2021

Self-Help and Business Development Media

FINDING PRACTICAL APPLICATIONS

ANDREW LINDBERG

Abstract

Amazon claims over 70,000 results for books in the category of self-help and over 80,000 results in business and money. Aspiring entrepreneurs, business leaders, and individuals seeking *remarkable results* or *success* in their personal or professional life would be silly to disregard the wisdom of prominent economists, psychologists, and individuals who have achieved the success that they aspire to achieve.

The sheer volume of this content has created the possibility for other media forms and industries to be built out of this quest for knowledge - the most prominent of which being podcasting. There are shows like *How To Be Awesome At Your Job*, where host Pete Mockaitis interviews authors about the content of their books; and other shows, where thought leaders like Wharton professors, Adam Grant and Katy Milkman, explore the foundational science behind behavior and behavior change that may support success (or failure).

Unfortunately, a lot of this content can seem like common sense, sharing themes like the importance of building good habits, or it can be redundant, when the same guest speaker shares the same ideas on different shows. There are many tools for information seekers, like Blinkist that seek to maximize the amount of information gained per unit time by summing the key points and delivering the message in a quicker format.

This project applies natural language processing techniques to transcripts from the podcast, *Dare to Lead*. In doing so, it aims to identify common themes and novel ideas that may have practical applications in business leadership.

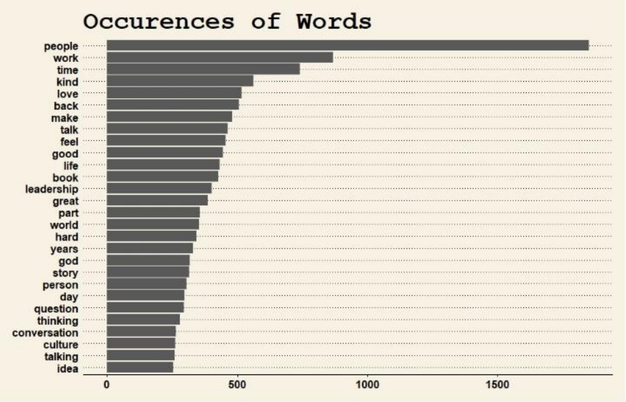
Data

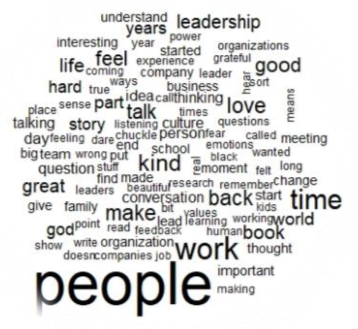
The *Dare to Lead* podcast posts copies of their transcripts on their website. This information was scraped using the RSelenium and rvest packages. Rselenium allows users to access information from websites that is gated by javascript encoding, and rvest was used to read html and isolate transcript text. 32 episode transcripts were gathered including the titles, host, guest speaker names, website urls, and publication dates. One episode was removed because it was a trailer and offered no semantic value.

The tidyverse suite of packages, forcats, and reshape2 were used to organize and structure the data in a tibble. The lubridate and zoo packages were used to clean and process date information. Ggplot2, ggthemes, igraph, ggraph, widyr, and wordclouds were used to convert tabular data into graphical representations. Tidytext, tm, topicmodels, and ldatuning were used to process text data and build natural language processing models.

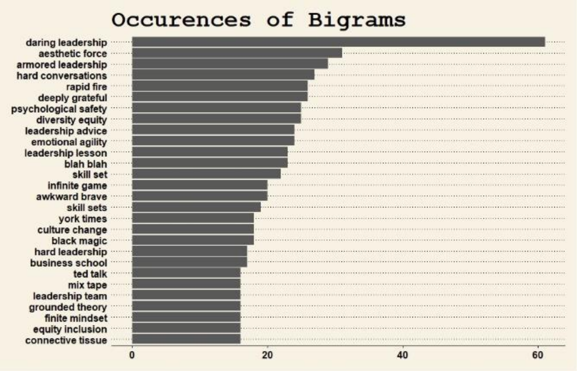
Results

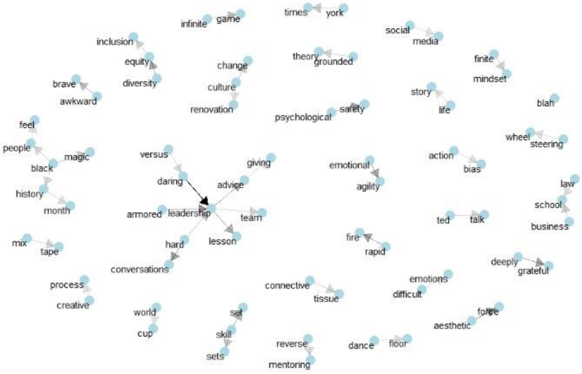
The data was tokenized into one word per line, tidytext format, stopwords were removed and a bag of words model was used to start exploratory data analysis.





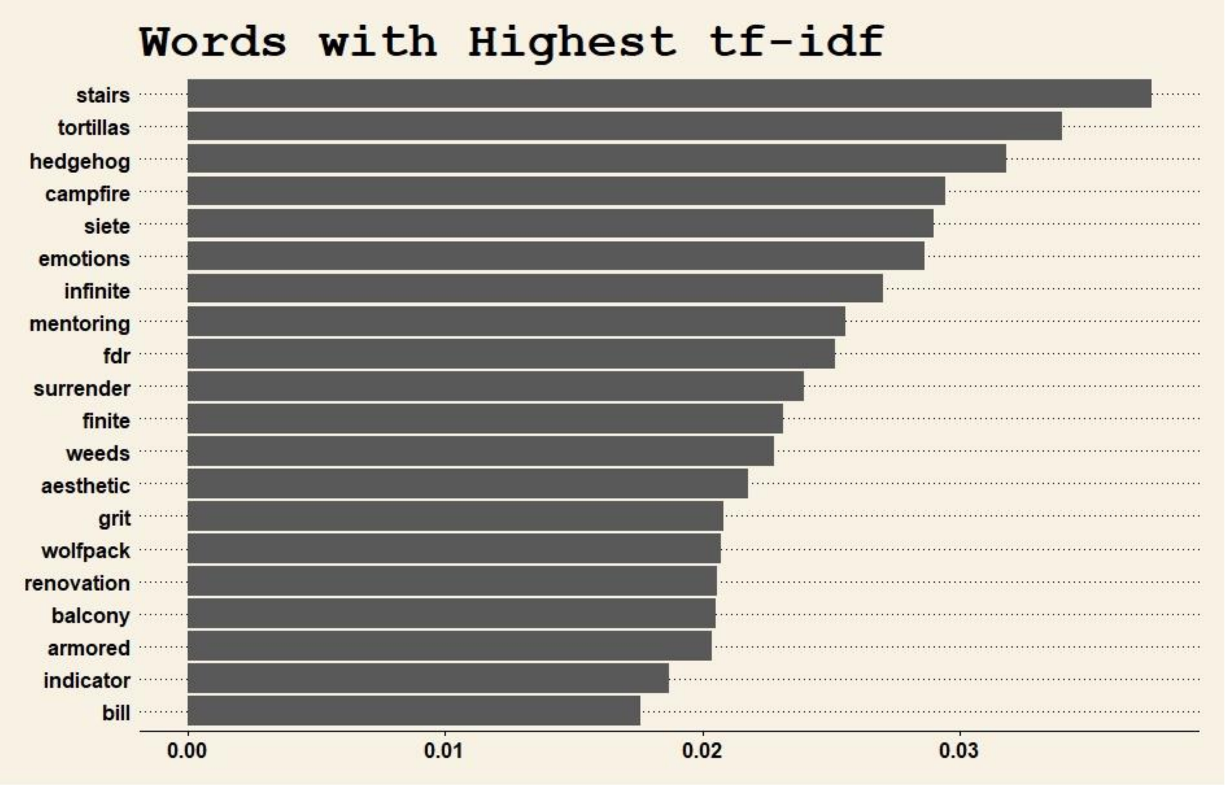
While this validates that the show is primarily about people, work, and time – little insight can be gained from this model. Maintaining the bag of words approach and separating by bigrams offers a little more interesting insight.



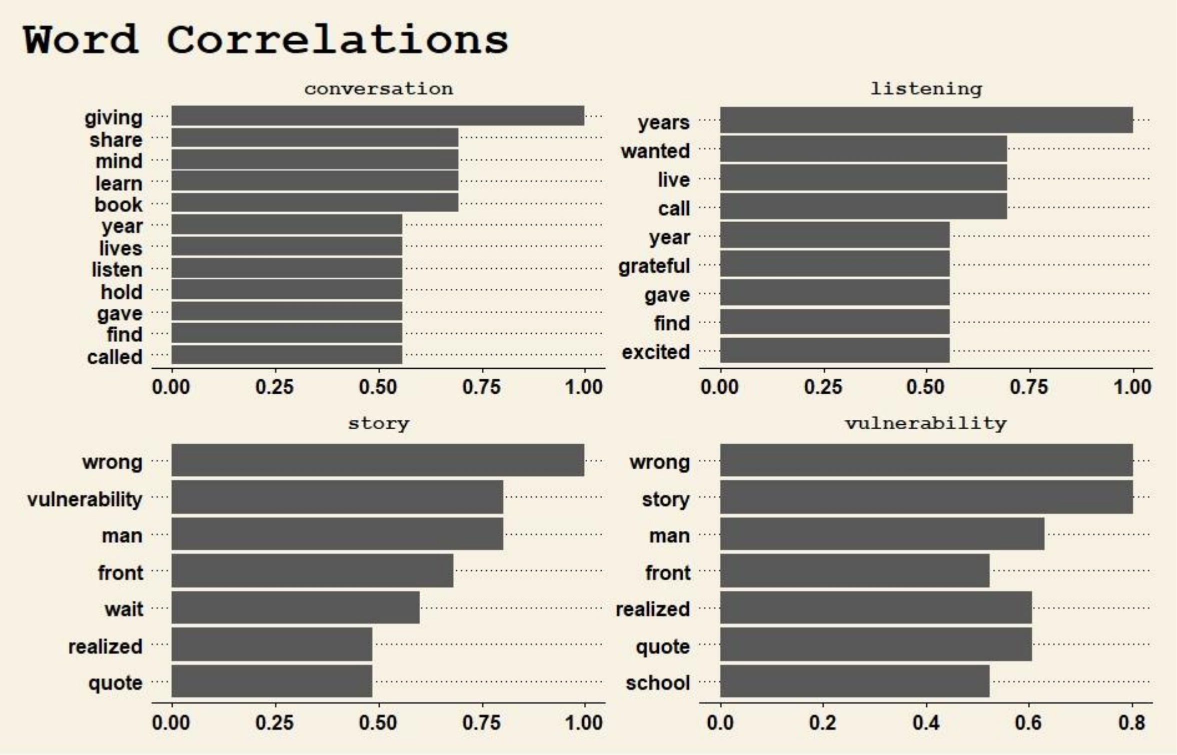


In particular, the directional bigrams shows some common subjects like psychological safety, difficult emotions, and being deeply grateful. Avid followers of Brene Brown are likely be able to make these deductions; but individuals lacking the meta knowledge might get a better understanding of the show from these bigrams.

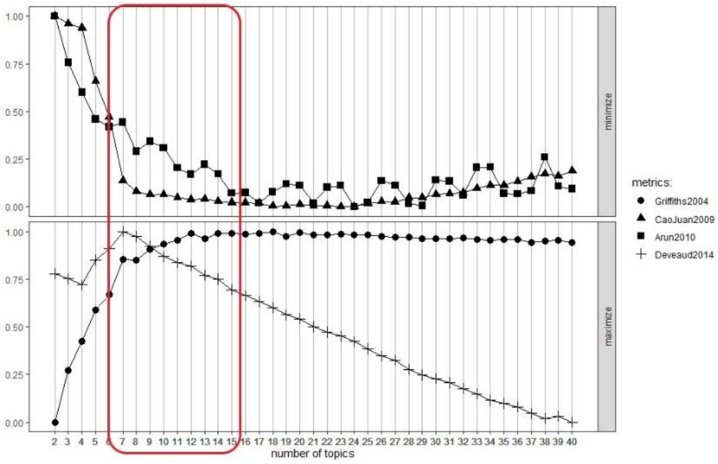
In the quest for deeper, more meaningful insights, we considered the words with the highest tf-idf (term frequency-interdocument frequency). This model tries to identify the most impactful words across the corpus. Some words make sense as they're uniquely tied to the subject of some episodes. For example, one episode interviews the founders of Siete Foods who make specialty tortillas. Emotions and mentoring make sense for the corpus. Stairs and weeds raise more questions.



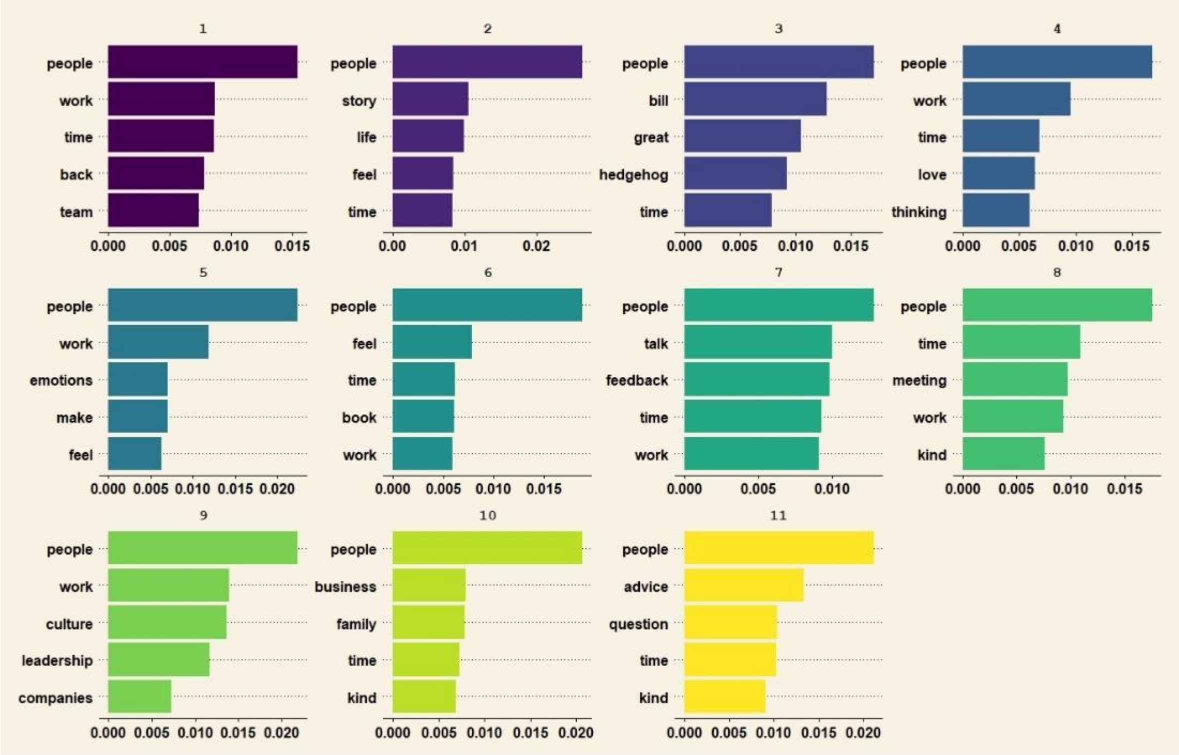
Word correlation models were used with the words “conversation”, “listening”, “story”, and “vulnerability”. These words were chosen due to the subject matter of the corpus and offer some interesting relationships. Conversation is related to giving and share. Story, vulnerability, and wrong are also very related.



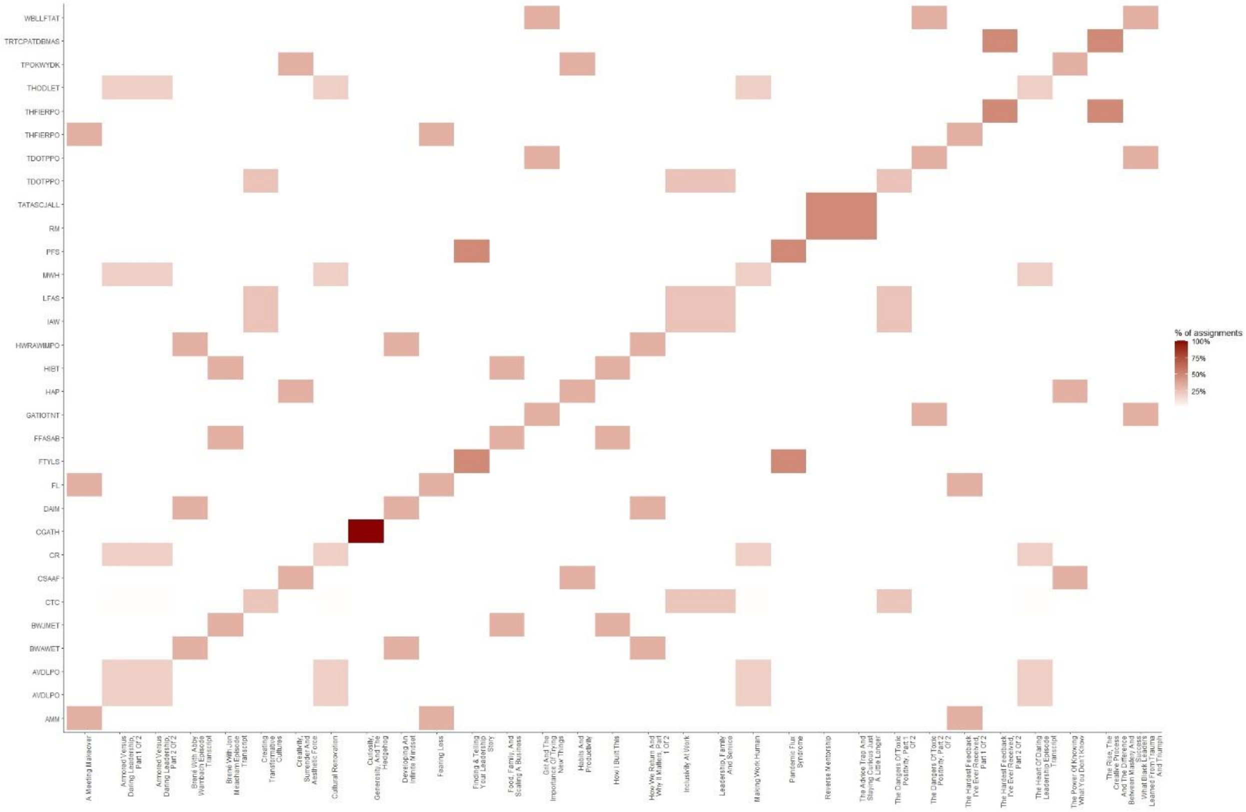
While there are 31 episodes, this research seeks to gain insight regarding the topics. As such, empirical processing was used to identify the ideal number of topics across the corpus. Based on the Griffiths2004, CaoJuan2009, Arun2010, and the Devaud2014 models, somewhere between 7 and 15 topics is ideal.



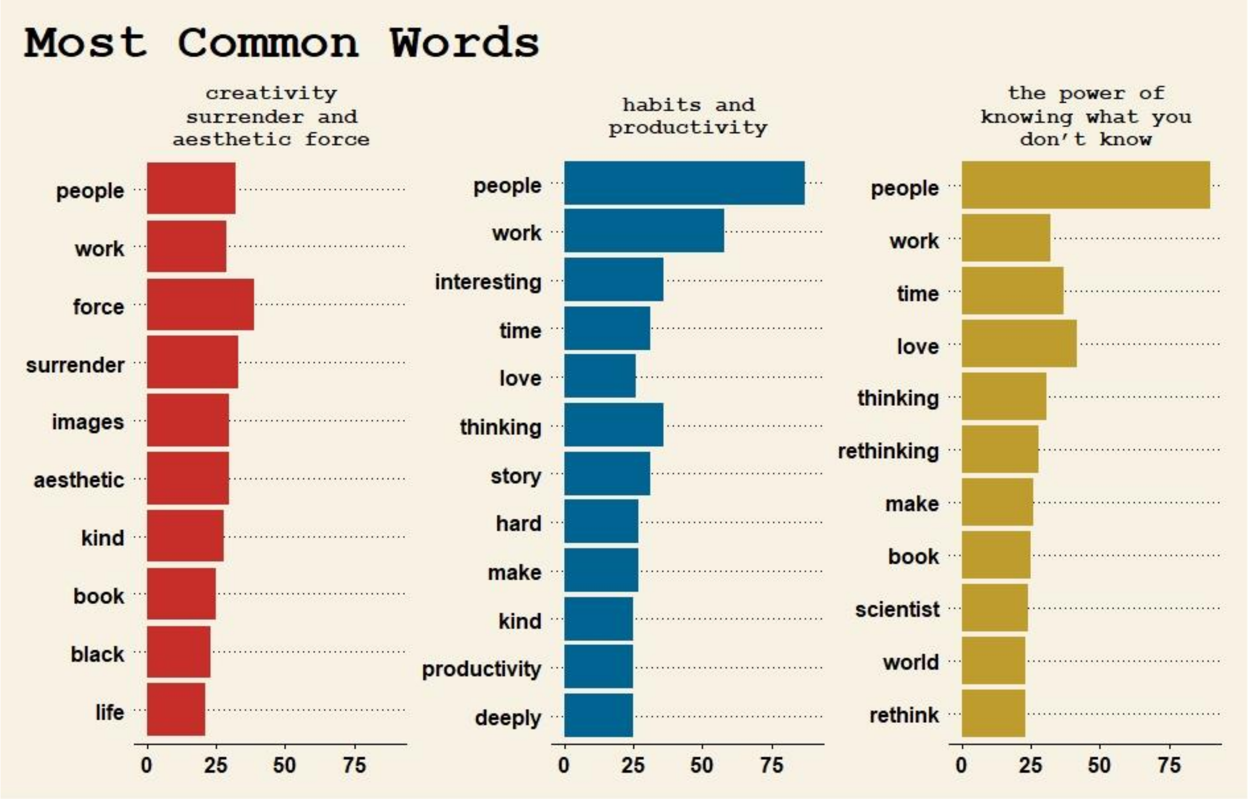
Using 11 models, the following groups of words represent each topic.



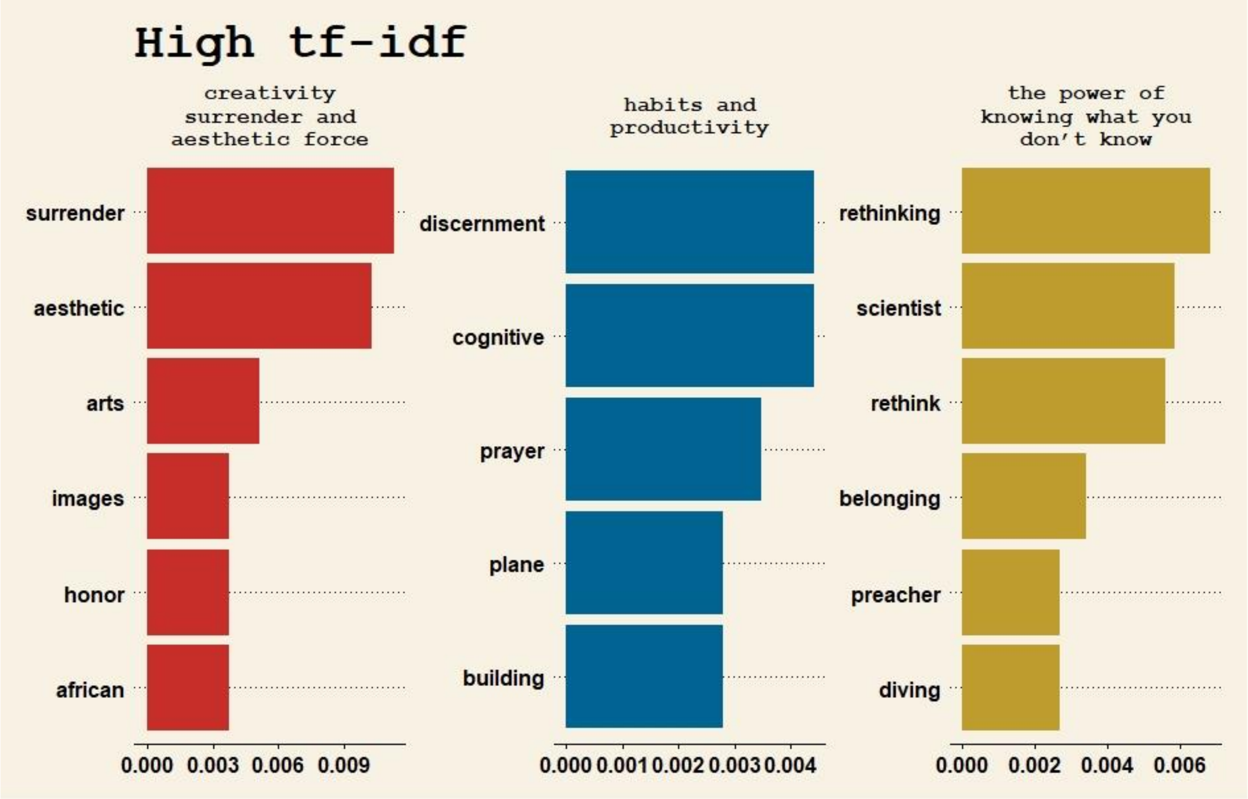
Using the 11 models, the words from the document on the y-axis were assigned to the documents on the x-axis. The episodes, “Creativity, Surrender, and Aesthetic Force,” “Habits and Productivity”, and “The Power of Knowing What You Don’t Know” were the most mislabeled between each other.



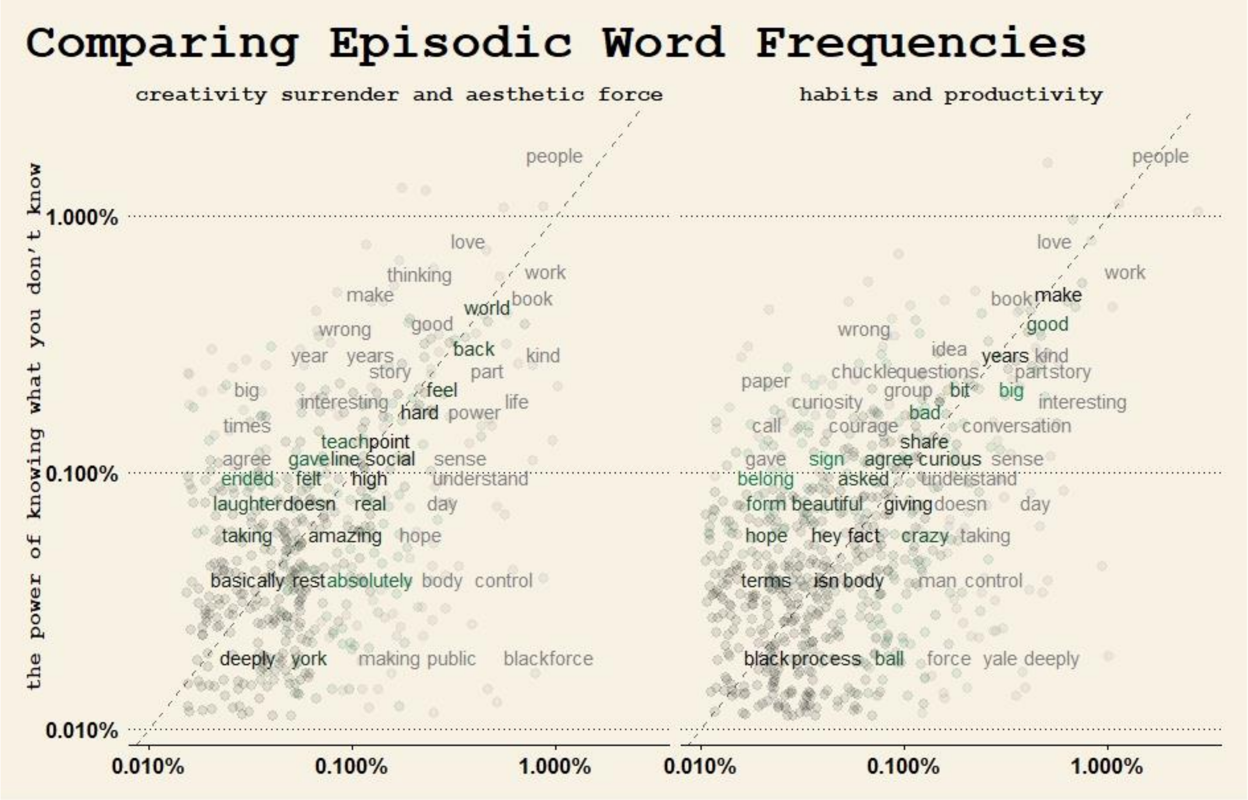
Isolating these episodes and applying bag of words models, the most common words per episode show the similarities between the three.



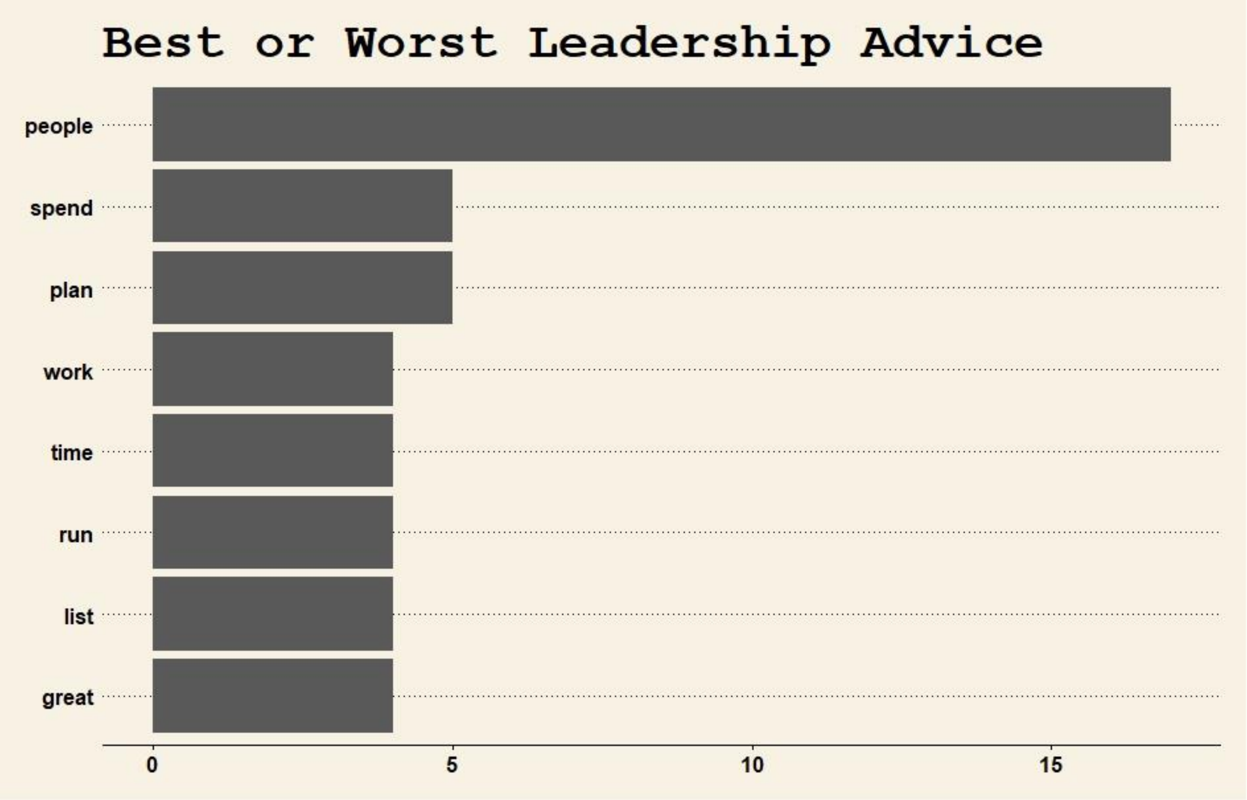
The highest TF-IDF words show the differences between the episodes.

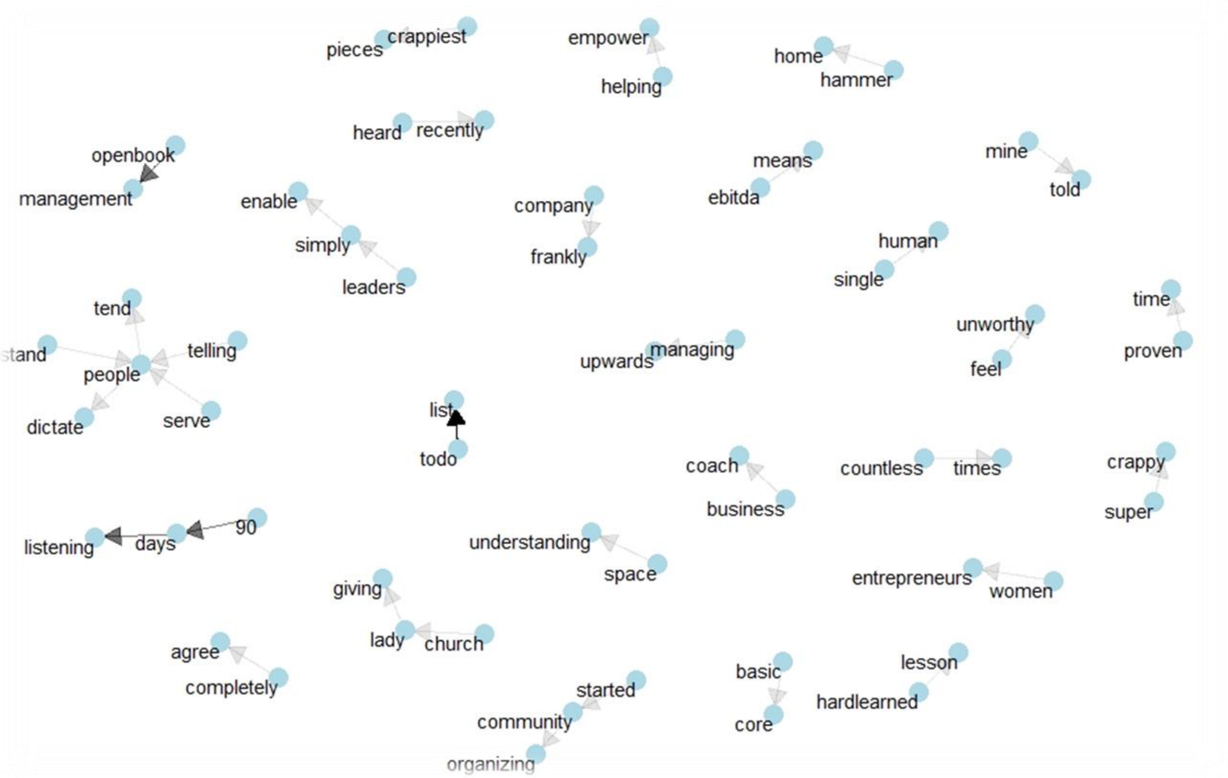


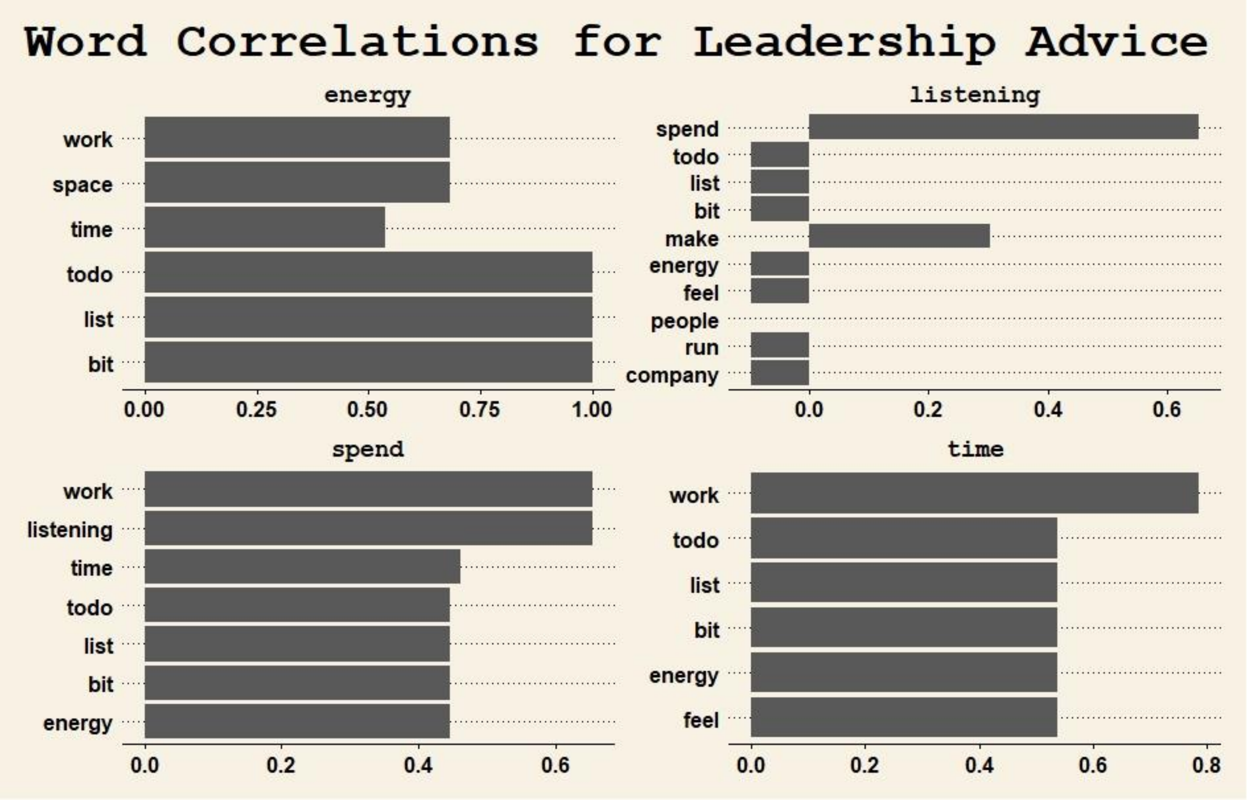
Using word frequencies, we can see more of the similarities between the episodes.



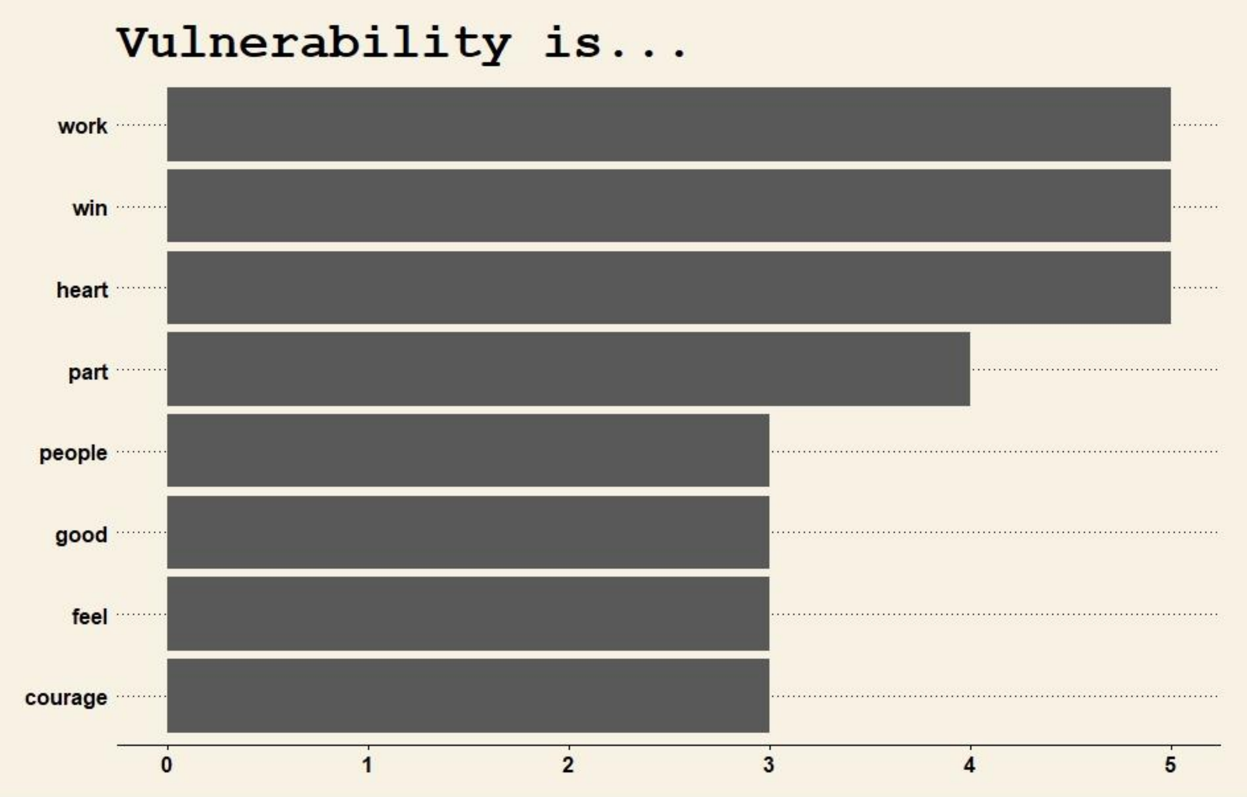
Finally, we dug into the rapid-fire questions that Brene asks every guest at the end of each episode. One question asks the guest for the best leadership advice they've ever received or advice that's so bad, they need to warn listeners. Bag of words, word correlations and bigrams were applied to the next response from a guest.

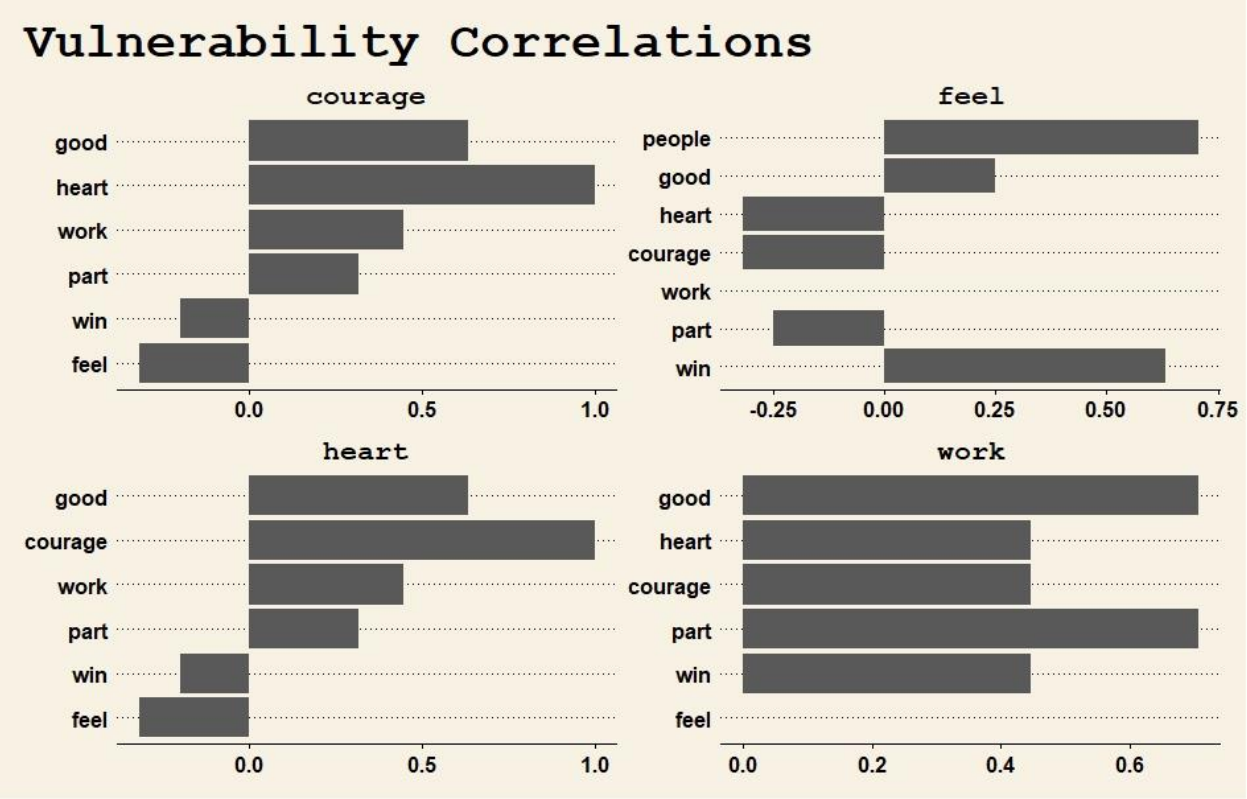


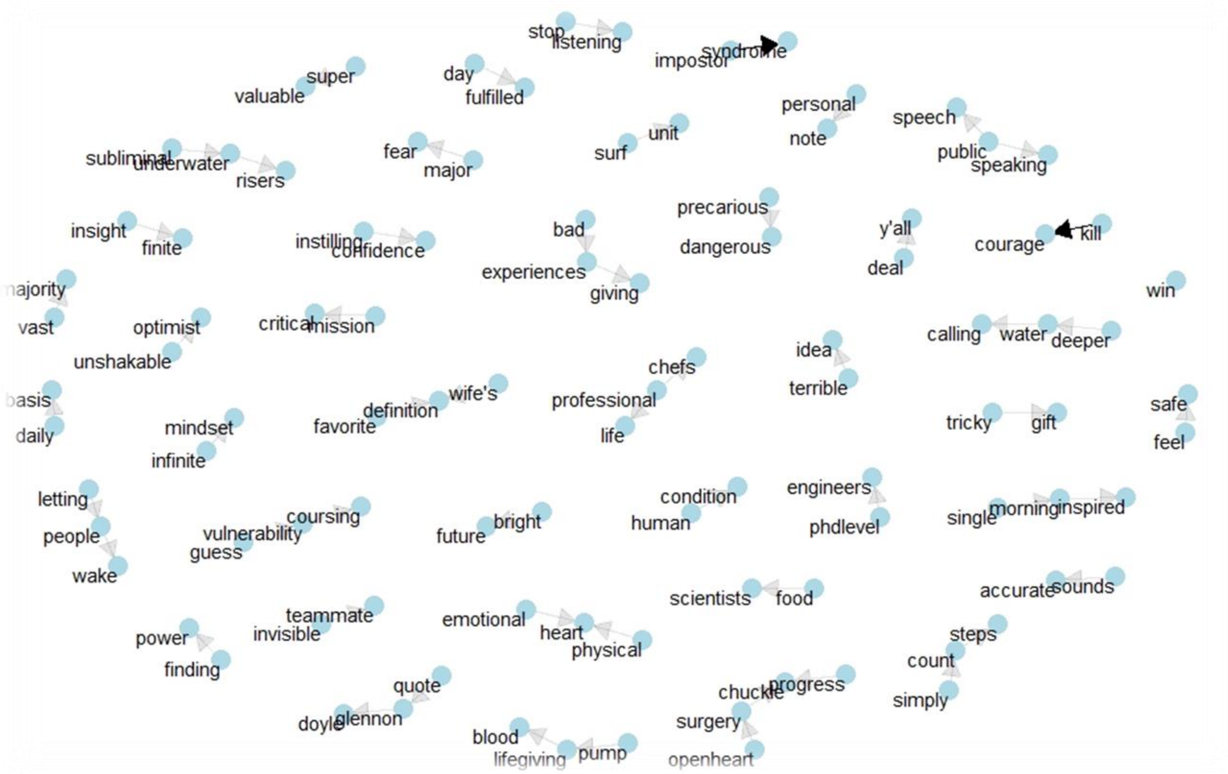

pieces
crappiest
empower
helping
home
hammer
heard
recently
openbook
management
enable
simply
leaders
company
frankly
means
ebitda
mine
told
human
single
tend
stand
telling
people
dictate
serve
upwards
managing
unworthy
feel
time
proven
list
todo
coach
business
countless
times
crappy
super
listening
days
90
understanding
space
entrepreneurs
women
giving
lady
church
agree
completely
started
community
organizing
basic
core
lesson
hardlearned


Word Correlations for Leadership Advice
energy
work
space
time
todo
list
bit
0.00
0.25
0.50
0.75
1.00
listening
spend
todo
list
bit
make
energy
feel
people
run
company
0.0
0.2
0.4
0.6
spend
work
listening
time
todo
list
bit
energy
0.0
0.2
0.4
0.6
time
work
todo
list
bit
energy
feel
0.0
0.2
0.4
0.6
0.8

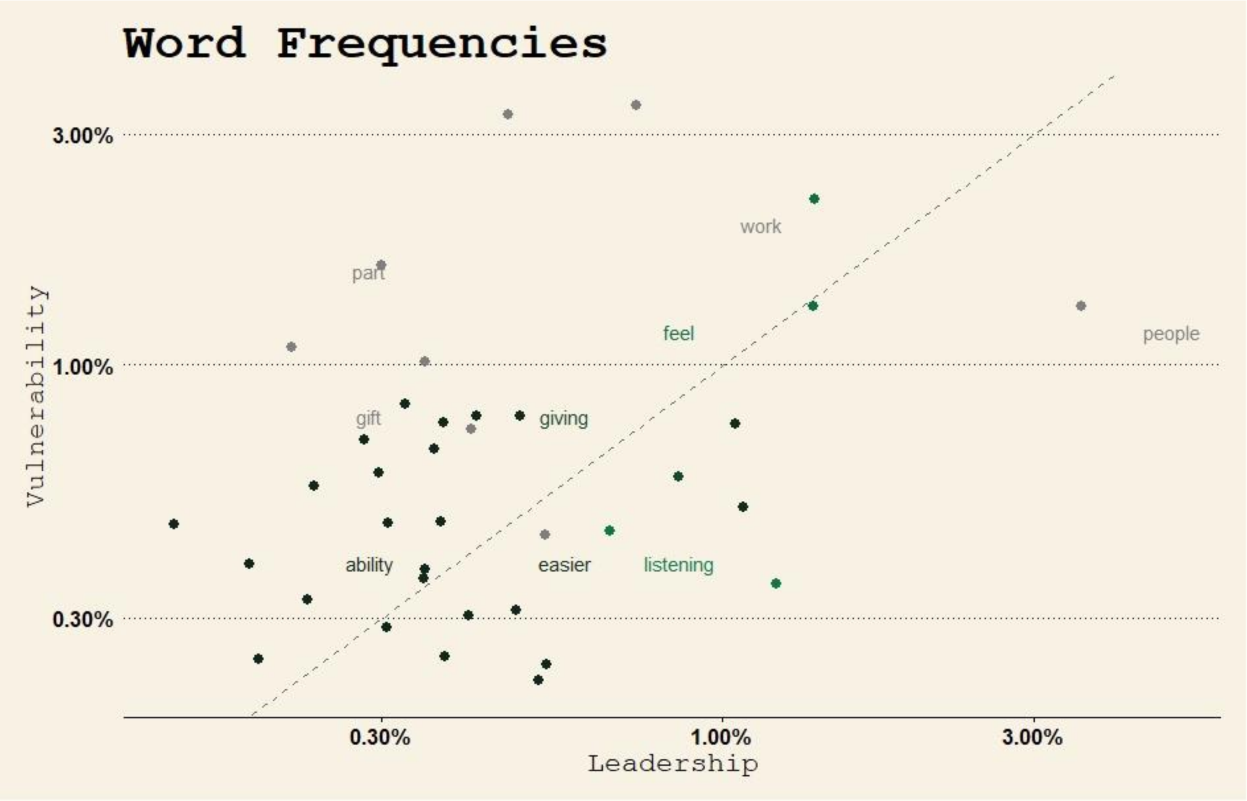
The same approaches were applied to the question, “Vulnerability is…”






stop
listening
impostor
syndrome
super
valuable
day
fulfilled
personal
note
speech
public
speaking
subliminal
underwater
risers
fear
major
surf
unit
precarious
dangerous
insight
finite
instilling
confidence
bad
experiences
giving
y'all
deal
courage
kill
majority
vast
optimist
unshakable
critical
mission
win
idea
terrible
calling
water
deeper
chefs
professional
life
basis
daily
mindset
infinite
definition
wife's
favorite
tricky
gift
safe
feel
letting
people
wake
coursing
vulnerability
guess
bright
future
condition
human
engineers
phdlevel
single
morning
inspired
scientists
food
accurate
sounds
teammate
invisible
emotional
heart
physical
power
finding
steps
count
simply
quote
glennon
doyle
chuckle
progress
surgery
openheart
blood
lifegiving
pump

Finally, word frequencies were compared between the answers for both questions.



Limitations & Next Steps

This research was limited by the data processing, the amount of the data, and the skills of the researcher. In future research, the word “people” should be removed to hopefully isolate more distinct topics. Several episodes are monologues from Brene Brown, and given the conversational nature of the show, analysis of just the guest responses is likely to be less skewed towards the hosts perspective.

The answers to the rapid-fire questions seeks to indulge this idea; however, there isn’t enough data. Next steps would be to further explore this dataset until satisfactory models can be built, and then compare models to other podcasts. Specifically, topic modelling from this show compared to other shows with similar subject matter could be very interesting.

Conclusions

This project succeeded in starting the exploratory data analysis necessary to answer the initial research question: What common themes and novel ideas can be learned from self-help and business development podcasts?

Unfortunately, most of the models offered little or no insight into the *Dare to Lead* podcast beyond what could be gathered from an abstract description of the show. Bag of words modeling using bigrams and topic models seemed to be the most effective in offering novel answers to the initial research question.

Ultimately, the most time consuming and restrictive part of this process was manipulating the data so that we could more easily filter, sort, and label different content. While I did learn a lot through the process, I learned more about what I still need to learn. Many functions seem intuitive, especially when applied to pre-processed, readily available datasets.

This dataset was particularly convenient because each speaker section was easily identifiable with a simple code and was frustrating because of UTF-8 encoding and small inconsistencies between episodes. I applied many different strategies that worked for this corpus but are not likely to work if the content were expanded with new episodes over time and would not scale beyond this show.

As I continue to learn, I hope to better understand the underlying behavior of packaged functions, and more effective techniques that would both be less time intensive and more scalable.

References & Packages

https://brenebrown.com/podcast-show/dare-to-lead/

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing,
Vienna, Austria. URL https://www.R-project.org/.

Silge J, Robinson D (2016). “tidytext: Text Mining and Analysis Using Tidy Data Principles in R.” _JOSS_, *1*(3). doi: 10.21105/joss.00037 (URL: https://doi.org/10.21105/joss.00037), <URL: http://dx.doi.org/10.21105/joss.00037>.

Wickham et al., (2019). Welcome to the tidyverse.
Journal of Open Source Software, 4(43), 1686,
https://doi.org/10.21105/joss.01686

Ingo Feinerer and Kurt Hornik (2020). tm: Text Mining
Package. R package version 0.7-8.
https://CRAN.R-project.org/package=tm

Ingo Feinerer, Kurt Hornik, and David Meyer (2008).
Text Mining Infrastructure in R. Journal of Statistical
Software 25(5): 1-54. URL:
https://www.jstatsoft.org/v25/i05/.

Garrett Grolemund, Hadley Wickham (2011). Dates and
Times Made Easy with lubridate. Journal of Statistical
Software, 40(3), 1-25. URL
https://www.jstatsoft.org/v40/i03/.

Ian Fellows (2018). wordcloud: Word Clouds. R package
version 2.6.
https://CRAN.R-project.org/package=wordcloud

H. Wickham. ggplot2: Elegant Graphics for Data
Analysis. Springer-Verlag New York, 2016.

Hadley Wickham and Dana Seidel (2020). scales: Scale
Functions for Visualization. R package version 1.1.1.
https://CRAN.R-project.org/package=scales

Jeffrey B. Arnold (2021). ggthemes: Extra Themes,
Scales and Geoms for 'ggplot2'. R package version
4.2.4. https://CRAN.R-project.org/package=ggthemes

Achim Zeileis and Gabor Grothendieck (2005). zoo: S3

Infrastructure for Regular and Irregular Time Series. Journal of Statistical Software, 14(6), 1-27. doi:10.18637/jss.v014.i06

Hadley Wickham (2021). forcats: Tools for Working with Categorical Variables (Factors). R package version 0.5.1. https://CRAN.R-project.org/package=forcats

Hadley Wickham (2007). Reshaping Data with the reshape Package. Journal of Statistical Software, 21(12), 1-20. URL http://www.jstatsoft.org/v21/i12/.

Csardi G, Nepusz T: The igraph software package for complex network research, InterJournal, Complex Systems 1695. 2006. https://igraph.org

Thomas Lin Pedersen (2021). ggraph: An Implementation of Grammar of Graphics for Graphs and Networks. R package version 2.0.5. https://CRAN.R-project.org/package=ggraph

David Robinson (2021). widyr: Widen, Process, then Re-Tidy Data. R package version 0.1.4. https://CRAN.R-project.org/package=widyr

Grün B, Hornik K (2011). “topicmodels: An R Package for Fitting Topic Models.” _Journal of Statistical Software_, *40*(13), 1-30. doi: 10.18637/jss.v040.i13 (URL: https://doi.org/10.18637/jss.v040.i13).

Murzintcev Nikita (2020). ldatuning: Tuning of the Latent Dirichlet Allocation Models Parameters. R package version 1.0.2. https://CRAN.R-project.org/package=ldatuning