TOPIC MODELING ON MARKETING PODCAST

Team 8:
Ashley Keung, Billy Choy, Boli Qiao, Jonathan Chu, Juan Figini

AGENDA

https://podcasts.apple.com/lb/podcast/culture-ate-my-brand-improving-company-culture-results/id373596600?i=1000542336192

1. Project Introduction
2. Data Collection And Transformation
3. Topics Extraction And Descriptive Analysis
4. Subtopics Extraction
5. Conclusion

Part 1: Introduction Of Project

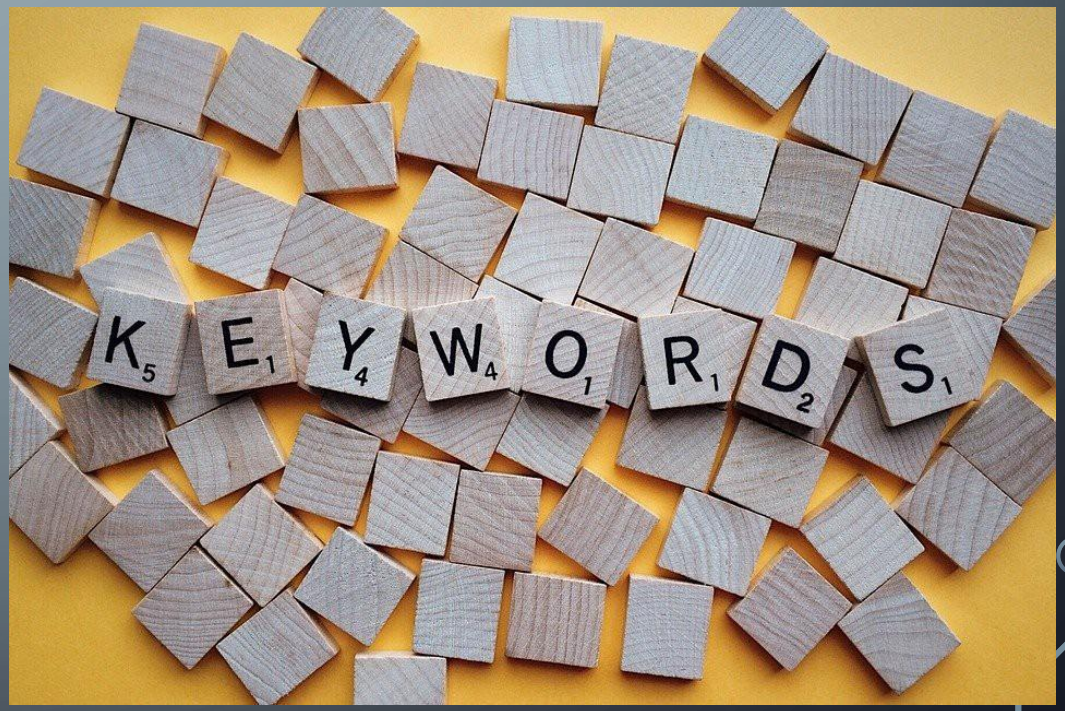
Introduction : Project Objectives

1. Perform **topic modelling** to identify key **topics** and **subtopics** from the digital marketing podcast.
2. Compare our topics and subtopics with **podcast summaries** (human knowledge).

https://www.kdnuggets.com/2019/11/understanding-nlp-topic-modeling-part-1.html

Introduction:
How Does Topic Modeling (Key Phrase Extraction) Work ?

1. Lemmatize Text
2. Select Potential Phrases
3. Score Each Phrase



https://towardsdatascience.com/keyword-extraction-with-bert-724efca412ea

Exploration on Methods for Topic Modelling:
3 Potential Methods To Do Topic Modeling

METHOD 1: Topic modeling with decision tree

Step1:
Sentiment analysis with Textblob and Vader
Step2:
Structural Topic Modeling with stm package in R, one-vs-rest binarization strategy to deal with sparse matrices issues
Step3:
SMOTE for imbalance data
Further steps:
Decision Tree for further study within the authors' topic which is not relevant to our project

METHOD 2: Structural Topic Model

Step 1:
Data Cleaning
Step 2:
used STM, implemented with the stm R package

METHOD 3: STM with topic aggregation and topic segmentation

Step 1:
Data Preprocessing
- Remove the emoji, link
- Language detection
- Lemmenization
- Filtering for stop

Step 2:
STM by Additive Regularization of Topic Models
Step 3:
Topic aggregation
Step 4:
Topic Segmentation

Introduction: Our Method For This Project

1. Transcribe Podcast Audio into the transcript by using Microsoft Word for Web.
2. Perform data cleansing and conduct descriptive analysis on transcribed files.
3. Extract keywords using Rapid Keyword Extraction (RAKE) Algorithm in Natural Language Processing.
4. Use another package, Gensim and LDA, to perform vectorization and identify subtopics.
5. Based on the finding, cross-check with the podcast summary (human knowledge).

Part 2: Data collection and transformation

Data collection: The podcast episode we selected

32 min PLAY

Culture Ate My Brand - Improving Company Culture & Results

The Digital Marketing Podcast

Marketing

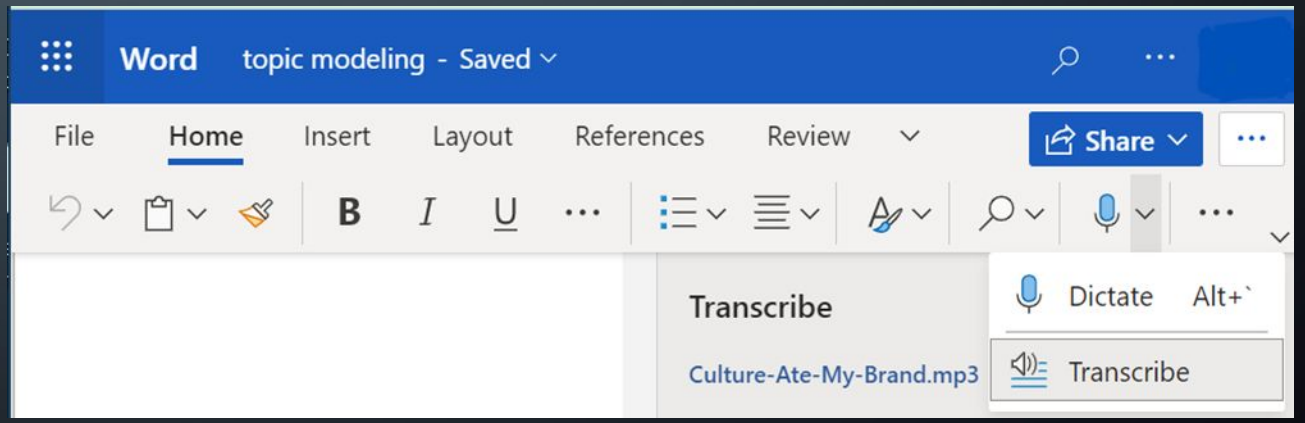
Listen on Apple Podcasts ↗

If culture is the foundation of every successful brand, is your organisation fundamentally doomed? Does your organisation settle for inadequate employee performance, mediocre outcomes, and unremarkable earnings? It doesn't have to be like this. We speak to Mark Miller and Ted Vaughn, the cofounders of Historic Agency about their work as a brand strategy company that helps ambitious brands do more good. They believe in mediocrity's potential for greatness through understanding culture at all levels with methods like measuring it so you can understand how your organisation functions from top-down or bottom-up, or both. The Culture of a CompanyWe explore the concepts in their Book 'Culture Built My Brand'. Discover ways to break through the inertia to engage your team, drive better results, and attract a tribe of loyal customers—by tapping into the greatest driver of brand success: your internal company culture. Passionate about helping ambitious businesses do more good in their communities they love tackling tough problems like how best to measure culture or what actions can lead an organisation towards greatness with its customers at heart. What is Company Culture?Learn how your company culture can help you win more customers and turn them into super fans. Filled with practical steps and case study examples of culture at work, Mark and Ted share the know-how you need to tap into your company culture to create an authentic brand that stands out from the competition. Useful LInks Historic Agency Culture Built my brand Patagonia Patagonia Daycare Nasa Pumpkin Carving Contest The Netflix Keeper Test (It is a more positive than you think) Tidal High Fidelity Music Sharing service

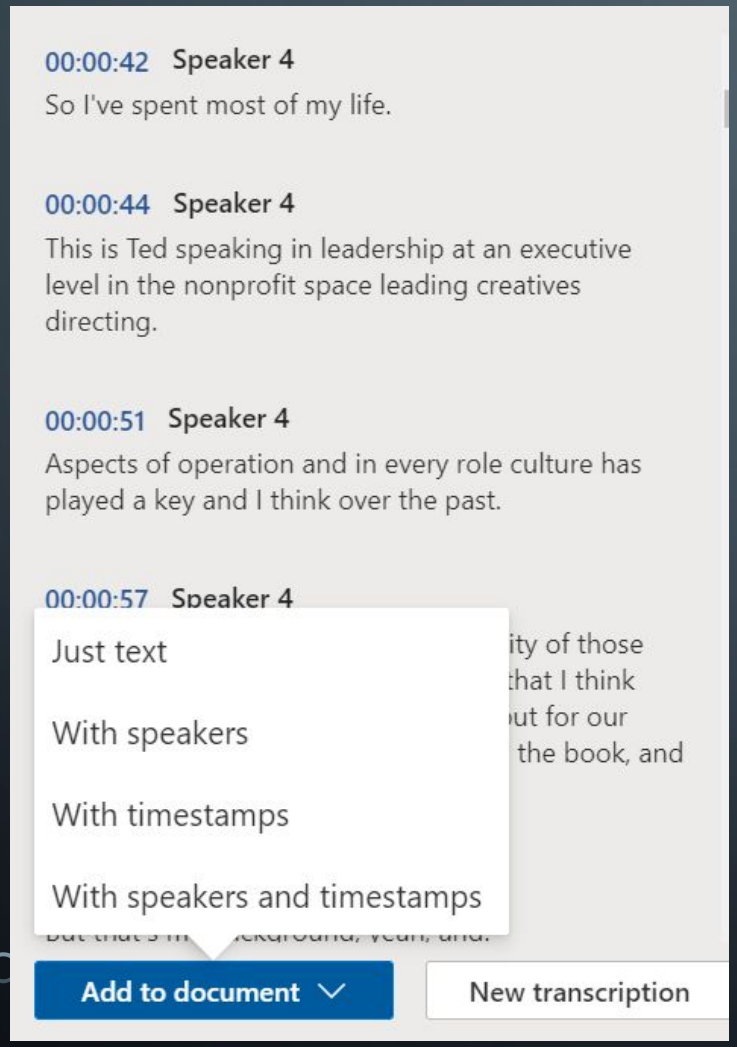
Data collection: How Did We Get The Text Data From Podcast Audio File?

- Podcast Audio mp3 File:

- Transcribe Audio file to text:
  Using Microsoft Words for Web



Data Collection: Transcribed Audio To Text




```
Transcript
00:00:00 Speaker 2
Welcome to the Digital marketing podcast. Brought to you by targetinternet.com.
00:00:10 Speaker 2
Hello and welcome back to the Digital Marketing podcast.
00:00:17 Speaker 2
My name is Kieran Rogers and today listeners we are talking about culture and.
00:00:22 Speaker 2
How it can actually eat your?
00:00:24 Speaker 2
And and I have to help me with this subject.
00:00:26 Speaker 2
A couple of real culture brand experts we have Mark Miller and Ted Vaughn.
00:00:30 Speaker 2
Welcome to the podcast guys.
00:00:32 Speaker 4
Thanks for having us.
00:00:32 Speaker 3
Thanks for having me.
00:00:33 Speaker 4
Yes, very excited to be.
```

As we can see every odd row is conversation content and every even row is time stamp and speakers information.

Data Transformation:

- Step 1: split dataset - **Text** and **Time-Stamp** datasets

```
odd_rows.head(5)
```

	Transcript
1	Welcome to the Digital marketing podcast. Brou...
3	Hello and welcome back to the Digital Marketin...
5	My name is Kieran Rogers and today listeners w...
7	How it can actually eat your?
9	And and I have to help me with this subject.

```
even_rows.head(5)
```

	Transcript
0	00:00:00 Speaker 2
2	00:00:10 Speaker 2
4	00:00:17 Speaker 2
6	00:00:22 Speaker 2
8	00:00:24 Speaker 2

Data Transformation:

- Step 2: Split column in even_row dataset -

  Timestamp and **Speaker** separately

```
odd_rows.head(5)
```

	Transcript
1	Welcome to the Digital marketing podcast. Brou...
3	Hello and welcome back to the Digital Marketin...
5	My name is Kieran Rogers and today listeners w...
7	How it can actually eat your?
9	And and I have to help me with this subject.

```
df_even.head(5)
```

	Time	Speaker
0	00:00:00	2
2	00:00:10	2
4	00:00:17	2
6	00:00:22	2
8	00:00:24	2

Part 3: Topics (Keywords) Extraction

Topics Extraction:

RAKE

- We applied Rapid Keyword Extraction (RAKE) package to extract topics from transcript message column.

- **What is Rapid Keyword Extraction (RAKE)?**
  - It is a well-known keyword extraction method which uses a list of stopwords and phrase delimiters to detect the most relevant words or phrases in a piece of text.

	Keyword	Score
0	digital marketing podcast	9.0
1	digital marketing podcast	9.0
2	kieran rogers	4.0
3	eat	1.0
4	subject	1.0

Topics Extraction:

- RAKE return the top score keywords in each sentence.

- Score: It is calculated by combining the word's frequency and the number of connections (degree).

Input sentence

Welcome to the Digital marketing podcast.

Keyword identified after remove stopword

Welcome Digital marketing podcast

Vectorise the Word

Welcome Digital marketing podcast

Degree of Word Digital marketing podcast

Welcome	1
Digital marketing podcast	9

Frequency of Word

Welcome	1
Digital marketing podcast	1

Score

Welcome	1 / 1 = 1
Digital marketing podcast	9 / 1 = 9

RAKE Score:

There are three components of the scoring:

- Word Frequency (freq(w)) - occur frequently
- Word Degree (deg(w)) - occur often and in longer candidates.
- Ratio of degree to frequency (deg(w)/freq(w)) - favors the words that predominately occur in longer candidate keywords.

The final score for each candidate keyword is calculated as the sum of its member word scores.

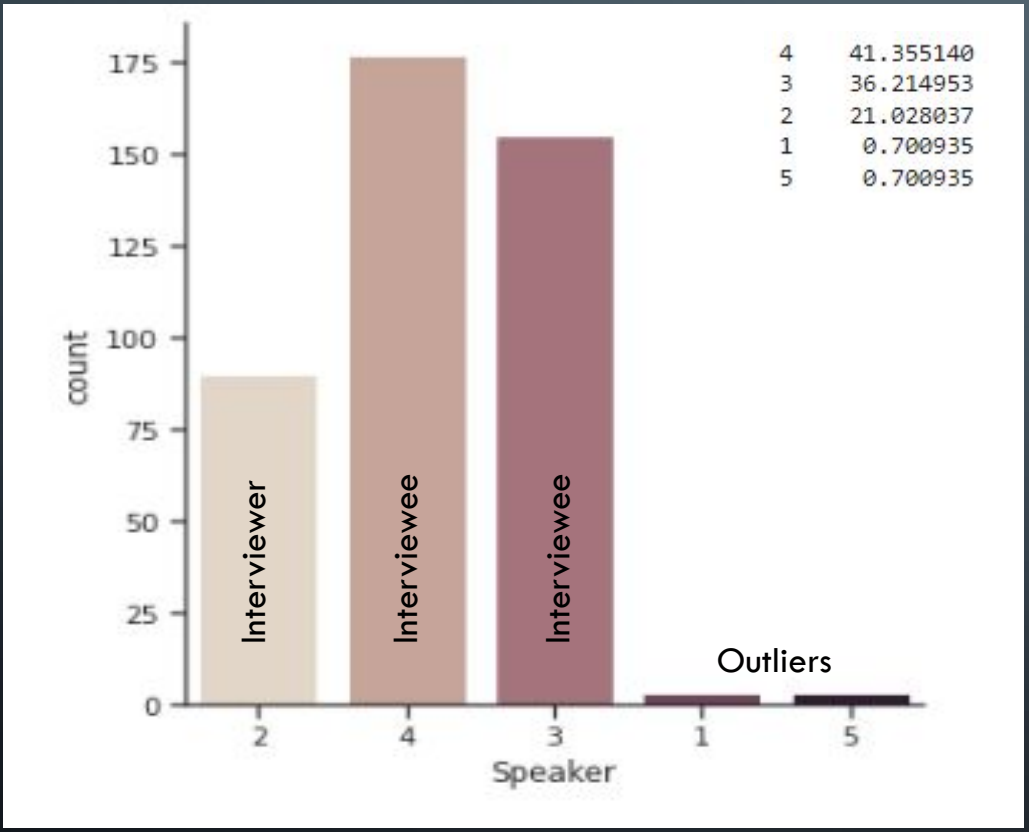
	Time	Speaker	Keyword	Score
0	00:00:00	2	digital marketing podcast	9.0
1	00:00:10	2	digital marketing podcast	9.0
2	00:00:17	2	kieran rogers	4.0
3	00:00:22	2	eat	1.0
4	00:00:24	2	subject	1.0
5	00:00:26	2	mark miller	4.0
6	00:00:30	2	podcast guys	4.0
9	00:00:33	4	excited	1.0
10	00:00:34	2	guys	1.0
11	00:00:42	4	spent	1.0

Topics Extraction:

- Merge keywords table with Timestamp and Speakers table
- Remove No speaker and No keywords observation and get final dataset

Topics Extraction: Descriptive Analysis

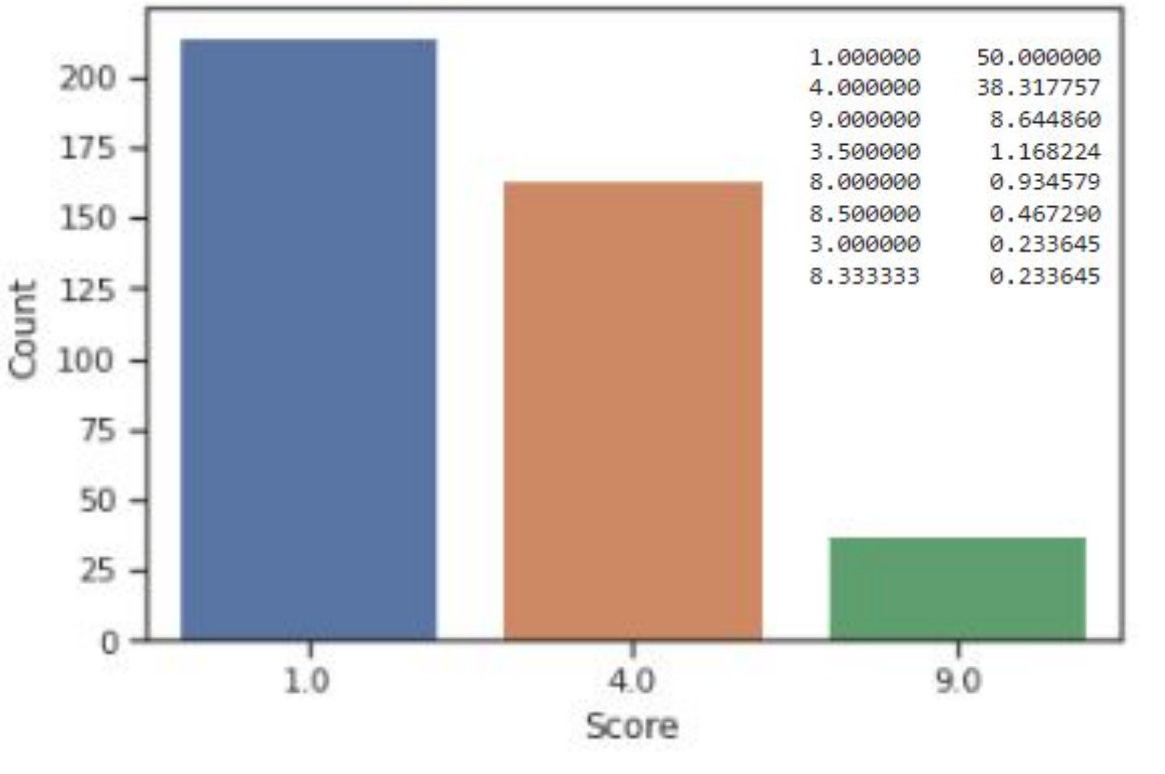
Frequency of Speakers



- The most frequent Speaker is Speaker 4, which occupies 41%, followed by Speaker 3 (26%) and Speaker 2 (21%).
- The least frequent speaker is Speaker 1 and Speaker 5 which speaker 1 is mis-transcribed (when voice overlapped) and Speaker 5 is the Advertising pitch at the end

Topics Extraction: Descriptive Analysis

Frequency of Score



- The highest score (number 9) has 8.6% of the total; this group represents the words most frequent and more connected of the podcast.
- After score 9, there are scores 4 with 38% and 1 with 50%.
- Other scores have a representation inferior to 1%.

```
df_topics_unique
```

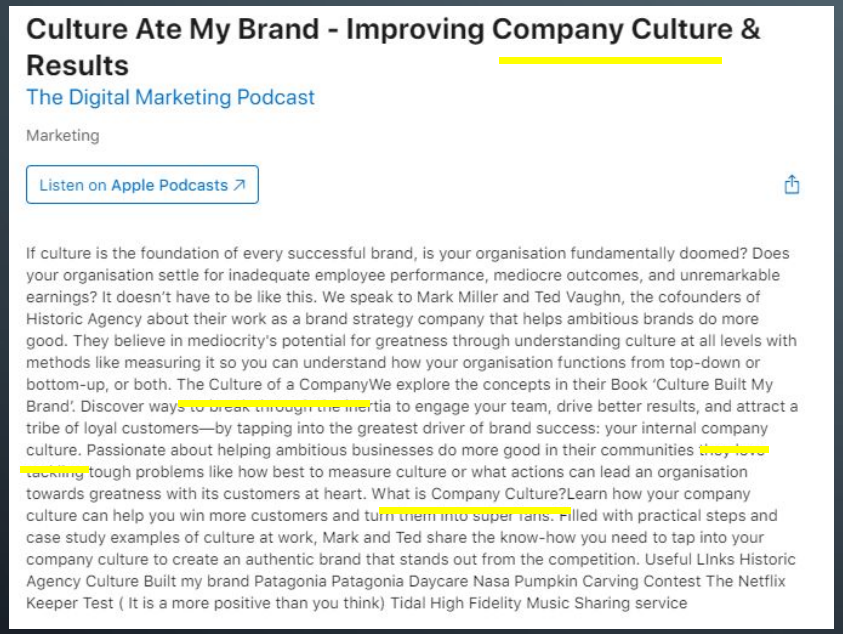
```
['digital marketing podcast',
 'social sector nonprofits',
 'pretty cynical guys',
 'bad stuff happening',
 'amazing opportunities tank',
 'stems directly back',
 'c-suite leaders found',
 'define company culture',
 'random inconsistent definitions',
 'create great products',
 'toxic culture leaked',
 'company culture defines',
 '#2 rituals lower',
 'essentially organizational structure',
 'decision making power',
 'make life easy',
 'decision making authority',
 'repeated experiential activities',
 'jet propulsion laboratory',
 'make actual rockets',
 'in-house creative teams',
 'cruise ship director',
 'false loyalty culture',
 'senior leadership deleted',
 'pumpkin carving contest',
 "bad stories 'cause",
 'great talent ended',
 'mountain climbing hardware',
 'making reusable hardware',
 'deluded senior leaders',
 'called sun shining',
 'project management software',
 'taco bell fan',
 'feel greater ownership',
 'call staff camp',
 'create camp badges']
```

Topics Extraction: Topics Summary (Score = 9 keywords)

- After getting the final clean keyword dataset, we filter out only high frequency keywords which score is 9 as screenshot shows

Topics Extraction: Compare Keywords Extraction with Ground truth

Human knowledge summary from Ground truth:



Culture Ate My Brand - Improving Company Culture & Results

The Digital Marketing Podcast

Marketing

Listen on Apple Podcasts ↗

If culture is the foundation of every successful brand, is your organisation fundamentally doomed? Does your organisation settle for inadequate employee performance, mediocre outcomes, and unremarkable earnings? It doesn't have to be like this. We speak to Mark Miller and Ted Vaughn, the cofounders of Historic Agency about their work as a brand strategy company that helps ambitious brands do more good. They believe in mediocrity's potential for greatness through understanding culture at all levels with methods like measuring it so you can understand how your organisation functions from top-down or bottom-up, or both. The Culture of a CompanyWe explore the concepts in their Book 'Culture Built My Brand'. Discover ways to break through the inertia to engage your team, drive better results, and attract a tribe of loyal customers—by tapping into the greatest driver of brand success: your internal company culture. Passionate about helping ambitious businesses do more good in their communities they love tackling tough problems like how best to measure culture or what actions can lead an organisation towards greatness with its customers at heart. What is Company Culture?Learn how your company culture can help you win more customers and turn them into super fans. Filled with practical steps and case study examples of culture at work, Mark and Ted share the know-how you need to tap into your company culture to create an authentic brand that stands out from the competition. Useful LInks Historic Agency Culture Built my brand Patagonia Patagonia Daycare Nasa Pumpkin Carving Contest The Netflix Keeper Test (It is a more positive than you think) Tidal High Fidelity Music Sharing service

This podcast help companies to understand their own **company culture** in order to build their **brand**. This brand strategy will engage **teams** and attract loyal customers.

Our Keywords Extraction result as below:

'digital marketing podcast', 'social sector nonprofits', 'pretty cynical guys', 'bad stuff happening', 'amazing opportunities tank', 'stems directly back', 'c-suite leaders found', 'define company culture', 'random inconsistent definitions', 'create great products', 'toxic culture leaked', 'company culture defines', '#2 rituals lower', 'essentially organizational structure', 'decision making power', 'make life easy', 'decision making authority', 'repeated experiential activities', 'jet propulsion laboratory', 'make actual rockets', 'in-house creative teams', 'cruise ship director', 'false loyalty culture', 'senior leadership deleted', 'pumpkin carving contest', "bad stories 'cause", 'great talent ended', 'mountain climbing hardware', 'making reusable hardware', 'deluded senior leaders', 'called sun shining', 'project management software', 'taco bell fan', 'feel greater ownership', 'call staff camp', 'create camp badges'

Topics Extraction: Compare Keywords Extraction with Ground truth

- As we can see, our keywords (subtopics) did cover the majority of the summary of ground truth, such as "company culture, teams, fan"
- However, there are too many keywords extracted , which is more than what we need.
- Next we need to track subtopics from topics (keywords) to make our topic modeling more accurate.

Part 4: Subtopics Extraction

Subtopics Extraction:

- After extracting Topics, next we are going to look at sub topics from the result of Topics (Keywords which score is 9.0) dataset.

- We perform topic modeling using **Gensim** package and LDA in this part.

  What does Gensim do? Automatically extract clear, segregated and meaningful topics from large volumes of text

Subtopics Extraction: Method1

Step 1: Store all messages into a list in order to perform vectorisation.

Step 2: Remove Stop words

```
'digital marketing podcast digital marketing podcast social sector nonprofits pretty cynical guys bad stuff happening amazing opp
ortunities tank stems directly back csuite leaders found define company culture random inconsistent definitions create great prod
ucts toxic culture leaked company culture defines 2 rituals lower essentially organizational structure decision making power make
life easy decision making authority repeated experiential activities jet propulsion laboratory make actual rockets inhouse creati
ve teams cruise ship director false loyalty culture senior leadership deleted pumpkin carving contest bad stories cause great tal
ent ended mountain climbing hardware making reusable hardware deluded senior leaders called sun shining project management softwa
re taco bell fan feel greater ownership call staff camp create camp badges'
```

Subtopics Extraction: Method1

Step 3 : Create Word Cloud

Subtopics Extraction: Method1

Step 4: Get most_common words by using FreqDist

We used FreqDist to find the number of occurrences of each word in the text. By getting len(vocab) we get the number of unique words in the text (including punctuation). And then got the most common words too.

Subtopics Extraction: Method1

Step 4: Get most_common words by using FreqDist

We filtered only the ones with at least 3 characters, then sorted them descending by number of occurrences.

```
93
[('culture', 4), ('making', 3), ('digital', 2), ('marketing', 2), ('podcast', 2), ('leaders', 2), ('company', 2), ('create', 2),
```

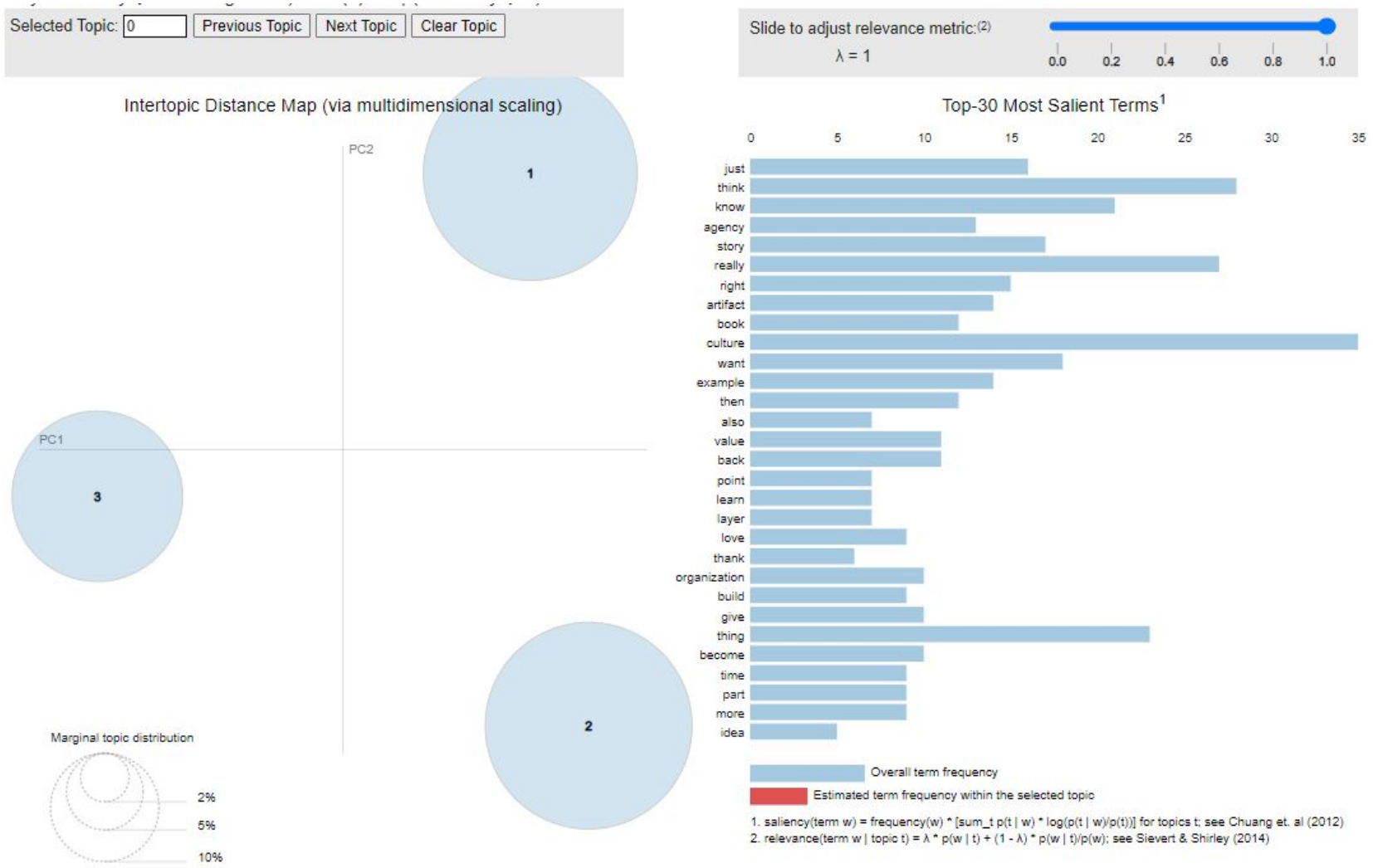
4 main sub-topics:

```
text.collocations()
```

```
digital marketing; marketing podcast; decision making; company culture
```

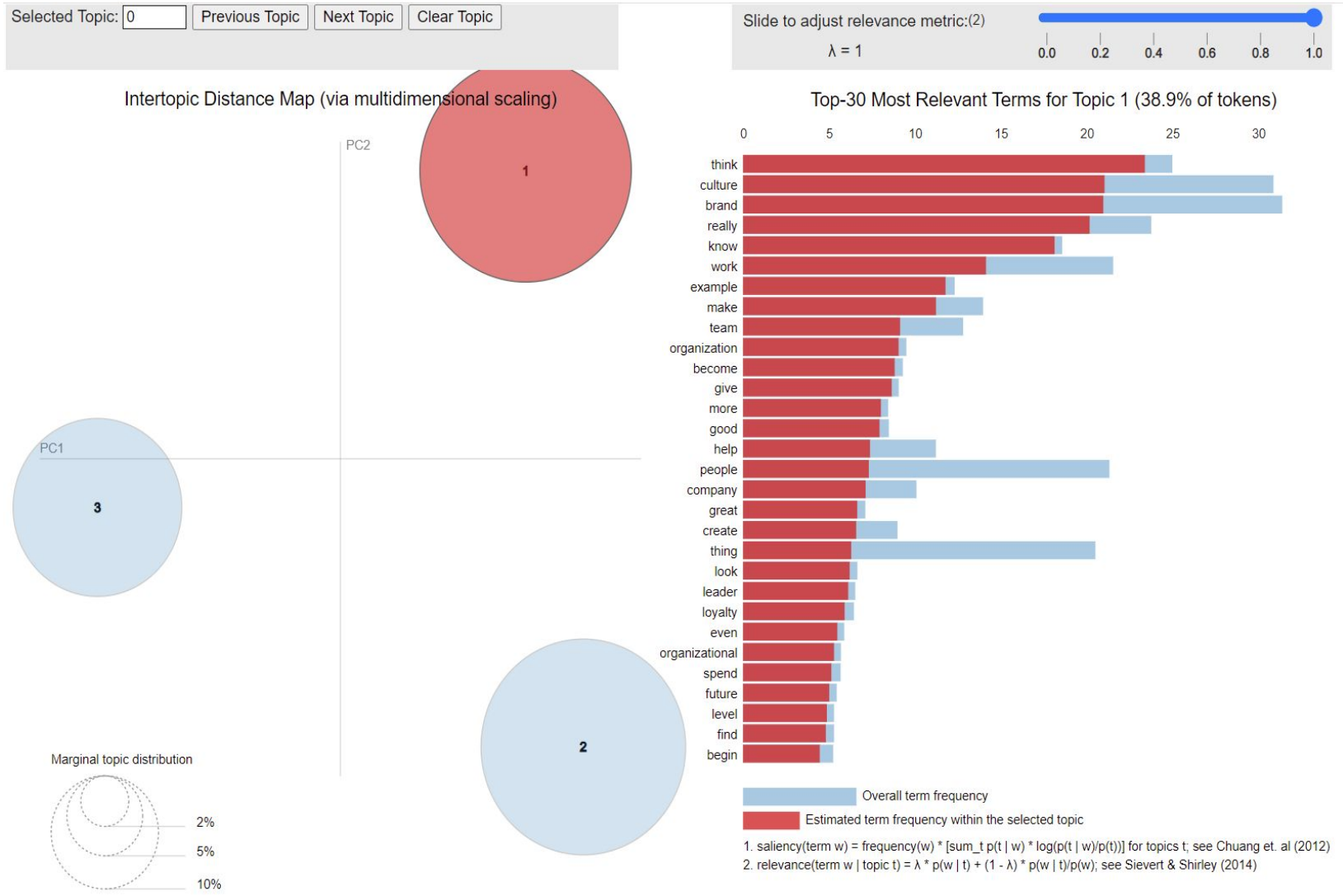
Subtopics Extraction: Method2

Latent Dirichlet Allocation (soft clustering for sub-topics)



Subtopics Extraction: Method2

Mouse over LDA1, the value change on the right



Subtopics Extraction: Method 2

Latent Dirichlet Allocation (soft clustering for sub-topics)

```
lda_model.show_topics()
```

```
[(0,
  '0.028*"just" + 0.022*"agency" + 0.018*"want" + 0.014*"people" + 0.013*"also" + 0.013*"love" + 0.012*"point" + 0.012*"learn" + 0.011*"layer" + 0.011*"thank"'),
 (1,
  '0.030*"think" + 0.027*"culture" + 0.027*"brand" + 0.026*"really" + 0.023*"know" + 0.018*"work" + 0.015*"example" + 0.014*"make" + 0.012*"team" + 0.012*"organization"'),
 (2,
  '0.020*"story" + 0.019*"thing" + 0.018*"right" + 0.017*"artifact" + 0.015*"book" + 0.014*"then" + 0.013*"value" + 0.013*"culture" + 0.013*"back" + 0.012*"actually"')]
```

Conclusion/Summary:

Topic modelling on an unsupervised data is hard to make an conclusion whether the findings are correct of not. Validation process by human is required and will have bias on the results interpretation.

Transcript tooling is not smart enough to extract speaker name from context.

```
00:00:17 Speaker 2
My name is Kieran Rogers and today listeners we are
talking about culture and.
00:00:22 Speaker 2
How it can actually eat your?
00:00:24 Speaker 2
And and I have to help me with this subject.
00:00:26 Speaker 2
A couple of real culture brand experts we have Mark Miller
and Ted Vaughn.
00:00:30 Speaker 2
Welcome to the podcast guys.
00:00:32 Speaker 4
```

Reference

- https://www.analyticsvidhya.com/blog/2021/10/rapid-keyword-extraction-rake-algorithm-in-natural-language-processing/#free-courses
- https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/
- https://towardsdatascience.com/extracting-keyphrases-from-text-rake-and-gensim-in-python-eefd0fad582f
- https://podcasts.apple.com/lb/podcast/culture-ate-my-brand-improving-company-culture-results/id373596600?i=1000542336192

Thank You