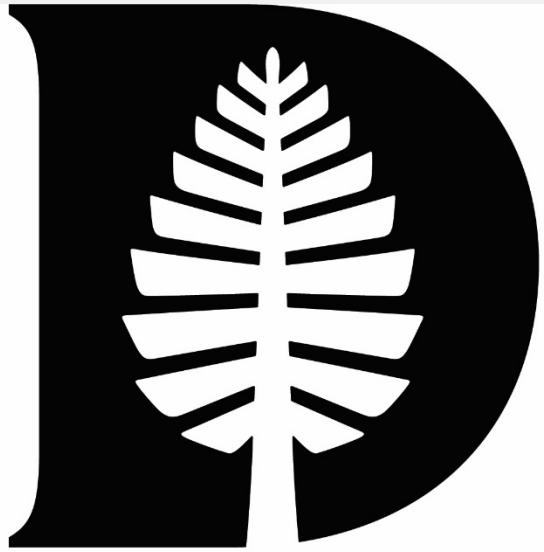

DARTMOUTH

Rise of the Machines:
Text Mining Survey Comments

HEIR Conference, September 2018

Alicia Betsinger, Associate Provost of Institutional Research

Objectives

- To understand survey context at a small private institution.
- To review text mining options on open-ended comments, including sentiment analysis and topic modeling.
- To discuss how other institutions are leveraging machine-learning techniques with their open-ended survey comments.

Dartmouth College

- 4-year, private, Ivy League institution, located in Hanover, New Hampshire
- Primarily residential campus for traditional-aged students (18-22)
- Undergraduates = 4,400
- Graduate/Professional = 2,200
  - Guarini School of Graduate and Advanced Studies
  - Geisel School of Medicine
  - Thayer School of Engineering
  - Tuck School of Business

Surveys at Dartmouth

- Member of consortium of private institutions. Suite of student surveys intended to be examined longitudinally.
  - New Student
  - Enrolled Student
  - Senior
  - Alumni
- Internal surveys, most often for Student Affairs.
  - Advising
  - Health, including Sexual Misconduct
  - House Communities

Structured vs. Open-ended

	Open-ended	Structured (close-ended)
Pros	• Responses are from respondents' perspective/own words • "Top-of-mind" • Gain new insights (unanticipated results)	• Quick and easy for respondents to answer • Easy to quantify/tabulate • Easy to make comparison of results
Cons	• More burden on respondents (60% will not answer) • Hard to interpret/quantify (coding)	• Need to have the right choices

Qualitative Insights **Quantitative Insights**

Traditional Open-ended Approach

Listing of all comments.

Redaction considerations.

Q23. What was the biggest challenge you encountered in planning your curriculum and associated academic plans?
N/A
Not much
scheduling
The fact that classes are not planned/ tentatively planned until we graduate so I had to guess which class I might be able to take my junior and senior years. This was an issue due to how light my schedule is up until my senior spring.
looking ahead for classes
Still have not declared
I still don't have it planned. Thayer is so confusing and my advisor is not knowledgable about it and did not refer me to anyone who could help.
Balancing D-Plan (on/off-terms) with classes of interest
Would've been nice to see generally what classes might be offered in the years to come
Fitting in other degrees
Finding available classes offered only in certain terms
Department websites don't show all the classes that will be offered until I graduate so it's hard to anticipate what I'll be taking to put into my major worksheet.
Finding time to study abroad and get credit and still take classes of interest
Classes not listed many terms out
Don't know who to meet with.
Lack of knowledge of class schedule
Too many requirements to think about
There are many interesting classes, and sometimes it is hard to decide which one I prefer.
nobody helped me and i still don't know what im doing
The process was tedious.

Three Examples

- Community Study: Campus Climate Survey
  - Text mining & Topic modelling
- Senior Surveys
  - Text mining & Thematic analysis
- Finance Survey
  - Text mining & Sentiment analysis

Community Study

- Campus Climate survey conducted in 2015. Returned to open-ended comments as part of Inclusive Excellence efforts in AY 17-18.
  - Do you have specific **recommendations for improving the climate** at Dartmouth?
- Hired outside consultant:
  - Survey design in conjunction with campus working group
  - Survey administration
  - Survey analysis: quantitative & qualitative

Text Mining

- Read/Load documents: Word vs. Excel
- Process Documents
  - **Tokenize**: Breaking up a sequence of strings into pieces such as words, keywords, and phrases.
    - Non letters
    - Linguistic sentences
  - **Filter Stopwords**: filter out common words (e.g., a, the, is, are, etc.) and custom stopwords (e.g., Dartmouth, dartmouth)
  - **Transform Cases**: all lower case
  - **N-grams**: Simply a sequence of N word. Co-occurring words.

Text Mining (cont.)

- Create Wordlists
- FP-Growth and Association
  - **FP-Growth**: Frequent pattern growth. Algorithm for calculating frequently co-occurring items.
    - Default: Highest support items at top
    - Sort: Ability to sort by any items set
  - **Association**: If-then statements that help uncover relationships.
    - Sort: Highest confidence at top
    - Graph: Visual representation of associations

Topic Modeling

- Want to learn something about a large amount of text (corpus) that's too big to read.
- Intended to go beyond word item sets. Discover abstract "topics" or hidden semantic structure.
  - **Supervised & unsupervised**
- Unsuccessful in this example. Found no clear pattern due to relatively small number of comments (~900) and wide variation in the messages.
  - To detect patterns, either number of comments should be large or the number of topics should be small.

2016 & 2018 Senior Surveys

2016	2018
Atlas.ti	Atlas.ti & RapidMiner
Partnered with qualitative researchers in teaching and learning area.	Only IR researchers
Five open-ended comments. Lead analyst and then reliability check. Total of 6 individuals	Four open-ended comments. Total of 2 individuals
Time and labor-intensive	Less time and labor-intensive and discovered similar patterns to 2016.

Thematic Analysis: Senior Surveys

- Content analysis to identify core meanings through patterns or themes.
- Concept coding:
  - **RapidMiner:** Provided initial scan of data and scope of responses.
  - **Qualitative:** Subsumed codes into broader codes or categories. Re-coding &/or recategorizing during analysis.
    - Many rounds of coding → themes emerged, solidified, and gained a rounded sense of what they encompassed.
    - Meetings to discuss categories and themes. Continued until arrived at agreement.
    - Code map was drawn.

Finance Survey

- Internal survey conducted every three years for Finance & Administration area.
- Conducted same text mining using RapidMiner for open-ended comment but then utilized AYLIEN add-on for sentiment analysis.
- Need API key. Only set a few parameters.
- Helpful to focus clients on more than just negative comments. Polarity (negative, neutral, positive) and polarity confidence.

Summary & Discussion

- Text mining saved time but higher-order categorization is still elusive.
  - Responses to survey prompt perhaps limits unsupervised techniques?
  - Combine open-ended comments across time.
- Natural Language Processing (NLP) will continue to evolve and IR professionals need to determine how these techniques can benefit survey research efforts in our office.
- How has your office utilized any of the noted techniques?
  - Text mining, topic modeling, sentiment analysis.

Thank You

Alicia.M.Betsinger@dartmouth.edu