

Revealing Relations between Open and Closed Answers in Questionnaires through Text Clustering Evaluation

Magnus Rosell and Sumithra Velupillai



2008-05-28

Content

- ▶ Introduction and Motivation
- ▶ Text Set and Exploration Tool
- ▶ Method
- ▶ Example
- ▶ Evaluation
- ▶ Conclusion

Introduction and Motivation

- ▶ Questionnaires – an important source for research
- ▶ Hidden information in open free-text answers
- ▶ Time-consuming to analyze manually
- ▶ Text clustering could aid

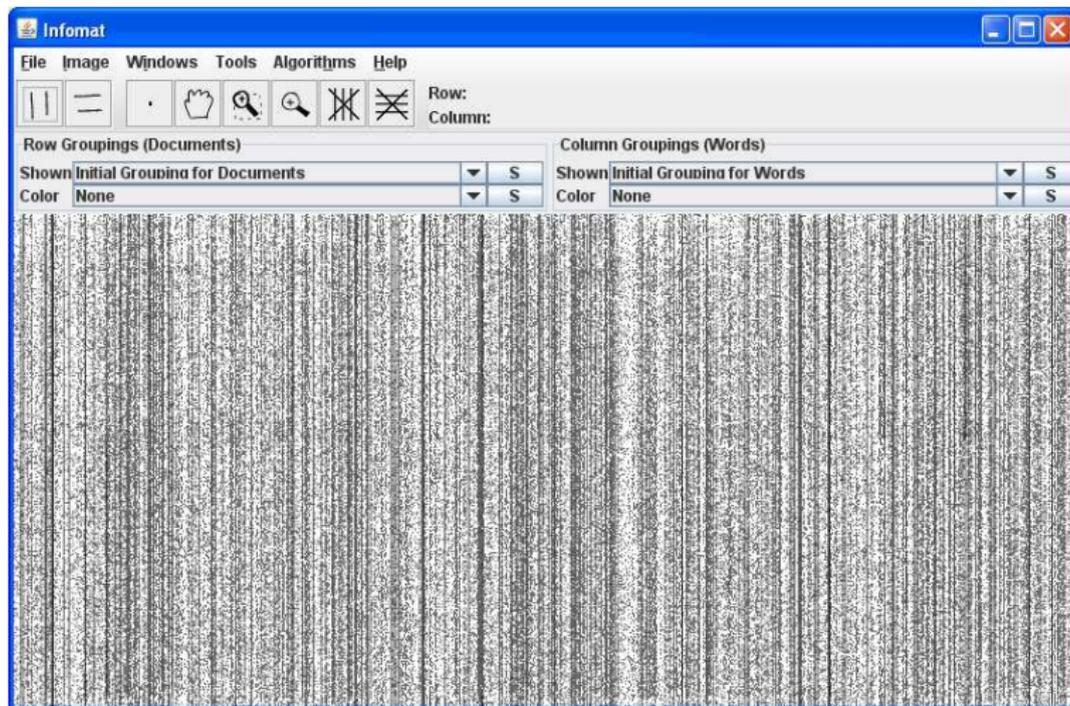
Text Set and Exploration Tool

- ▶ The Swedish Twin Registry
- ▶ A questionnaire
 - ▶ An open answer: occupation (41 549)
 - ▶ Vector space model
 - ▶ A closed answer: smokers (29%)
- ▶ Infomat – A vector space exploration tool
 - ▶ <http://www.csc.kth.se/tcs/humanlang/tools.html>

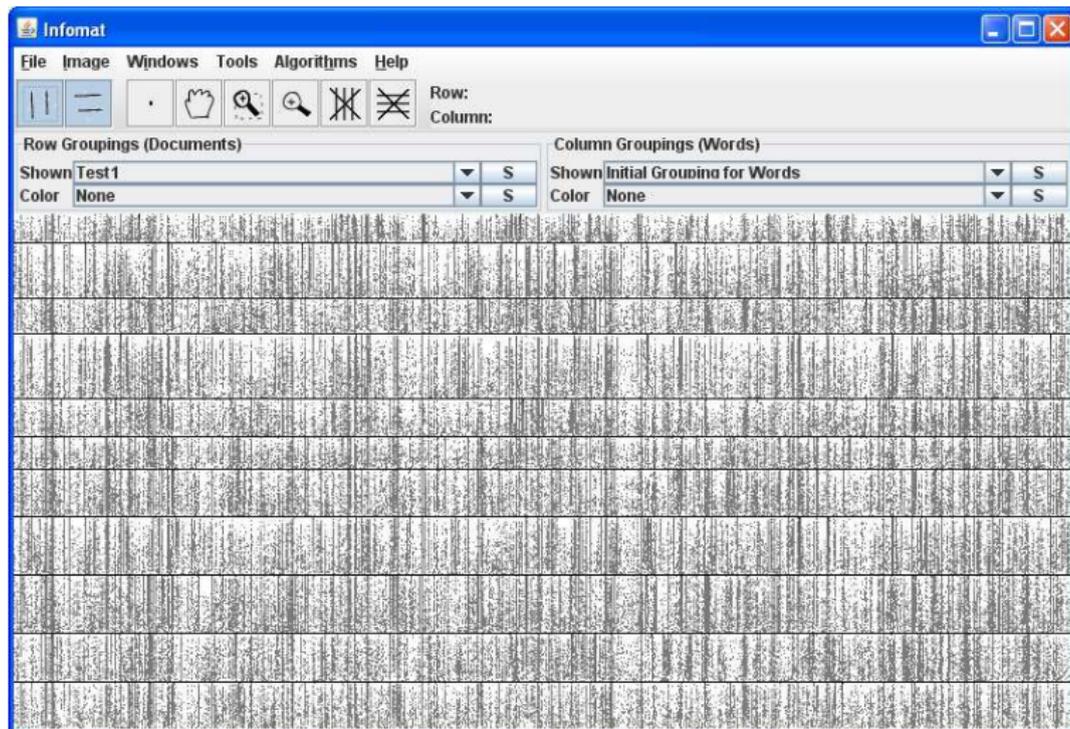
Method

1. Cluster the text set
2. Identify interesting clusters
3. Explore cluster contents
4. Formulate potential hypotheses
 - ▶ Iterate
 - ▶ Interactive exploration
 - ▶ Pursue hypotheses further

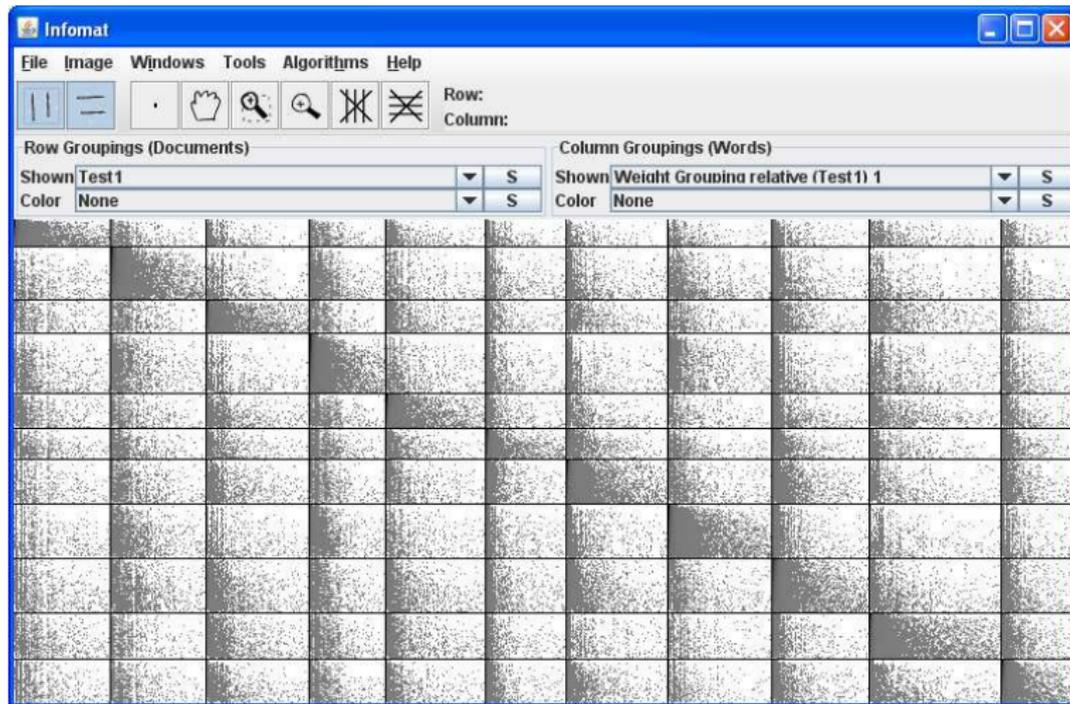
Example – Text Set



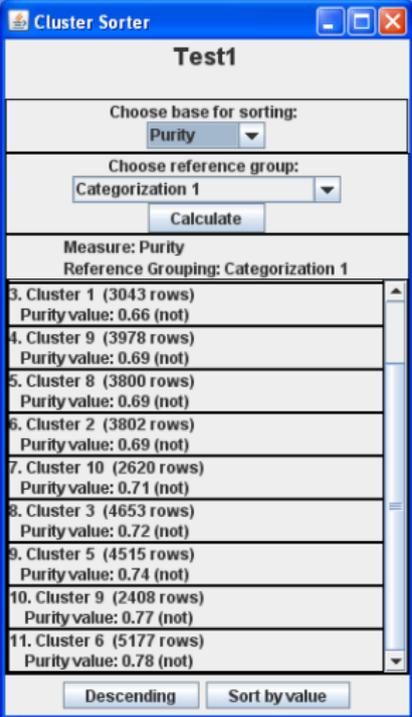
Example – Clustering



Example – Relative Clustering



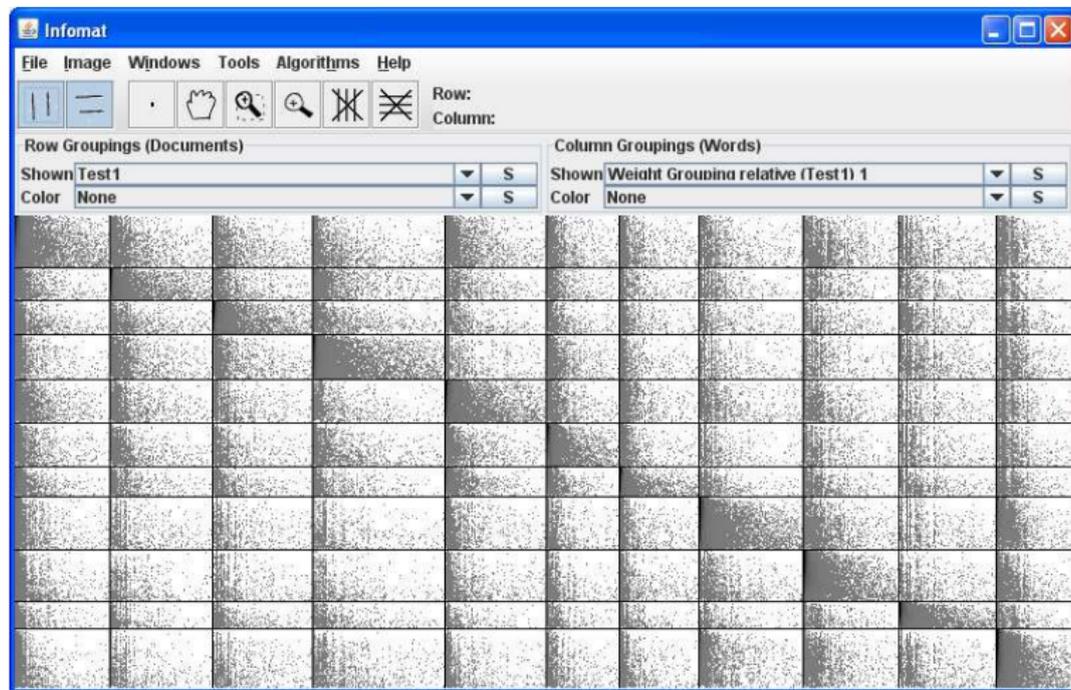
Example – Evaluated Clusters



The screenshot shows a software window titled "Cluster Sorter" with a sub-header "Test1". The interface includes a "Choose base for sorting:" dropdown menu set to "Purity", a "Choose reference group:" dropdown menu set to "Categorization 1", and a "Calculate" button. Below these controls, the current settings are displayed as "Measure: Purity" and "Reference Grouping: Categorization 1". A list of clusters is shown, sorted by their purity values in descending order. Each cluster entry includes its ID, the number of rows it contains, and its purity value followed by "(not)". At the bottom of the window, there are two buttons: "Descending" and "Sort by value".

Cluster ID	Number of Rows	Purity Value	Status
3. Cluster 1	3043	0.66	(not)
4. Cluster 9	3978	0.69	(not)
5. Cluster 8	3800	0.69	(not)
6. Cluster 2	3802	0.69	(not)
7. Cluster 10	2620	0.71	(not)
8. Cluster 3	4653	0.72	(not)
9. Cluster 5	4515	0.74	(not)
10. Cluster 9	2408	0.77	(not)
11. Cluster 6	5177	0.78	(not)

Example – Sorted Clustering



Example – Zoom on an Interesting Cluster

The screenshot displays the Infomat software interface. On the left, the 'Pixel Info' panel shows details for a selected object: 'Object: id(19581)', 'Group: Smokers10', 'Object: lant', and 'Group: Weight Group rel...'. Below this, a 'Pixels' section contains a list of items with their weights, frequencies, and counts:

Weight	Frequency	Count
w: 0.7800455	f: id(5030)	c: gård
w: 0.69715357	f: id(1650)	c: gård
w: 0.62941206	f: id(8768)	c: lant
w: 0.53614277	f: id(1100)	c: lant

The main window, titled 'Infomat', features a menu bar (File, Image, Windows, Tools, Algorithms, Help) and a toolbar with icons for zooming and selection. The interface is divided into 'Row Groupings (Documents)' and 'Column Groupings (Words)'. The 'Row Groupings' section shows 'Shown: Test1' and 'Color: None'. The 'Column Groupings' section shows 'Shown: Weight Groupina relative fT...' and 'Color: None'. The central area displays a grid of document clusters, with a mouse cursor hovering over a specific cluster in the middle row and middle column.

Evaluation

- ▶ Farmers smoke less than the average
 - ▶ A few hours of exploration
 - ▶ No prior knowledge on smoking habits in occupation groups
- ▶ Comparable surveys
- ▶ Hypotheses can be generated

Conclusion

- ▶ No need to avoid free-text answers
 - ▶ valuable
 - ▶ analyze with our method
- ▶ Previously unknown relations were revealed
- ▶ Interaction!
- ▶ Pursue hypothesis further
- ▶ Future work
 - ▶ Questionnaires from other domains
 - ▶ Similar text sets, e.g. electronic medical records