

# Data Analytics in Business

## Text Analytics

**Sridhar Narasimhan**

*Professor*

Scheller College of Business

Introduction to Text Analytics

GTx

## Lessons

- A. Introduction to Text Analytics
- B. Tokenization
- C. Word Count Analysis
- D. Sentiment Analysis
- E. Topic Modelling Using Latent Dirichlet Allocation (LDA)

GTx

## From Structured to Unstructured Data

- So far we have looked at analyses of **structured data**, which is numbers and strings that can be stored as columns in relational databases or dataframes.
- However, **unstructured data** – text heavy data, such as email, chats, social media etc. – will make up 70 to 80% of data by 2020.
- The question now is: how do we analyze unstructured data from....
  - Social media – Facebook, Snapchat, Twitter, etc.
  - Message and chat forums – Messenger, Whatsapp, etc.
  - Digital media – newspapers, magazines, blogs, etc.
  - Documents – emails, reports, etc.

Unstructured data. (2018, April 23). Retrieved May 4, 2018, from [https://en.wikipedia.org/wiki/Unstructured\\_data](https://en.wikipedia.org/wiki/Unstructured_data)

GT<sub>x</sub>

## What is Text Analytics?

- Text analytics is about converting textual data into a structured format suitable for analysis.
- Some Applications of text analytics include:
  - Predicting stock market returns (Chen, De, Hu, and Hwang (2014))
  - Predicting customer churn (Coussement and Poe (2008))
  - Competitor analysis (Thorleuchter, Dirk, and Dirk Van Den Poel. (2012))
  - Analysis of finance and accounting documents (Loughran and McDonald (2016))
  - Online reviews (Yin, Mitra, and Zhang (2016), Chevalier and Mayzlin (2006))

GT<sub>x</sub>

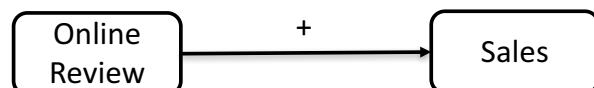
## Opportunities and Challenges

- Social media, social networks, messaging services, etc. all contain textual data that can help us to understand people's behavior.
- Artificial Intelligence based assistants (Siri, Alexa) process voice commands to recommend news, products etc.
- Text analytics and Natural Language Processing (NLP) techniques are growing.
- Challenges
  - Numerous languages, e.g. Chinese, English, French, etc.
  - Variations in usage, grammar, dialects, etc.
  - Hard to understand context, resolve ambiguity, formally encode rules of language, etc.

GT<sub>x</sub>

## Application of Text Analytics: Online Reviews

- Chevalier and Mayzlin (2006) show the impact of reviews on sales
- Online reviews contains electronic Word of Mouth (WOM)
- How to extract meaning – volume, valence, etc. – from textual data?
  - Summary, visualization, sentiment analysis, topic modeling, etc.



Volume – quantity of reviews  
 Valance – mean rating of reviews  
 Variance – distribution of ratings  
 Emotions – including positive/negative attitude, anger, anxiety etc.  
 .....

Chevalier, J., & Mayzlin, D. (2003). The Effect of Word of Mouth on Sales: Online Book Reviews. *Journal of Marketing Research*. doi:10.3386/w10148

GT<sub>x</sub>

# Application of Text Analytics: Online Reviews



GTx

## Quiz (True/False)

- In online WOM, *volume* denotes the number (or quantity) of reviews received by a product.

Answer: **TRUE**

- Unstructured data** – text heavy data, such as email, chats, social media etc. – will make up 70 to 80% of data by 2020.

Answer: **TRUE.**

GTx

# Data Analytics in Business

## Text Analytics

**Sridhar Narasimhan**

*Professor*

Scheller College of Business

Tokenization

GTx

## Tokenization

- Hard to analyze large documents. Therefore, use the *bag of words* approach:
  - Each document (bag) is a collection of tokens (words).
  - The order of words is ignored.
  - Long strings are split into smaller pieces or “tokens.”
- “A token is a meaningful unit of text, most often a word, that we are interested in using for further analysis, and tokenization is the process of splitting text into tokens.” – Silge and Robinson (2018)

Silge, J., & Robinson, D. (2018, April 02). Text Mining with R. Retrieved May 4, 2018, from <https://www.tidytextmining.com/tidytext.html>

GTx

# Tokenization

- Only keep tokens that are useful for analysis.
  - Remove punctuation.
  - Convert to lower case.
  - Remove **stop words** – commonly occurring words such as *a, an, and, after, by, why, your, we*, etc. that are not informative about the document.
  - **Stemming** – families of related words with similar meanings can be considered as a single unit by reducing words to their “stem,” base, or root form.

Silge, J., & Robinson, D. (2018, April 02). Text Mining with R. Retrieved May 4, 2018, from <https://www.tidytextmining.com/tidytext.html>

GT<sub>x</sub>

# Stemming

- For example, take the case of taste, tasted, tasteful, tastefully, tastes, tasting...
- Use the R SnowballC library for stemming. This is one library for stemming. There are others!

```
> library(SnowballC)
> wordStem(c('taste', 'tasted', 'tasteful', 'tastefully', 'tastes', 'tasting'), language = "english")
[1] "tast" "tast" "tast" "tast" "tast" "tast"
```

GT<sub>x</sub>

## R Package: Tidytext

- Use Tidytext package developed by Julia Silge and David Robinson that can be found at <https://www.tidytextmining.com/>
- Use “unnest\_tokens” function to achieve “one-token-per-document-per-row” or tidy text format.
- Pre-processing
  - Removes punctuation
  - Converts text to lower case
  - Retains other columns in dataframe for each token
  - Removes stop words
  - Applies stemming

GT<sub>x</sub>

## Example: Fine Foods

- Amazon reviews for fine foods
- URL: <http://snap.stanford.edu/data/web-FineFoods.html>
- About 500K reviews
- Data format is

```
productId: B001E4KFG0
userId: A3SGXH7AUHU8GW
profileName: delmartian
helpfulness: 1/1
score: 5.0
time: 1303862400
summary: Good Quality Dog Food
text: I have bought several of the Vitality canned dog food products and have found them all to be of
good quality. The product looks more like a stew than a processed meat and it smells better. My
Labrador is finicky and she appreciates this product better than most.
```

Mcauley, J. J., & Leskovec, J. (2013). From amateurs to connoisseurs. *Proceedings of the 22nd International Conference on World Wide Web - WWW 13*.  
doi:10.1145/2488388.2488466

GT<sub>x</sub>

```

tidy_amzn <- review %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>% # Removes stop words
  filter(word != "br") mutate(word = wordStem(word)) #HTML tag <br /><br /> results in the word "br"
  mutate(word = wordStem(word)) # stemming

> head(tidy_amzn,100)
  rev_id productid  userid rating   time   word
1    1 B001E4KFG0 A3SGXH7AUHU8GW 5 1303862400 bought
2    1 B001E4KFG0 A3SGXH7AUHU8GW 5 1303862400 vital
3    1 B001E4KFG0 A3SGXH7AUHU8GW 5 1303862400 can
4    1 B001E4KFG0 A3SGXH7AUHU8GW 5 1303862400 dog
5    1 B001E4KFG0 A3SGXH7AUHU8GW 5 1303862400 food
6    1 B001E4KFG0 A3SGXH7AUHU8GW 5 1303862400 product
7    1 B001E4KFG0 A3SGXH7AUHU8GW 5 1303862400 found
8    1 B001E4KFG0 A3SGXH7AUHU8GW 5 1303862400 qualiti
9    1 B001E4KFG0 A3SGXH7AUHU8GW 5 1303862400 product
10   1 B001E4KFG0 A3SGXH7AUHU8GW 5 1303862400 stew
11   1 B001E4KFG0 A3SGXH7AUHU8GW 5 1303862400 process
12   1 B001E4KFG0 A3SGXH7AUHU8GW 5 1303862400 meat
13   1 B001E4KFG0 A3SGXH7AUHU8GW 5 1303862400 smell
14   1 B001E4KFG0 A3SGXH7AUHU8GW 5 1303862400 labrador
15   1 B001E4KFG0 A3SGXH7AUHU8GW 5 1303862400 finicki
16   1 B001E4KFG0 A3SGXH7AUHU8GW 5 1303862400 appreci
17   1 B001E4KFG0 A3SGXH7AUHU8GW 5 1303862400 product

```

GT<sub>x</sub>

## Quiz (True/False)

- Tokenization is the process of reducing words to their root or base form.  
Answer: **FALSE.**
- Stemming is the process of removing commonly used words such as *a*, *the*, *there*, etc.  
Answer: **FALSE.**
- Stemming is the process of reducing words to their “stem,” base, or root form.  
Answer: **TRUE.**

GT<sub>x</sub>



# Data Analytics in Business

## Text Analytics

**Sridhar Narasimhan**

*Professor*

Scheller College of Business

Word Count Analysis

GTx

## How to Describe Text?

- Summary statistics (mean, variance, etc.) and visualization like scatter plot are ways of describing structured data.
- How do we describe text?
  - **Word count analysis** – table with tokens in descending order of frequency.
  - **Word count chart** – bar chart showing frequency of top N tokens.
  - **Word cloud** – visual representation of frequency (or importance) of words in a corpus.
- How are they useful?
  - Help us to understand what the keywords are.
  - Allow us to visually understand overall sense or “theme” for results of customer survey, reviews, etc.

GTx

## Word Count

- Count frequency of tokens and sort by frequency of occurrence
- Tabulate results

```
tidy_amzn %>%
  count(word, sort = TRUE) %>%
  slice(1:10)
```

A tibble: 10 x 2

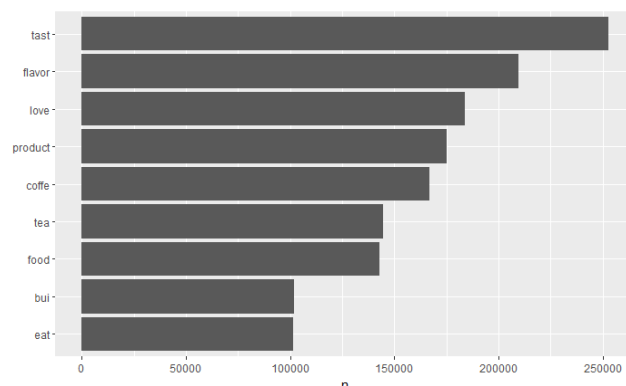
	word	n
	<chr>	<int>
1	tast	252881
2	flavor	209758
3	love	183847
4	product	175255
5	coffe	166978
6	tea	144552
7	food	143102
8	bui	101903
9	eat	101346
10	dog	99550

GTx

## Word Count Visualization

- Visualize frequency of top words as bar chart
- Customers are talking mostly about taste and flavor of food
- Common food type are coffee, tea, etc.

```
tidy_amzn %>%
  count(word, sort = TRUE) %>%
  filter(n > 40000) %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(word, n)) +
  geom_col() +
  xlab(NULL) +
  coord_flip()
```



GTx



# Data Analytics in Business

## Text Analytics

**Sridhar Narasimhan**

*Professor*

Scheller College of Business

Sentiment Analysis

GTx

## Text: Emotions and Sentiments



American Airlines got me F'd up. First they cancel my flight after delaying it 3 times. Then they rebook me so that I have to connect thru Chicago to get to Washington and offer no voucher no nothing 🙄

#AmericanAirlines

10:55 AM - 27 Apr 2018

vs.



We all love to publicly vent against airlines but today I would like to thank Southwest Airlines and the doctor who happened to be onboard. They emergency landed the plane for my dad. Because of their speed and professionalism, my father's life was saved.

#SouthWest #gratitude

5:40 PM - 26 Apr 2018

GTx

## What is Sentiment Analysis?

- “Sentiment analysis or opinion mining is the computational study of opinions, sentiments, and emotions expressed in text.” (Indurkha and Damerau (2010)).
- For a given text, unknown variable is the true sentiment (positive or negative).
- Our goal is to analyze the text to quantitatively predict its sentiment.
- Quantify sentiment in each token by using pre-built dictionaries.

Indurkha, N., & Damerau, F. J. (2010). *Handbook of natural language processing*. Boca Raton: CRC Press/Taylor & Francis Group.

GTx

## Why is Sentiment Analysis Important?

- Analyze textual data – product reviews, tweets, blogs, social media, digital documents such as filings/reports, etc. are used to **predict** whether the authors feel positive, negative, or neutral about the subject of the text. (Gentzkow, Kelly, and Taddy (2017))
- Finance – financial news is used to predict asset price movements.
- Media economics – news and social media are used to study the drivers and effects of political slant.
- Industrial organization and marketing – advertisements and product reviews are used to study the drivers of consumer decision making.
- Political economy – politicians’ speeches are used to study the dynamics of political agendas and debates.

Gentzkow, M., Kelly, B., & Taddy, M. (2017). Text as Data. *National Bureau of Economic Research*. doi:10.3386/w23276

GTx

## Sentiment Dictionaries

- Quantify sentiment of each “token” (word) in a document with the help of pre-built dictionaries
- Aggregate
- Tidyttext has in built dictionaries that give sentiment, emotions, etc. for each word
  - AFINN from Finn Årup Nielsen
  - Bing from Bing Liu and collaborators
  - NRC from Saif Mohammad and Peter Turney

GT<sub>x</sub>

## Sentiment Dictionaries

AFINN	Bing	NRC
<pre>&gt; get_sentiments("afinn") # A tibble: 2,476 x 2   word      score   &lt;chr&gt;    &lt;int&gt; 1 abandon    -2 2 abandoned  -2 3 abandons   -2 4 abducted   -2 5 abduction  -2 6 abductions -2 7 abhor      -3 8 abhorred   -3 9 abhorrent  -3 10 abhors    -3 # ... with 2,466 more rows</pre>	<pre>&gt; get_sentiments("bing") # A tibble: 6,788 x 2   word      sentiment   &lt;chr&gt;    &lt;chr&gt; 1 2-faced    negative 2 2-faces    negative 3 a+         positive 4 abnormal   negative 5 abolish    negative 6 abominable negative 7 abominably negative 8 abominate  negative 9 abomination negative 10 abort      negative # ... with 6,778 more rows</pre>	<pre>&gt; get_sentiments("nrc") # A tibble: 13,901 x 2   word      sentiment   &lt;chr&gt;    &lt;chr&gt; 1 abacus    trust 2 abandon   fear 3 abandon   negative 4 abandon   sadness 5 abandoned anger 6 abandoned fear 7 abandoned negative 8 abandoned sadness 9 abandonment anger 10 abandonment fear # ... with 13,891 more rows</pre>

GT<sub>x</sub>

## Sentiment Analysis

- Join sentiment table to pre-processed tokens to get sentiment for each token.

```
tidy_amzn_sentiment <- tidy_amzn %>%
  inner_join(get_sentiments("afinn"), by = "word")
```

rev_id	productId	userId	rating	time	word	score
1	2 B00813GRG4	A1D87F6ZCVE5NK	1	1346976000	error	-2
2	3 B000LQOCH0	ABXLMWJIXXAIN	4	1219017600	cut	-1
3	3 B000LQOCH0	ABXLMWJIXXAIN	4	1219017600	heaven	2
4	3 B000LQOCH0	ABXLMWJIXXAIN	4	1219017600	recommend	2
5	6 B006K2ZZ7K	ADT0SRK1MG0EU	4	1342051200	enjoy	2
6	7 B006K2ZZ7K	A1SP2KVKFXXRU1	5	1340150400	stuck	-2
7	7 B006K2ZZ7K	A1SP2KVKFXXRU1	5	1340150400	recommend	2
8	7 B006K2ZZ7K	A1SP2KVKFXXRU1	5	1340150400	love	3
9						

GTx

## Sentiment Analysis

- Aggregate “sentiment scores” at the document (or review) level to get sentiment of that document.

#Now get average sentiment score for each productId to plot rating vs. avg\_score

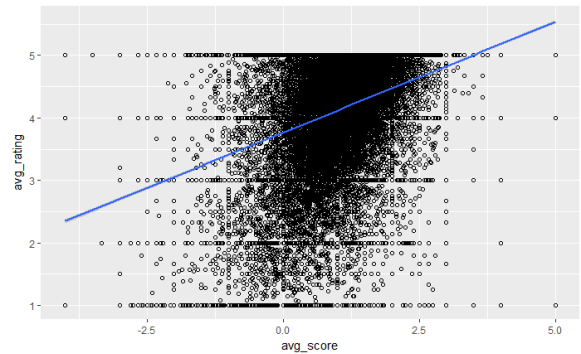
```
tidy_amzn_sentiment_prod <- tidy_amzn_sentiment %>%
  group_by(productId) %>%
  summarise(avg_score=mean(score),
            sum_score=sum(score),
            avg_rating = mean(rating))
```

GTx

## Visualize Results

```
ggplot(tidy_amzn_sentiment_prod,
  aes(x=avg_score, y=avg_rating)) +
  geom_point(shape=1) + # Use hollow circles
  geom_smooth(method=lm, # Add linear regression line
    se=TRUE) # Don't add shaded confidence region
```

- As the average sentiment score for a review increases, the average rating also increases in the positive direction.
- Reviews with rating of 5 often contain a combination of negative and positive sentiment. It is possible these reviews highlight both good and bad aspects of the product.

GT<sub>x</sub>

## Impact of Online Word Of Mouth (WOM)

- Chevalier and Mayzlin (2006) study impact of consumer reviews on relative book sales at Amazon and Barnes and Noble
  - Most reviews are positive
  - Better (positive) reviews lead to higher relative sales
  - Negative reviews have greater impact than positive reviews
- Yin, Mitra, and Zhang (2016) study Apple's app store review data to find that consumers have "confirmation bias" when evaluating helpfulness of positive vs. negative reviews. In general, positive (negative) reviews are more helpful when the majority of ratings is positive (negative).

Chevalier, J., & Mayzlin, D. (2006). The Effect of Word of Mouth on Sales: Online Book Reviews. *Journal of Marketing Research*. doi:10.3386/w10148

Yin, Dezhi, Sabyasachi Mitra, and Han Zhang. "Research note—When do consumers value positive vs. negative reviews? An empirical investigation of confirmation bias in online word of mouth." *Information Systems Research* 27, no. 1 (2016): 131-144.

GT<sub>x</sub>



## Quiz

- Which one of these is NOT a pre-built dictionary in tidytext package?
  - A. LIWC
  - B. AFFIN
  - C. NRC
  - D. Bing

Answer: **A. LIWC**

- True or False: Review valence and sales are positively related.

Answer: **TRUE**. A positive review leads to higher sales and vice versa.

GT<sub>x</sub>

## Data Analytics in Business

### Text Analytics

**Sridhar Narasimhan**

*Professor*

Scheller College of Business

Topic Modeling Using Latent  
Dirichlet Allocation (LDA)

GT<sub>x</sub>

## Use of Topic Modeling

- User Generated Content (UGC) such as social media, online reviews, etc. are an important source for understanding customers' experience with quality.
- But quality is multi-dimensional. How do we summarize reviews to get these hidden or "latent" dimensions?
- Goal of topic modeling is to discover the latent topics or factors from a large number of text documents.

Tirunillai, S., & Tellis, G. J. (2014). Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation. *Journal of Marketing Research*, 51(4), 463-479. doi:10.1509/jmr.12.0106



## History of Topic Models

- Probabilistic generative model is used primarily for text mining and information retrieval in recent years.
- Origin of topic modeling is latent semantic indexing (LSI) (Deerwester, et al. (1990)); probabilistic latent semantic analysis (PLSA) (Hofmann (2001)) was proposed by Hofmann as the first probabilistic topic model.
- "Each document in a given corpus is thus represented by a histogram containing the occurrence of words. The histogram is modeled by a distribution over a certain number of topics, each of which is a distribution over words in the vocabulary. By learning the distributions, a corresponding low-rank representation of the high dimensional histogram can be obtained for each document" (Hofmann (2001)).

Liu, L., Tang, L., Dong, W., Yao, S., & Zhou, W. (2016). An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, 5(1). doi:10.1186/s40064-016-3252-8  
Hofmann, T. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, 42(1/2), 177-196.



## Latent Dirichlet Allocation (LDA)

- LDA proposed by Blei et al. (2003) is an improvement over PLSA.
- It is an unsupervised learning method similar to cluster analysis (where we discover latent groups or clusters).
- Tirunillai and Tellis (2014) use LDA to discover these latent quality dimensions from reviews through LDA.
- Most common algorithm for topic modeling. It is based on two guiding principles:
  - Every document is a mixture of topics.
  - Every topic is a mixture of words.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3,993-1022.

GT<sub>x</sub>

## Documents, topics, mixture of words

- Every document is a mixture of topics. We imagine that each document may contain words from several topics in particular proportions. For example, in a two-topic model we could say “Document 1 is 90% topic A and 10% topic B, while Document 2 is 30% topic A and 70% topic B.”
- Every topic is a mixture of words. For example, we could imagine a two-topic model of American news, with one topic for “politics” and one for “entertainment.”
- The most common words in the politics topic might be “President”, “Congress”, and “government”, while the entertainment topic may be made up of words such as “movies”, “television”, and “actor”.
- Importantly, words can be shared between topics; a word like “budget” might appear in both equally.

GT<sub>x</sub>

# LDA Example

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be \$200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive \$400,000 each. The Juilliard School, where music and the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

Figure 8: An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3,993-1022.

GTx

## LDA Using Tidytext

- Convert Tidytext into another format "DocumentTermMatrix" (DTM) that is more suitable for LDA analysis.
- Use the method LDA to generate k topics for the given DTM.

```
amzn_dtm <- tidy_amzn %>%
  count(productId, word, sort = TRUE) %>%
  ungroup() %>%
  cast_dtm(productId, word, n)

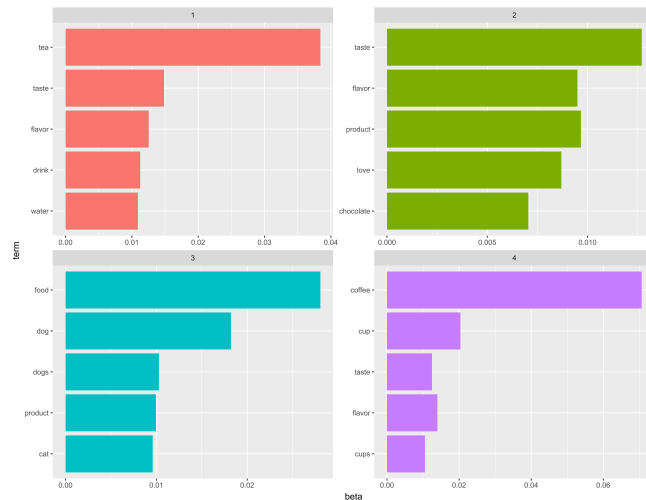
product_lda <- LDA(amzn_dtm, k = 4, control = list(seed = 1234))
```

GTx

# LDA with Four Topics

First topic is about **tea**, second about **taste/flavor**, third about **pet food**, and the last one about **coffee**

```
product_topics <- tidy(product_lda, matrix = "beta")
top_terms <- product_topics %>%
  group_by(topic) %>%
  top_n(5, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
top_terms %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip()
```



GTx

## Next Steps

- LDA should be run for at least 2 topics.
- Keep the number of topics small to improve interpretation.
- Alternatively, advanced methods are available to find optimal number of topics using LDA:

<https://cran.r-project.org/web/packages/ldatuning/vignettes/topics.html>

GTx

## Quiz (True/False)

- LDA is a supervised machine learning model.

Answer: **FALSE**. LDA is an unsupervised ML model similar to K-means clustering.

GTx

## Recap of Lessons

- A. Introduction to Text Analytics
- B. Tokenization
- C. Word Count Analysis
- D. Sentiment Analysis
- E. Topic Modelling Using Latent Dirichlet Allocation (LDA)

GTx

# Data Analytics in Business

## Text Analytics

**Sridhar Narasimhan**

*Professor*

Scheller College of Business

End



GT<sup>x</sup>