Using Text Analysis to Characterize Student Learning in an Introductory Statistics & Data Science Course

Ron Yurko Rebecca Nugent Philipp Burckhardt

Department of Statistics & Data Science
Carnegie Mellon University

eCOTS 2018

Revamp of introductory Statistics & Data Science course

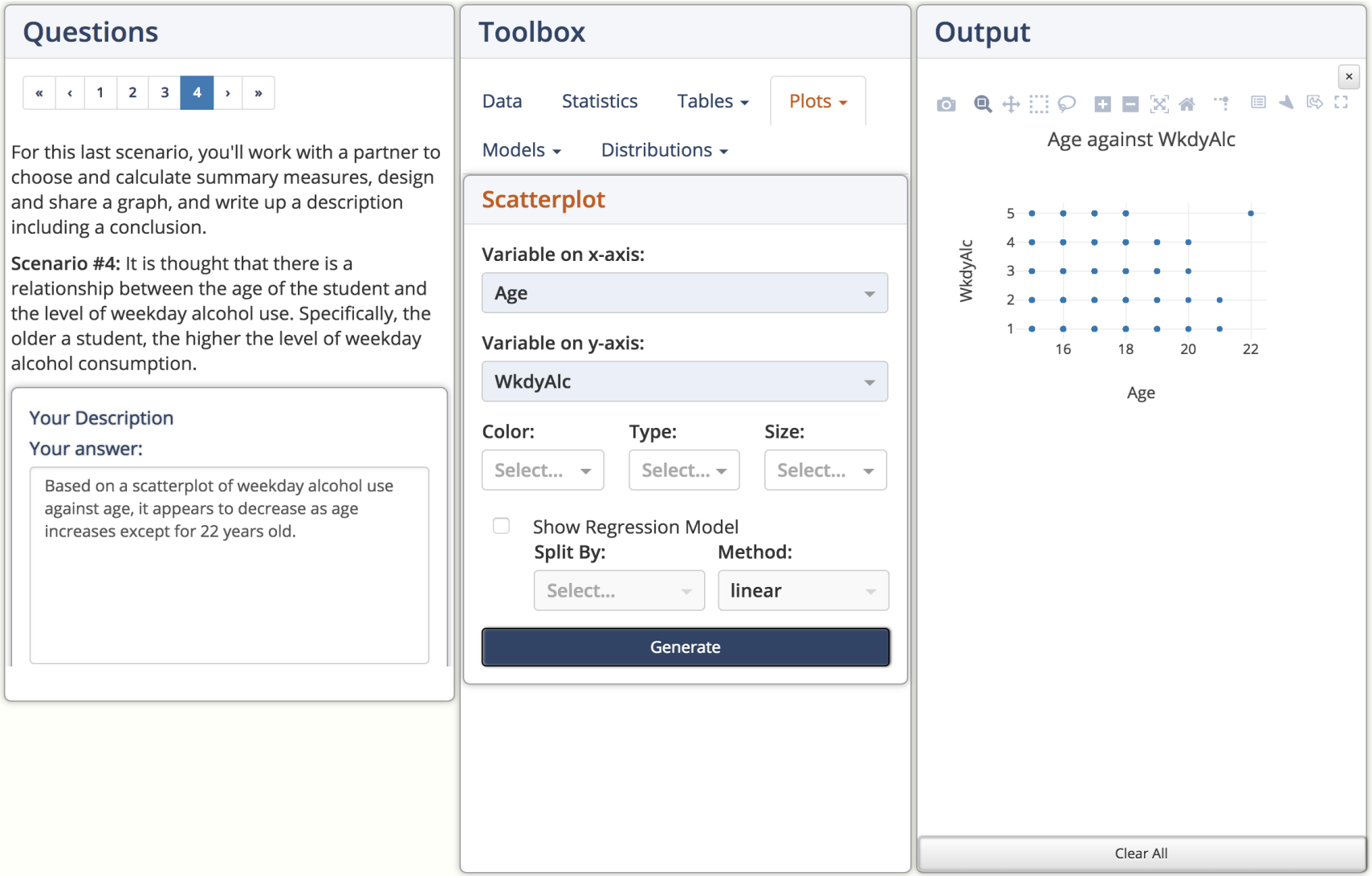
Required course in Dietrich College of Humanities and Social Sciences general education curriculum for students in several programs:

- Economics, English, History, Information Systems, International Relations, Modern Languages, Philosophy, Psychology, Social & Decision Sciences, and Statistics & Data Science
- Also taken by majors across campus

New emphasis on student inquiry and writing about data analysis with non-traditional data types and interdiscinplinary case studies

Students **interact with the ISLE e-learning framework** (browser-based Interactive Statistics Learning Environment) developed by Philipp Burckhardt

Students engage in the data analysis workflow with an interactive explorer that records answers and actions



Students engage in the data analysis workflow with an interactive explorer that records answers and actions

Time: 11:30:22 PM|**User:** ryurko@andrew.cmu.edu
ID: description_scenario4 |**Type:** FREE_TEXT_QUESTION_SUBMIT_ANSWER
Value: Based on a scatterplot of weekday alcohol use against age, it appears to decrease as age increases except for 22 years old.

Time: 11:24:33 PM|**User:** ryurko@andrew.cmu.edu
ID: schoolabsence |**Type:** DATA_EXPLORER:SCATTERPLOT
Value: {


```
  "xval": "Age",
  "yval": "WkdyAlc",
  "color": null,
  "type": null,
  "regressionLine": false,
  "regressionMethod": "linear",
  "lineBy": null
}
```

Students engage in the data analysis workflow with an interactive explorer that records answers and actions

Time: 11:30:22 PM|**User:** ryurko@andrew.cmu.edu
ID: description_scenario4 |**Type:** FREE_TEXT_QUESTION_SUBMIT_ANSWER
Value: Based on a scatterplot of weekday alcohol use against age, it appears to decrease as age increases except for 22 years old.

Time: 11:24:33 PM|**User:** ryurko@andrew.cmu.edu
ID: schoolabsence |**Type:** DATA_EXPLORER:SCATTERPLOT
Value: {


```
  "xval": "Age",
  "yval": "WkdyAlc",
  "color": null,
  "type": null,
  "regressionLine": false,
  "regressionMethod": "linear",
  "lineBy": null
}
```


Given these action logs, how can we characterize how student approach data analysis? The way they write about data?

Text analysis of students' answers

Represent student text answers in a matrix where rows are individual student answers and columns are unique words

Values in matrix could be:

- does the student write the word: Yes / No
- number of times the word appears in each answer

[1] [Salton and McGill, 1986]

Text analysis of students' answers

Represent student text answers in a matrix where rows are individual student answers and columns are unique words

Values in matrix could be:

- does the student write the word: Yes / No
- number of times the word appears in each answer
- penalized frequency by how many answers the word appears in

Term Frequency - Inverse Document Frequency (TF-IDF)[1]:

$$\text{TF-IDF} = \#\text{ of times word is in answer} \cdot \frac{\#\text{ of answers}}{\#\text{ of answers with word}}$$

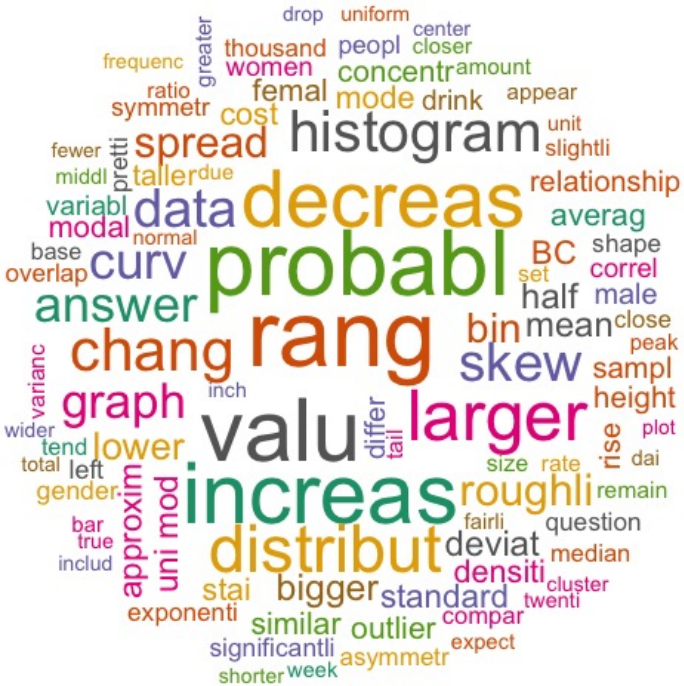
e.g. "the" would have a very low TF-IDF value

[1] [Salton and McGill, 1986]

Word cloud comparison for graphs

Create word clouds using TF-IDF values from answers where students made histograms compared to boxplots

Histograms



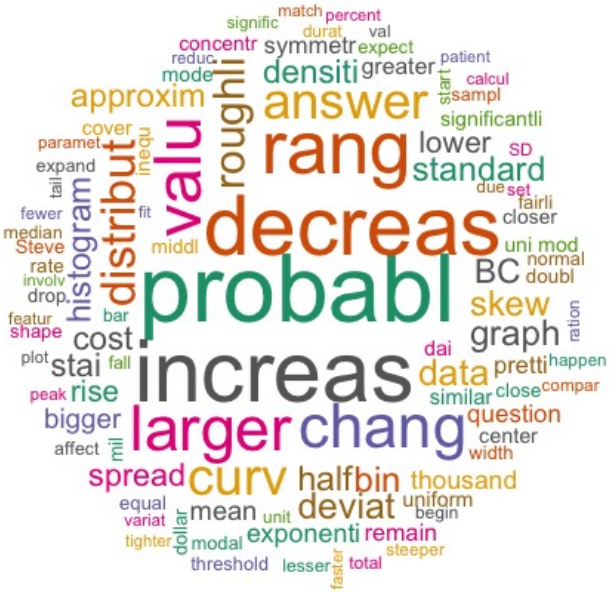
Boxplots

Change in word cloud during semester

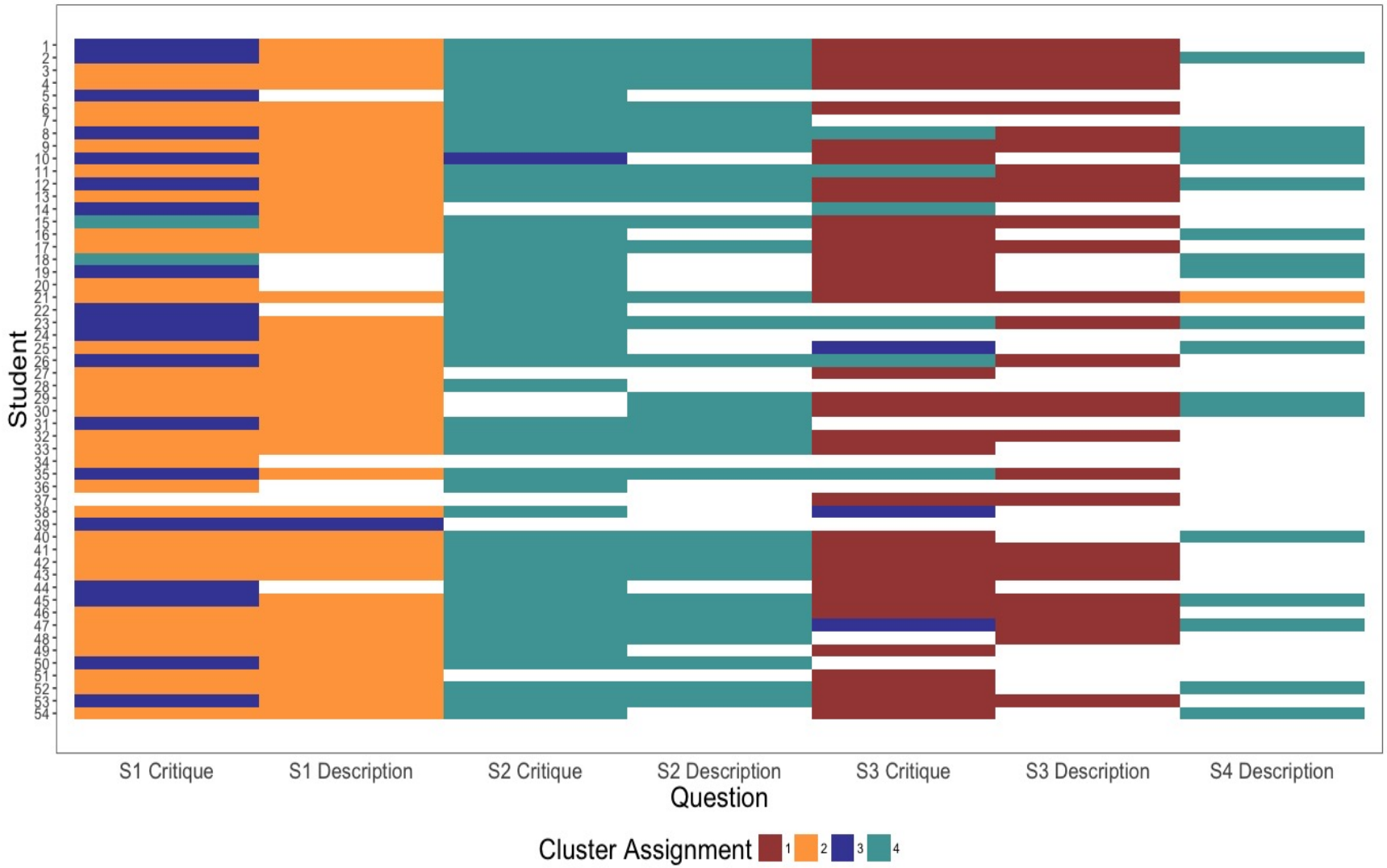
Word choice for histogram answers changed over the semester

First half of semester

Second half



Group students by words with TF-IDF (spherical k-means[2])



[2] [Dhillon and Modha, 2001]

Example of difference in answers for S1 Description

Student 1's answer:

> *"The mean absences for urban students is greater than the mean absences for rural students, but the variance is also significantly higher for urban students than for rural students. Therefore, our results are inconclusive based solely on mean and variance."*

Example of difference in answers for S1 Description

Student 1's answer:

> *"The mean absences for urban students is greater than the mean absences for rural students, but the variance is also significantly higher for urban students than for rural students. Therefore, our results are inconclusive based solely on mean and variance."*

Meanwhile, Student 39's?

Example of difference in answers for S1 Description

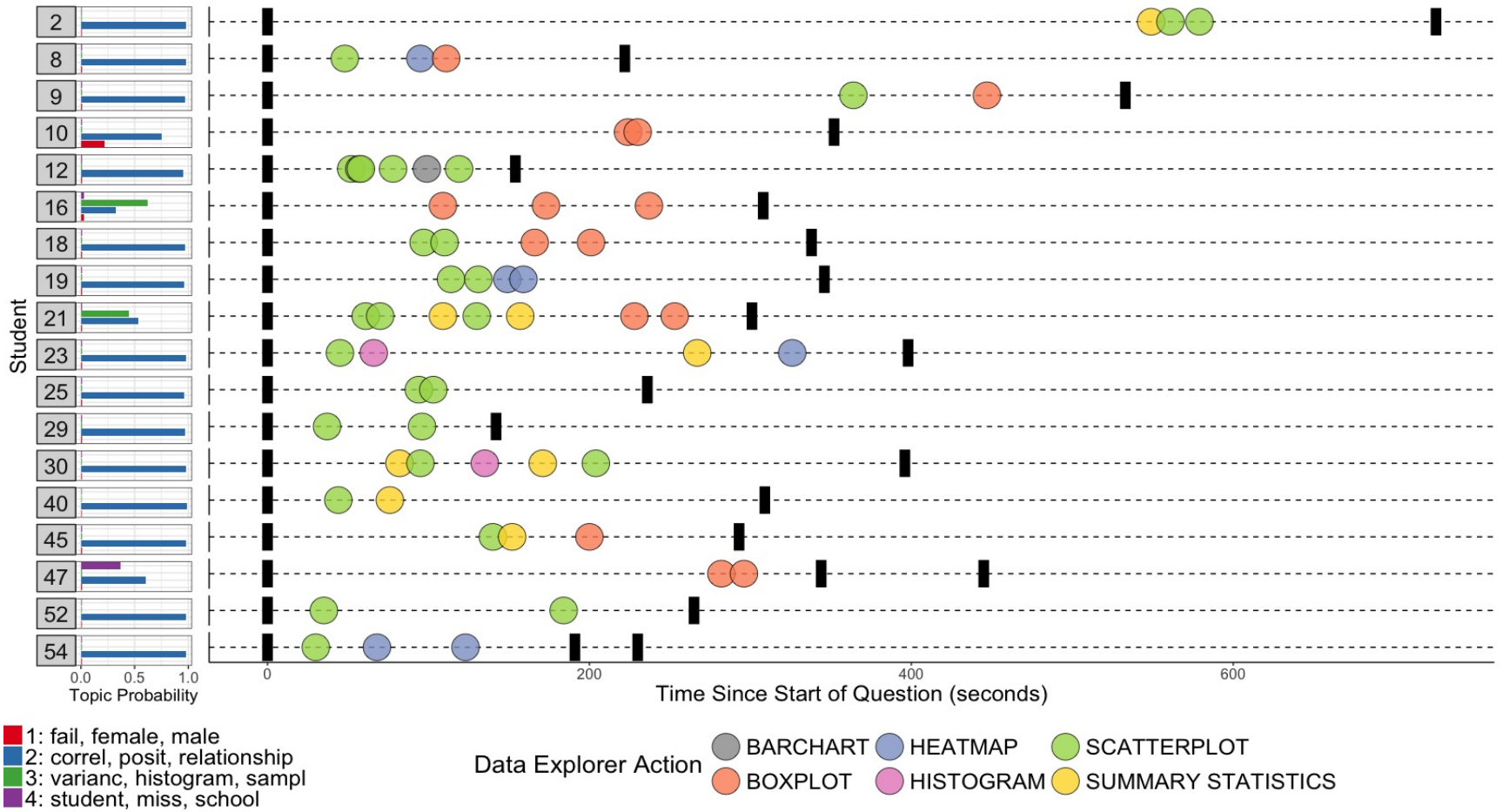
Student 1's answer:

> *"The mean absences for urban students is greater than the mean absences for rural students, but the variance is also significantly higher for urban students than for rural students. Therefore, our results are inconclusive based solely on mean and variance."*

Meanwhile, Student 39's?

> *"Skipped for time."*

Link topic modeling of student answers to the timeline of their actions to understand why they answered differently



This platform will provide us with new insight into statistical pedagogy and how students learn

CMU IRB gives us access to actions and text after semester is over

Over 14,000 actions made by 71 students in Fall '17 semester

Not just right or wrong, but full text of answers -
gaining insight into the process of student thinking

This platform will provide us with new insight into statistical pedagogy and how students learn

CMU IRB gives us access to actions and text after semester is over

Over 14,000 actions made by 71 students in Fall '17 semester

Not just right or wrong, but full text of answers -
gaining insight into the process of student thinking

Action logs all real-time checking of student analysis

Can lead to greater understanding of reproducible research

Check out Phillip Burckhardt's poster P3-01 for more info on ISLE

Thanks and References


Aggarwal, C. C. (2018).
Machine Learning for Text.
Springer International Publishing, 1 edition.


Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003).
Latent dirichlet allocation.
J. Mach. Learn. Res., 3:993–1022.


Dhillon, I. S. and Modha, D. S. (2001).
Concept decompositions for large sparse text data using clustering.
Machine Learning, 42(1):143–175.


Salton, G. and McGill, M. J. (1986).
Introduction to Modern Information Retrieval.
McGraw-Hill, Inc., New York, NY, USA.