and contact
info.; links to lecture slides. [46]

A list of homework items, along with important dates. A section for classroom notes, and a discussion area to have with other classmates on problems or questions. [47]

A lot like Microsoft Groove, where the downloads and updates just show up! [51]

A nice, available forum for class discussion that is optional, not mandatory. [59]

A professor would use it to give video lectures. Every individual would be connected live, with audio /visual capability. [63]

a quality video feed of the teacher, Quality voice feed, a seprate window for notes, No LAG.... [64]

A recording of the lectures. [65]

A schedule of class assignments and deadlines at a glance (all on one visual calendar). [66]

A small environment, a computer for every student. A small classroom. [69]

A small window to see the teacher, an open chat box to ask questions, and a powerpoint presentation. [70]

A supplemental tool for class- has lectures/powerpoints posted, links to external sites that are helpful,
study guides, examples, rubrics, stuff like that. [71]

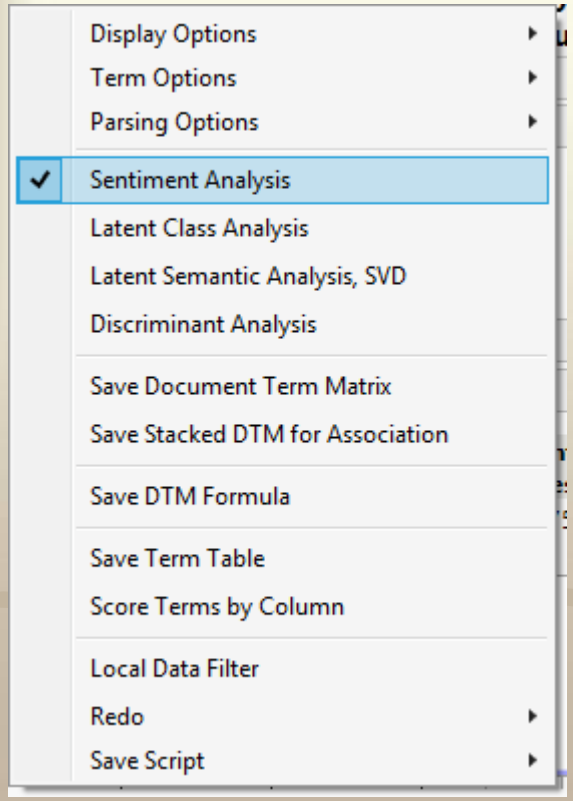
a teacher lecturing on webcam with students signed in to the chatroom/classroom [74]

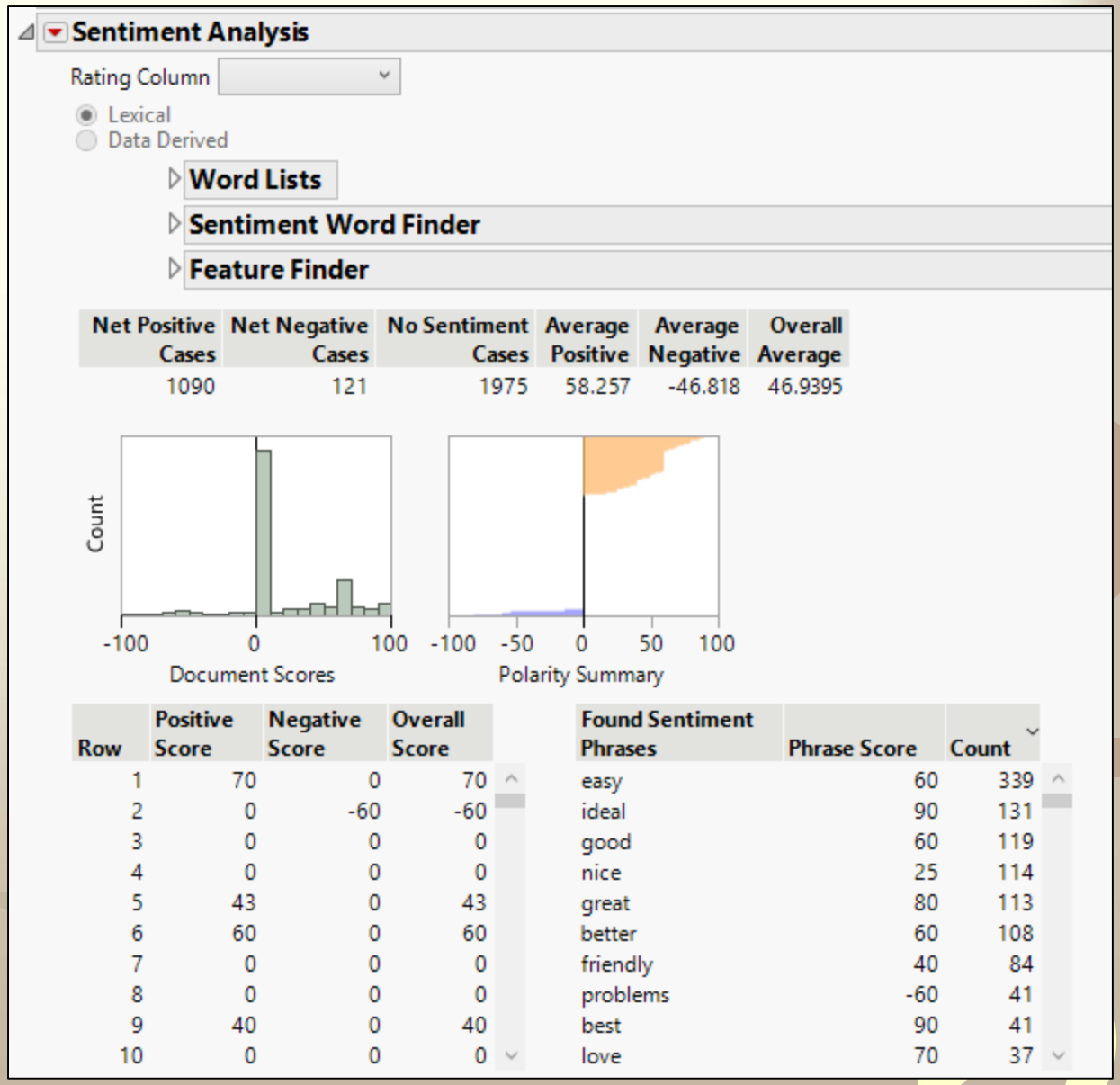
Name the cluster

- Concept 1: User friendliness or user interface
- Concept 2: Accessibility to course-related resources
- Concept 3, 4, 5…etc.

Sentiment analysis

0011





Net Positive Cases	Net Negative Cases	No Sentiment Cases	Average Positive	Average Negative	Overall Average
1090	121	1975	58.257	-46.818	46.9395

Row	Positive Score	Negative Score	Overall Score
1	70	0	70
2	0	-60	-60
3	0	0	0
4	0	0	0
5	43	0	43
6	60	0	60
7	0	0	0
8	0	0	0
9	40	0	40
10	0	0	0

Found Sentiment Phrases	Phrase Score	Count
easy	60	339
ideal	90	131
good	60	119
nice	25	114
great	80	113
better	60	108
friendly	40	84
problems	-60	41
best	90	41
love	70	37

IBM SPSS Modeler:

0011

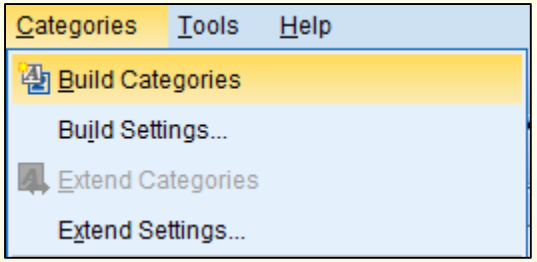
- Build the categories (concepts) based on the recurring terms and phrases.

Build Extend Score Display

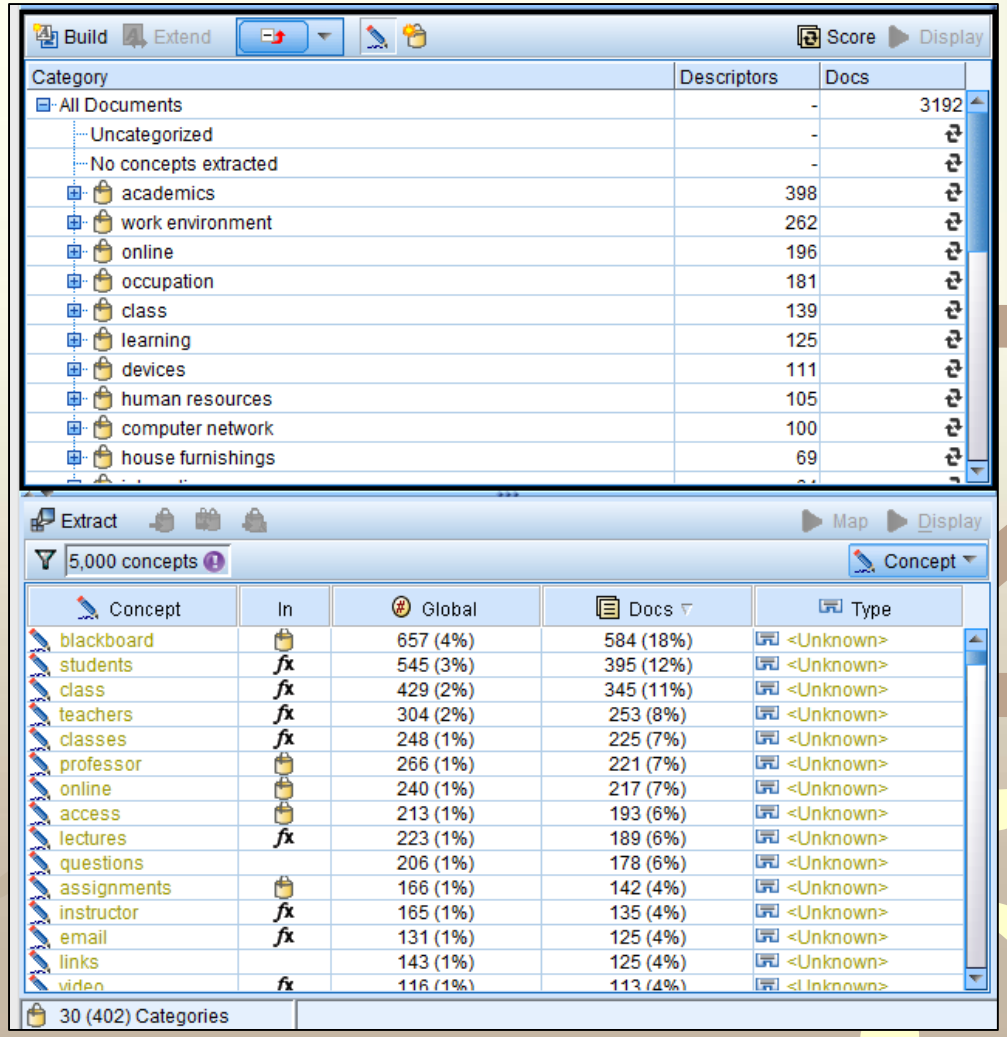
Category	Descriptors	Docs
All Documents	-	3192
Uncategorized	-	
No concepts extracted	-	
academics	361	
online	248	
consumer electronics	240	
occupation	189	
storage devices	134	
class	132	
learning	90	
human resources	83	
computer network	68	
house furnishings	61	
economics	60	
question	48	
course	46	
Neg: Service: Accessibility	44	

Categorization

0011



- Modeler counts the frequency of terms and phrases.
- Based on the words it builds categories.



Category Bar by frequency

0011

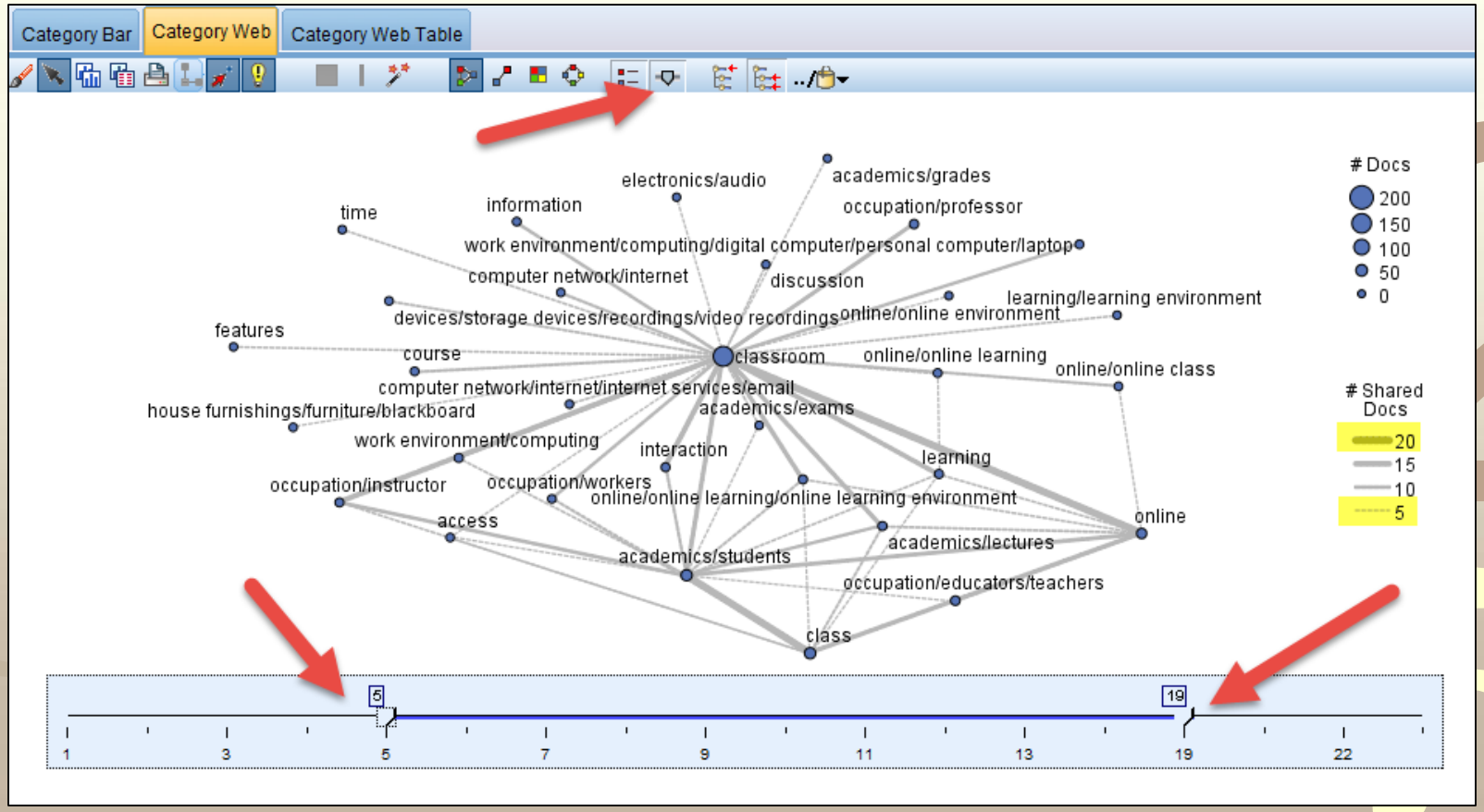
Category Bar | Category Web | Category Web Table

Category	Bar	Selection %	Docs
computer network		100.0	371
academics		41.2	153
occupation		35.3	131
class		28.3	105
house furnishings		28.0	104
work environment		27.2	101
online		21.3	79
access		15.4	57
devices		13.5	50
economics		12.1	45
course		10.8	40
learning		10.8	40
human resources		9.7	36
discussion		8.1	30
information		7.5	28
asu		6.7	25
optical device		6.2	23
time		5.9	22
features		5.7	21

Category Web:

0011

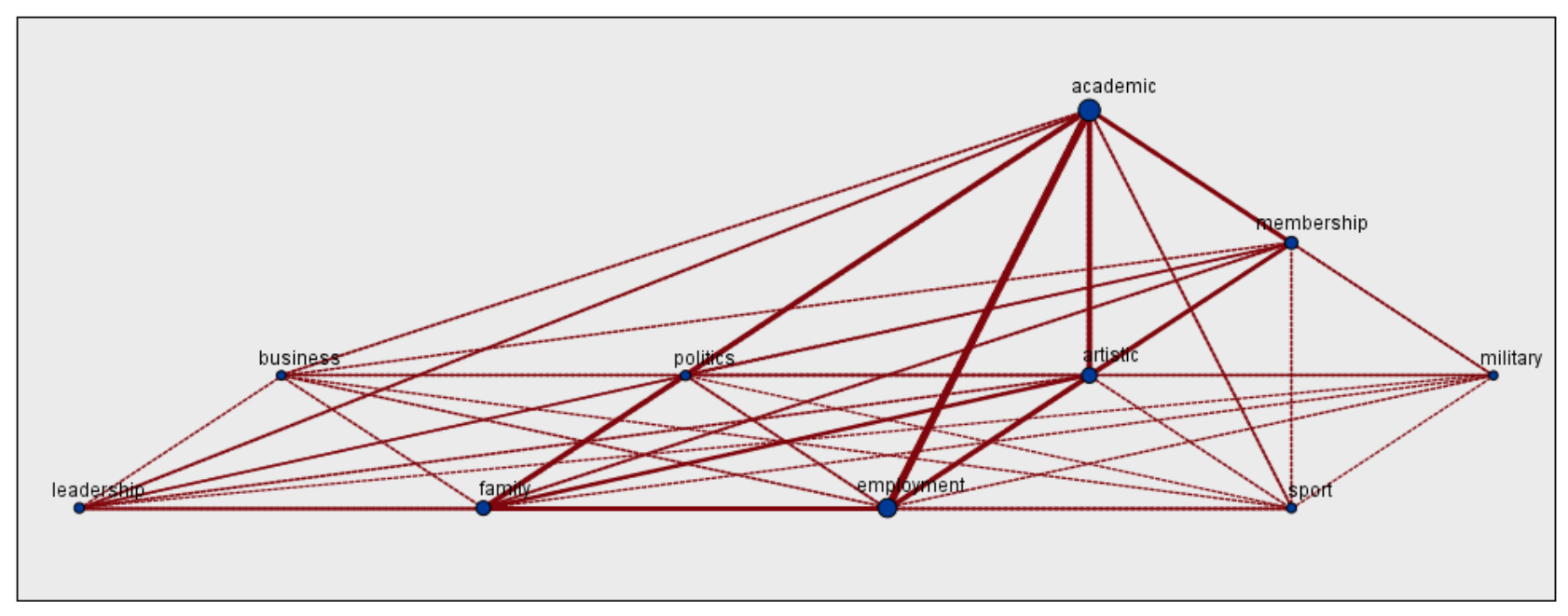
- Swanson process by text mining
- Show the inter-relationship between concepts



Category web

0011

- Similar to Code relation chart in MAXQDA
- Thicker line → stronger relationship (more co-occurrence)



Example of sub-categories

0011

- The researcher can drill down the category to view the sub-categories.
- The original responses are highlighted for the researcher to cross-examine.

Pos: Product: Usability (494)
- [easy *] (366)
- [user-friendly] (53)
- [<Usability> & <Positive>] (33)
- [accessible] (33)
- [* website * & <Positive>] (29)
- [<Registration> & <Positive>] (23)
- [able to access] (16)
- [able to connect] (3)
- [portable] (1)

	Id	Response	Categories
1	298628924	It would be nice to have an IM setup within the online learning enviroments for the people in the class to discuss things. Like I said before, in the online learning environment I would include the video of that day's class. So you can just go online and check out what happened that day. Another thing would be a calendar. All the classes that are setup on the learning environment would be integrated with a central calendar. It would display the tests, assignments, for each class. I know a lot of students who carry their calendar to class everyday. Make their life easy.	Fea: [record + .] Neg: Product: Functio... Neg: Product: Usability Pos: Product: Functio... Pos: Product: Usability Pro: [electronic devic...

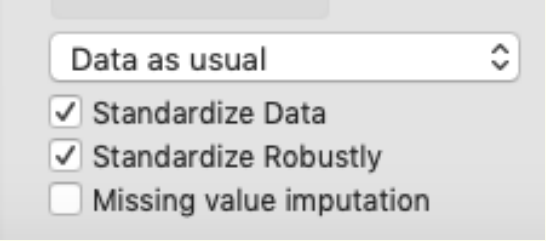
Hierarchical clustering for concept linking

Statement	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
S1	5	1	1	1	0	0	0	0	0	0
S2	1	5	0	2	1	1	1	1	0	0
S3	1	0	4	1	0	2	0	1	2	0
S4	1	2	1	5	1	1	0	2	1	0
S5	0	1	0	1	5	0	1	2	1	1
S6	0	1	2	1	0	5	2	1	3	0
S7	0	1	0	0	1	2	5	0	2	0
S8	0	1	1	2	2	1	0	5	1	0
S9	0	0	2	1	1	3	2	1	5	0
S10	0	0	0	0	1	0	0	0	0	5
S11	0	1	0	0	0	0	0	0	0	1

The numbers show the frequency of pairing: When people mentioned S2, how often do they mention S4?

Bonney, L., & Yu, C. H. (2018, January). *Sharing tacit knowledge for school improvement*. Paper presented at International Congress for School Effectiveness and Improvement, Singapore.

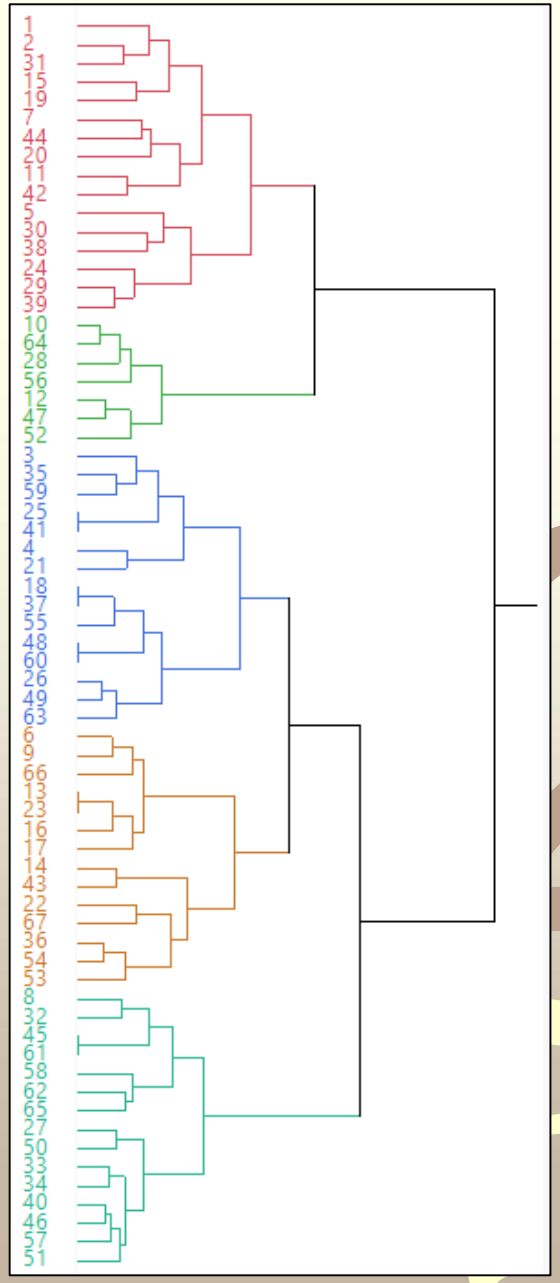
Hierarchical clustering



- Like dragging the regression line in regression modeling, the presence of outliers might skew the grouping patterns.
- if a column has outliers, then the estimate of the standard deviation is inflated.
- Choose Standardize robustly
- By doing so outliers stand out, resulting in more isolated clusters. However, those remaining columns can be used to form more accurate clusters.

Hierarchical clustering

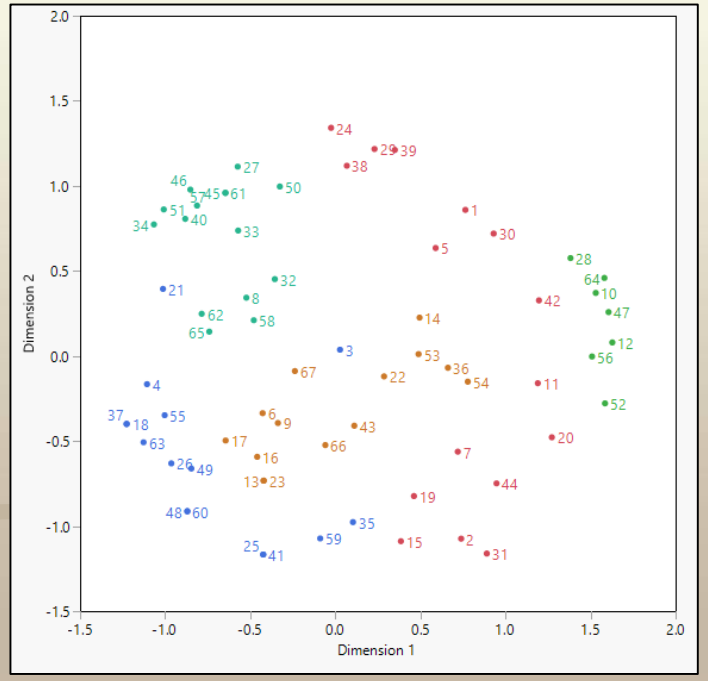
- The analyst can change the number of clusters on the graphical output.
- Based on the result it was decided that there should be 5 clusters for these terms.

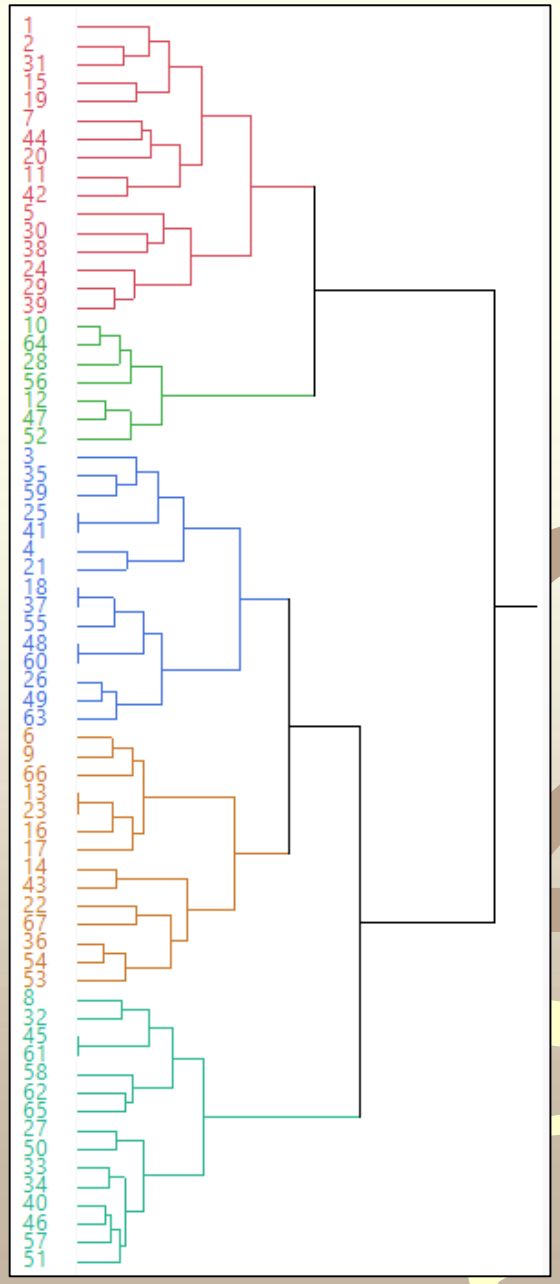


Compare HC and MDS

0011

- HC and MDS agree with each other to a large extent
- But there are some discrepancies

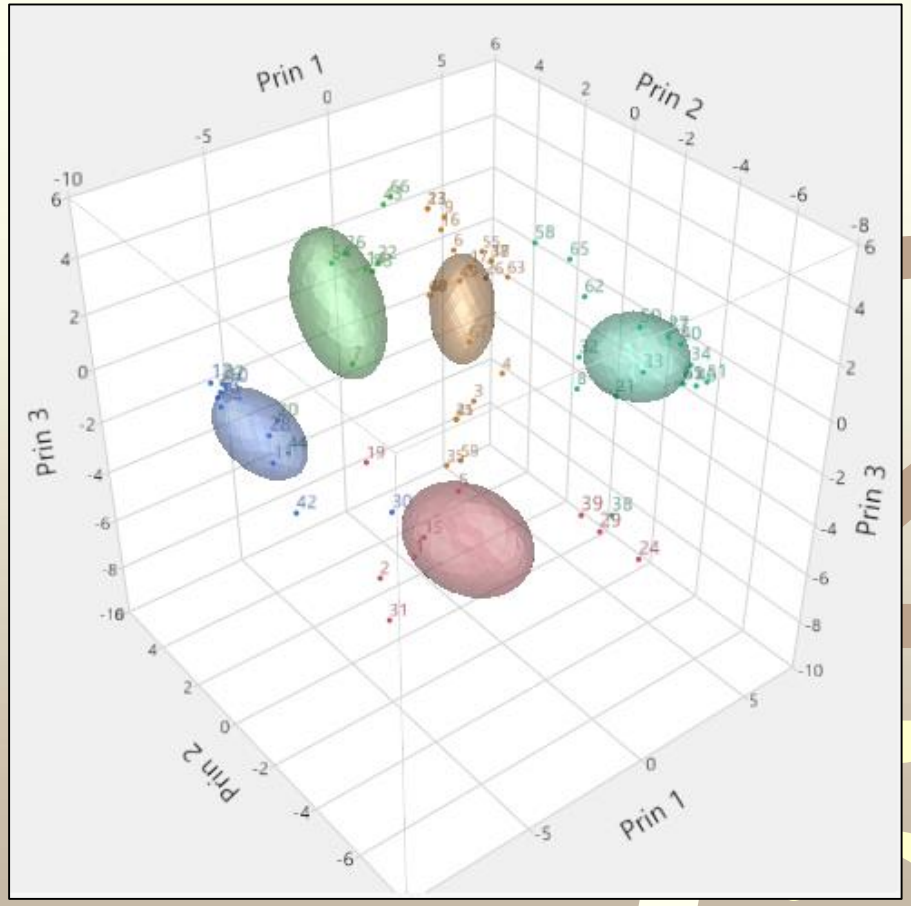




Biplot in K-mean clustering

0011

- To check HC and MDS, add K-mean clustering
- Use Biplot as the third verifier.
- But we still need to read the original text.



Summary, conclusion, & recommedations

- Content analysis, grounded theory, and Swanson process are precursors of text mining.
- Incorrect stop words, sarcastic expression, negation…etc. might fool NLP. The automated process must be checked by humans.
- Code relation matrix, category web, latent class analysis, hierarchical clustering, multi-dimensional scaling, k-mean biplot should be employed for triangulation.

Recommendations

- Some authors (e.g. Bennett, Dumais, & Horvitz, 2005) suggest ensemble methods, such as using multiple text mining tools and assigning reliability index to each of the results.
- Next, the researcher can select the best text classifier or combining all results to generate a meta-result.

Recommendations

- Triangulation between results coded by humans and concepts extracted by text mining is good for a small project.
- When there are too many documents, sample a subset for human coders for comparison or use text mining only.