An Investigation of Text Mining Techniques for the Analysis of Abstracts

David J. Marchette

dmarchette@gmail.com

Naval Surface Warfare Center

Code B10

< > - +

Organizing a Conference

Conference organizers have two difficult tasks:

- Combining contributed papers into sessions.
- Scheduling sessions to minimize the overlap of similar topics.

Often, the only information an organizer has is the abstracts for the talks. Some ideas from text mining can help with these tasks.

There are constraints:

- Some papers **must** go together (invited sessions).
- Some papers **cannot** be scheduled in parallel (speakers in two places at once).
- Some papers **should not** be scheduled in parallel sessions.

< > - +

Text Processing

We need to be able to compare abstracts. We will assume no structure to the abstracts, no keywords provided. Let $\mathcal{C}$ be the corpus (all abstracts) and $D \in \mathcal{C}$ be a document (an abstract). We will make use of the term/document mutual information, for word w in document D within corpus $\mathcal{C}$:

$$m_{\mathcal{C}}(w, D) = \log \frac{P(w \in D)}{P(w \in \mathcal{C})}$$

We estimate this as

$$\widehat{m}_{\mathcal{C}}(w, D) = \log \frac{\frac{\#w \in D}{|D|}}{\frac{\#w \in \mathcal{C}}{|\mathcal{C}|}}$$

where $\#w \in D$ is the number of times w appears in D and $|D|$ is the number of words in D.

Important Words

We can determine the important words in any document by thresholding the mutual information. We can then compare documents according to the number of “important” words they share:

$$s(d_1, d_2) = \frac{d_1 \cap d_2}{d_1 \cup d_2}$$

after thresholding on the mutual information. This defines an intersection graph (an edge between two documents if s is large) and we can use spectral graph methods to project to a low dimensional space.

Iterative Denoising

- Mutual information is a corpus dependent feature.
- Words from one domain can be “noise” in another.
- Iterative denoising aims to remove the noise, but only where it **is** noise:
  - Compute the MIs.
  - Threshold to throw out (document dependent) “stopper words”.
  - Cluster the documents.
  - Within each cluster, repeat.

< > - +

Science News Example

- 1160 articles from Science News.
- Grouped (by hand) into 8 categories.
- Words stemmed by a Porter stemmer.
- MI thresholded to remove document specific “stopper words”.
- Intersection graph defined, projected to 2D.

< > - +

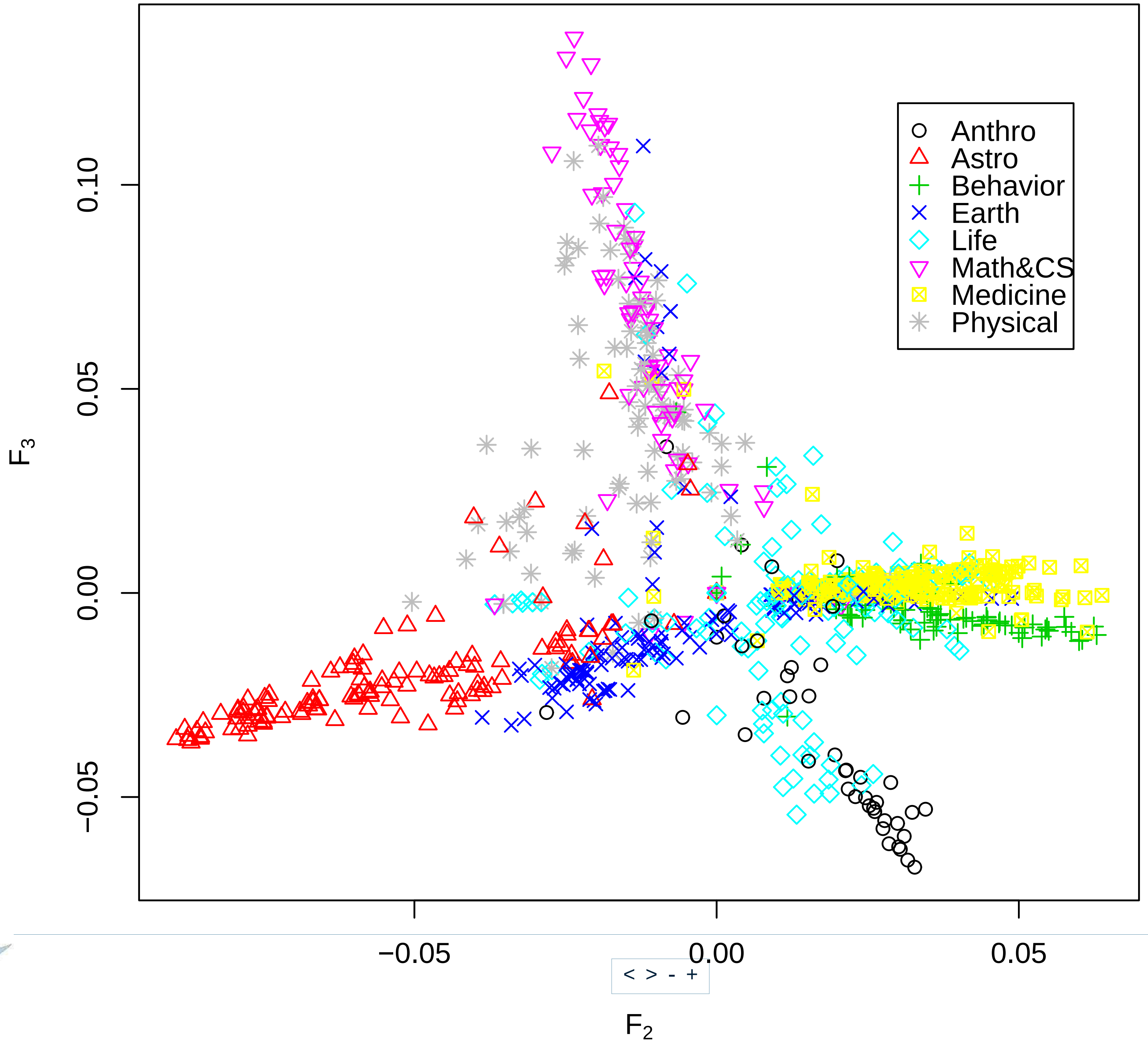
Visualizing the Data

Let A be the adjacency matrix of a graph G, and D be the diagonal matrix with $d_{ii} = degree(v_i)$. The Laplacian of G is: $L(G) = D - A$. Some authors use a scaled version:

$$\mathcal{L}(G) = D^{-\frac{1}{2}} L(G) D^{-\frac{1}{2}},$$

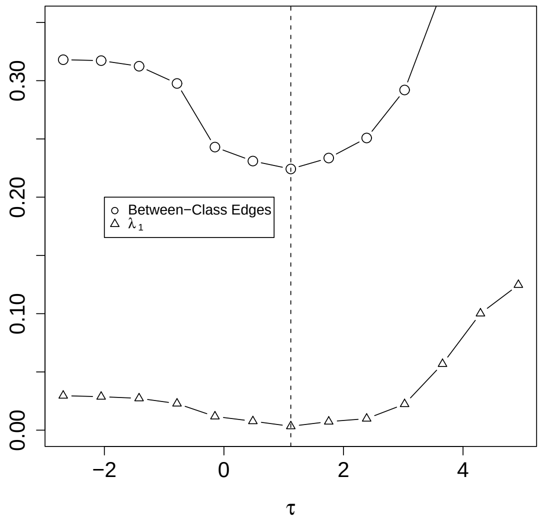
where $D_{ii}^{-\frac{1}{2}} = \frac{1}{\sqrt{dii}}$. We will use this latter definition. The eigenvectors associated with the smallest (nonzero) eigenvalues provide an analog to the multidimensional scaling projection of the graph.

Science News Example



How to Threshold?

In the Science News example we selected the threshold on the mutual information by observing the smallest nonzero eigenvalue of the intersection graph as a function of the threshold.



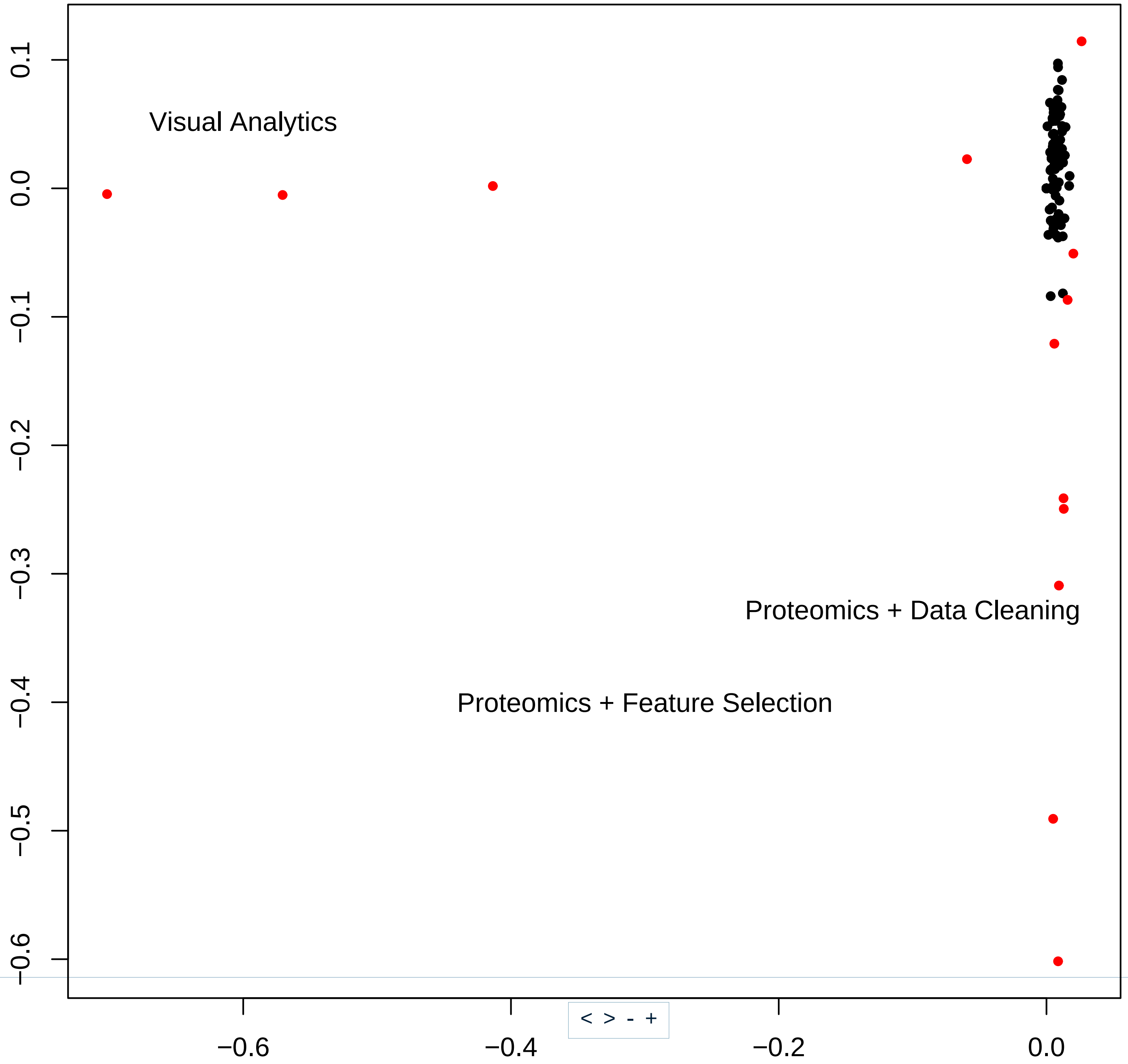
This does not work for the abstracts. Instead we chose an arbitrary threshold of 1.

< > - +

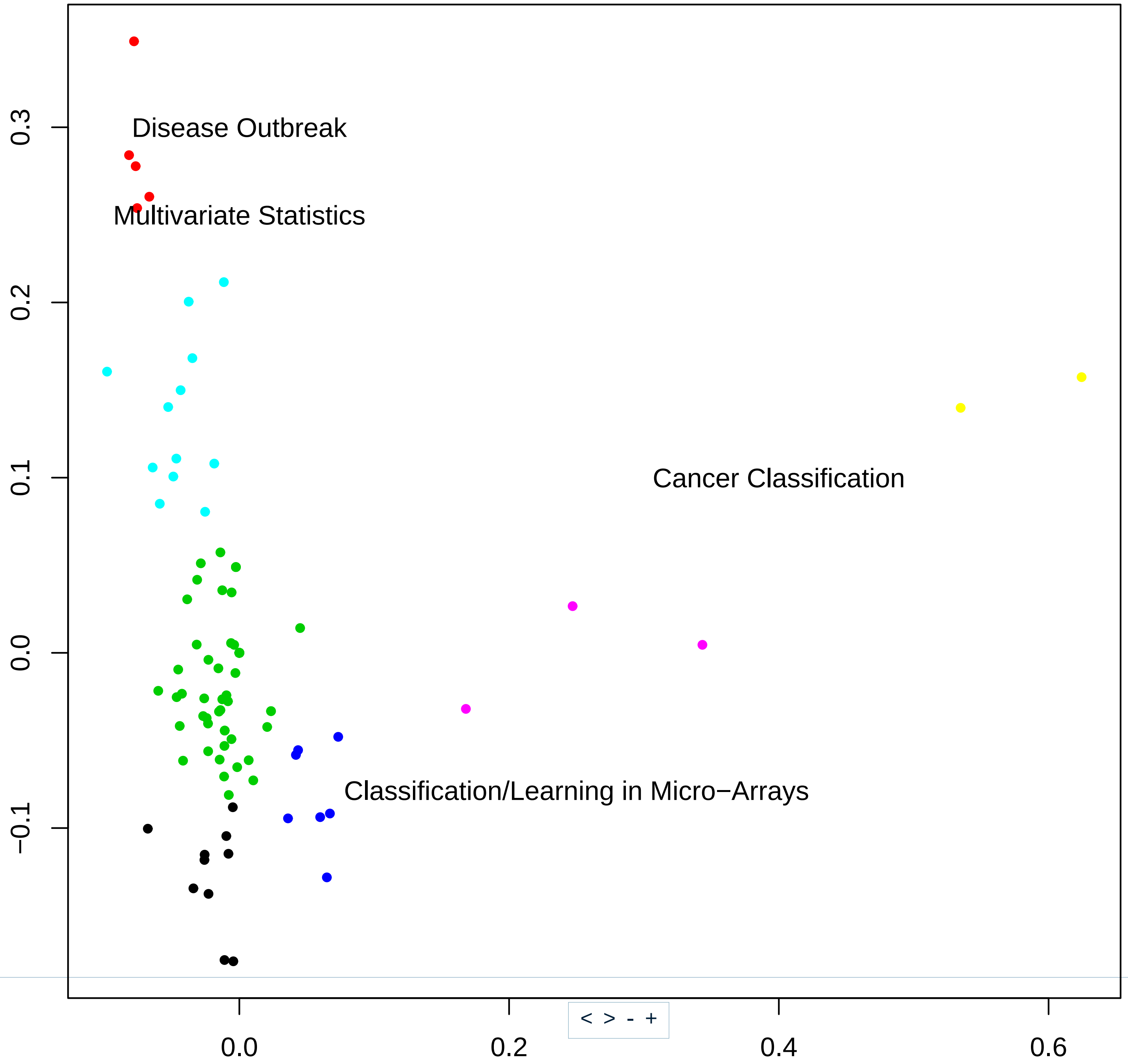
Interface Abstracts Example

- All abstracts from 2004 Interface Conference.
- 136 abstracts.
- Features consist of words, bigrams and trigrams.
- Mutual information computed independently for words, bigrams and trigrams.

Interface 2004 Abstracts



Interface 2004 Abstracts



Interface 2004 Abstracts

- At each stage we peel off the abstracts that group together.
- As we drill down, the abstracts continue to separate into sessions.
- Occasionally, a group does not seem to fit together (the words retained are content-free) and these are sent down the tree instead of being removed.
- There is quite a bit of user-interaction:

  Dimension to project.
  Thresholds (on MI and on intersection graph).
  Clustering method to use.
  Decision of which groups to keep.

Effects of Thresholding

Words, bigrams and trigrams in common between two documents:

the	and	of	that
ha	infect	viru	human
nile	west	been	an
a	to	first	howev
thi	mani	in	within
it	mosquito	indic	model
nile<>viru	west<>nile	and<>human	ha<>been
of<>west	to<>human	the<>first	of<>an
west<>nile<>viru	of<>west<>nile		

Effects of Thresholding

Words, bigrams and trigrams in common between two documents after thresholding:

ha	infect	viru	human
nile	west	been	an
first	howev	within	mosquito
indic	nile<>viru	west<>nile	and<>human
ha<>been	of<>west	to<>human	the<>first
of<>an	west<>nile<>viru	of<>west<>nile	

< > - +

Corpus Dependent Features

Note that the “important words” depend on the corpus. After removing one or more sessions, the words used for the remaining sessions have different MI.

- In a group of proteomics abstracts the word “proteomics” is not very useful for comparisons.
- In a group of bioinformatics abstracts, “proteomics” might be quite useful.

This is not an automated procedure. Because of the small size of the documents there is a lot of noise in the MI calculation. Stopper words need to be considered carefully (university affiliations).

Corpus Dependent Features

- Consider a document on the West Nile Virus.
- We look at the words with maximal mutual information when computed over the full corpus, compared to those computed over a small corpus of those abstracts containing “virus” or “epid”.
- Words change their importance in different contexts.
- We consider words with the highest MI in the smallest corpus as compared to those of the larger.

The Anatomy of a Bioevent: West Nile Virus in Washington, DC

In late 1999, West Nile fever, which is caused by a mosquito-vector ed virus, was identified in New York City during an epidemic involving 62 human cases and 7 fatalities. This was the first documentation of West Nile virus in the Western Hemisphere. It has been suggested the virus was transferred by accident via an infected human passenger arriving by airflight from the Middle East. The virus subsequently gained ecological establishment and now has been identified throughout the eastern United States, northward to Canada and as far south as Florida, with 50% of the continental U.S. effected within 18 months of initial introduction. The appearance of West Nile virus in a human community, like many insect-vectored pathogens, may be responsive to modulation by weather and climate variations (referred to as enviro-climatic modulation). Serious concern remains that other insect-vectored exotic pathogens may also gain entry to the U.S by way of accidental or intentional impor! tation. These pathogens may also be responsive to enviro-climatic modulation. Remotely Sensed Epidemic Surveillance (RSEPIS) of proxy indicators and warnings of epidemic initiation and propagation represent an area of promise for the public health community due to implications for proactive versus reactive epidemic control measures. Increases in incidence for some infectious diseases have been attributed to modulation of endemic ecology via enviro-climatic coupling, and delineation of climatic patterns thought to influence the appearance of an infectious disease in a community can be directly applied to the development of an RS infectious disease surveillance system. Significant obstacles exist, however, regarding basic understanding of the mechanisms of epidemic triggering, propagation, and conclusion that hinder rapid development of RS disease forecasting systems. Idealization of remotely sensed disease surveillance system development includes a comprehensive understan! ding ...

Corpus Dependent Features

Some words that are moved into the top MI by the smaller corpus:

epidem
system
identifi
remot
attribut
transmiss
mathemat

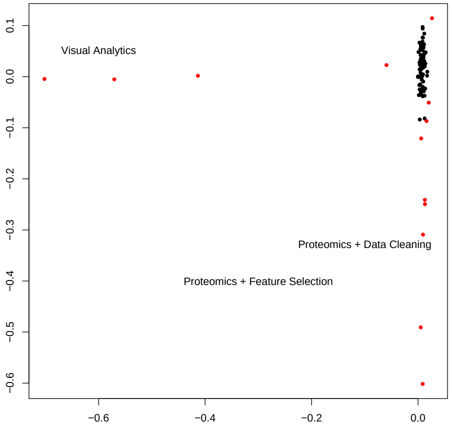
diseas
infecti
gain
mechan
public
highlight
forecast

surveill
involv
respons
environ
fever
refer
state

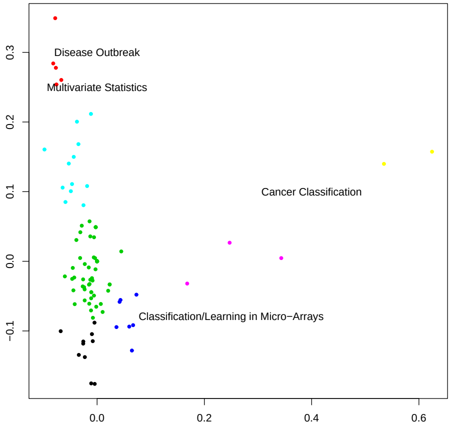
commun
bioevent
washington
implic
obstacl
influenc
far

Organizing the Abstracts

Compute the features, build the graph, project, investigate clusters or outliers:



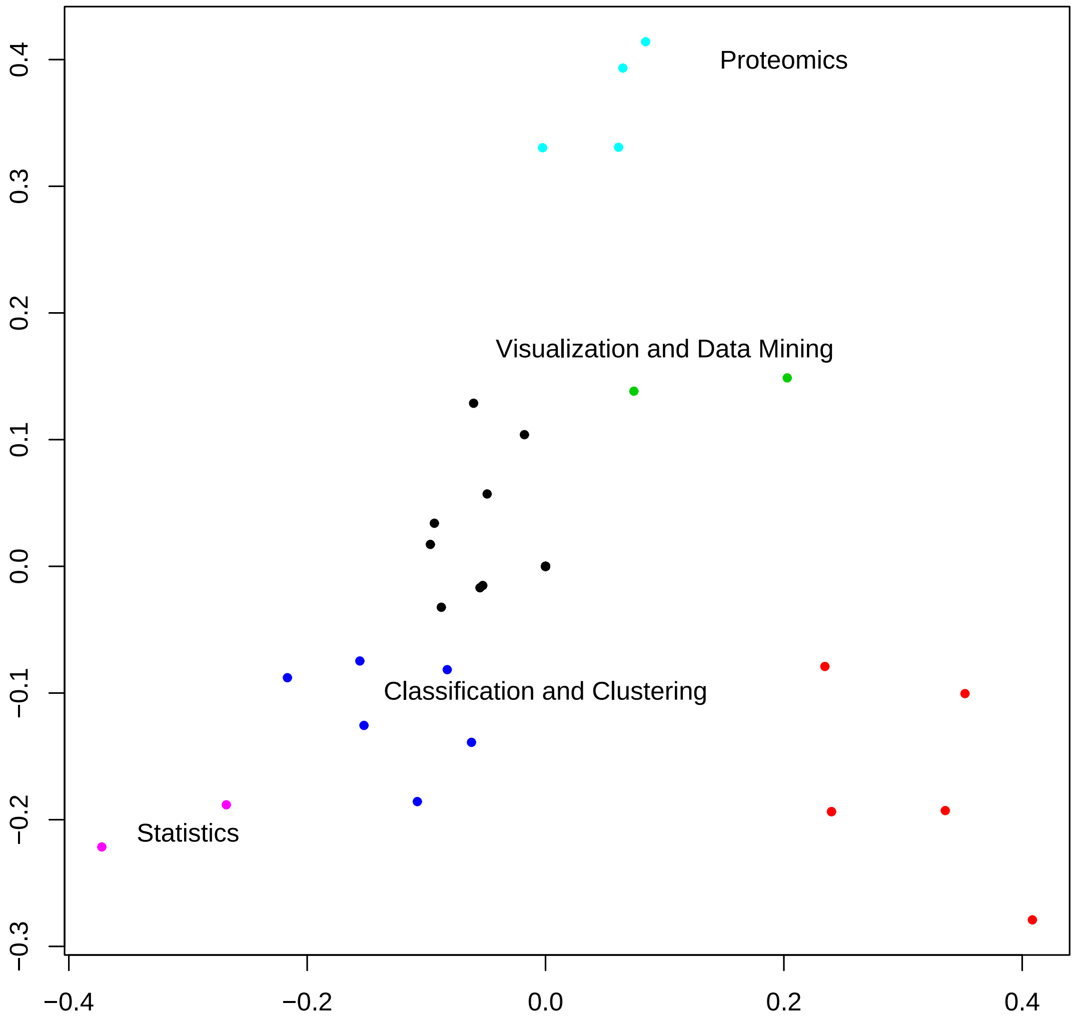
Iterate on the clusters:



Scheduling

- Each session becomes a document.
- Run the procedure on these.
- Now documents that cluster together should be scheduled in non-overlapping times.
- Speaker conflicts need to be handled separately.

Sessions


Proteomics
Visualization and Data Mining
Classification and Clustering
Statistics
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.2
0.0
0.2
0.4

< > - +

Difficulties

- Abstracts tend to be short.
- Removing “stopper” words is more critical.
- Short documents means MI estimate is noisy.
- Should incorporate document structure: Title, keywords (if given).
- Stemming may not be as critical: less flexibility of language.
- Abstracts sometimes deliberately vague.

A Larger Dataset

- Abstracts related to water purification taken from Web Of Science.
- 4378 abstracts total from three journals:

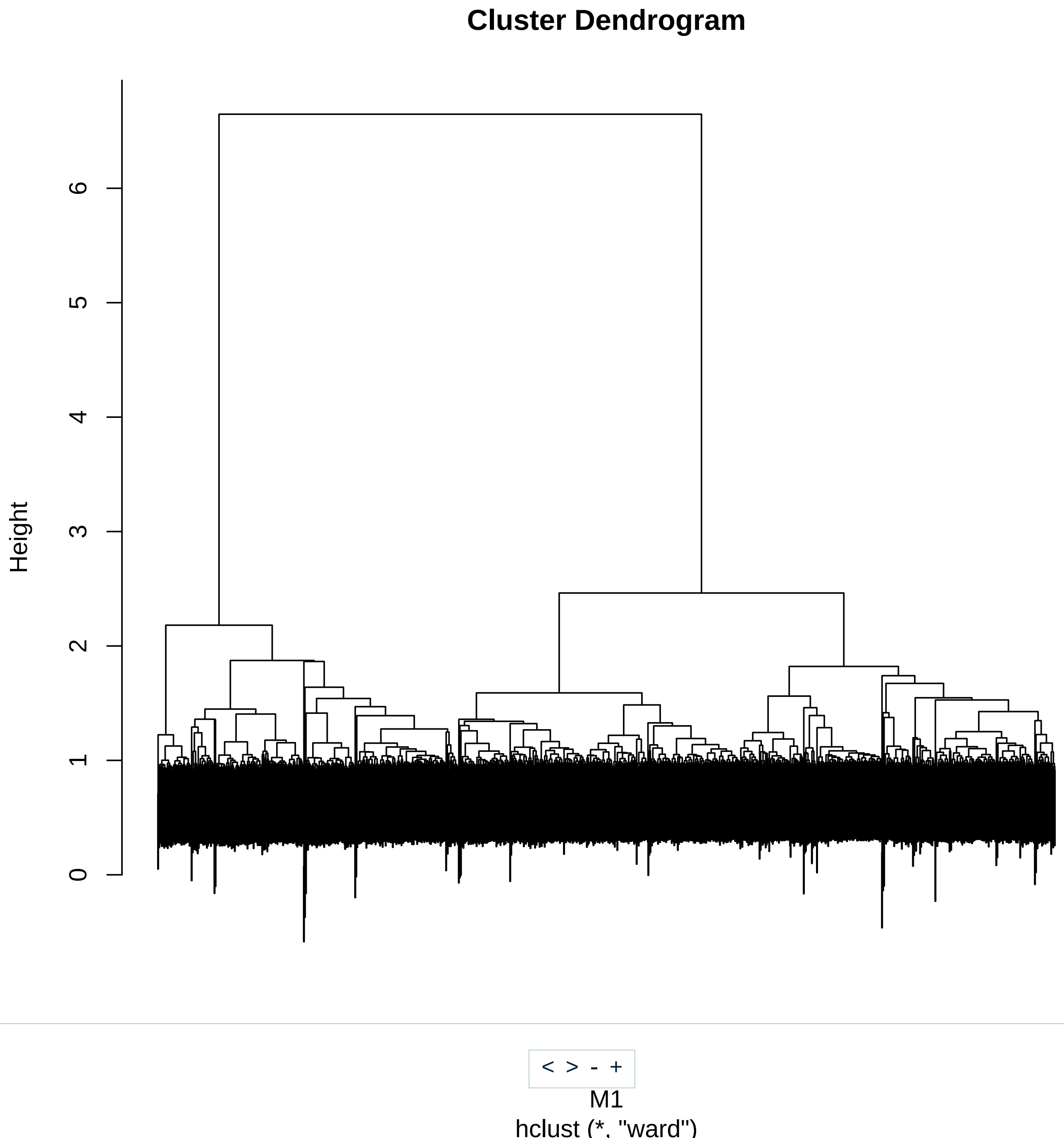
  Desalination

  Water Research

  Water Science and Technology

< > - +

First Clustering



Observations

- Top clustering is almost perfect: Desalination Journal vs Others.
- Iterative denoising was performed until leaves are small.
- Investigated the resulting clusters:

  Filters and Membranes

  Electro-dialysis

  Reverse Osmosis

  Flocculation

  Distillation

  Bioreactors
- This provides an overview of the different technology areas.
- Sub-classifications available within different leaves.

Discussion

- Text mining is useful for many applications. Can be used to
  - help schedule conference papers
  - identify key concepts
- Abstracts offer interesting challenges
  - short
  - heavy on buzzwords
  - often at a high level
- Iterative denoising can provide interesting insight and useful results.