# Essentials of Sentiment Analysis

**Cluster-An on Text Data**

Session # 2

Text Analytics for Batch 12

CBA @ ISB

Sudhir Voleti

1

## Session Plan

- Sentiment Analysis: Introduction
  - Three sentiment dictionaries (*tidytext*)
  - Valence shifters (*sentimentr*)
- Some useful Shiny apps to see
  - Basic text-an and sentiment-an apps
  - In-class exercise on Primary Data
- Basic Cluster-An Primer
- Clustering Text Data

2

2

# Introduction to Sentiment-An

3

## Sentiment Mining: What and Why

- What is *sentiment mining*?
- Attempt to detect, extract and assess *value judgments*, subjective opinion and *emotional content* in text data

- Why care about sentiment?
- It may have great explanatory and/or predictive power as a feature set in analytics models.

- How is sentiment measured?
- **Valence** is the technical term for subjective inclination of a document – measured along a Positive/ Neutral/ Negative continuum.

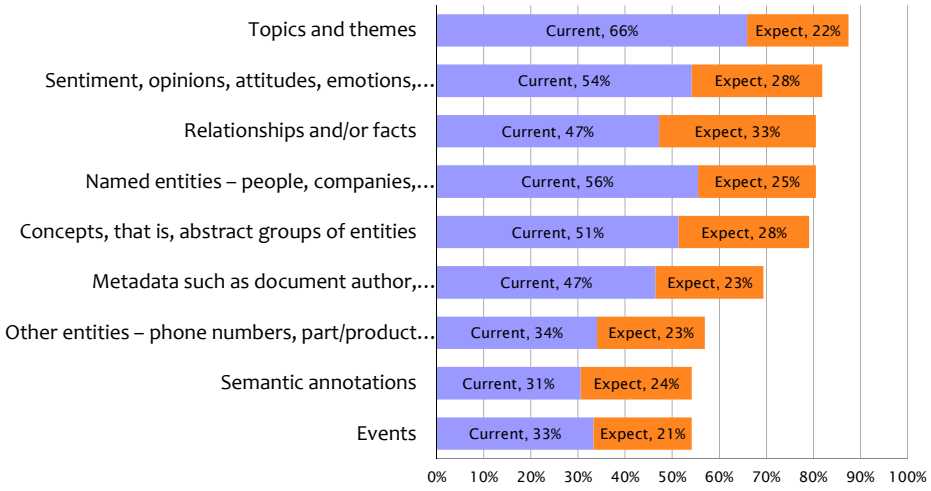
- Valence can be **measured and scored**. *How?*
- Can (and *should*) build our own context-specific sentiment scoring scheme given valence weights for the most common phrases…

4

4

# What are Firms Mining in text?

**Do you currently need (or expect to need) to extract or analyze…**




**Source: JDAT. JDAT is the Journal of Data Analysis Techniques*

5

# Sentiment Mining Steps

1. Start with the tokenized document

2. Match stem Document with **positive (or negative) word-lists**

3. **Score** each document in the corpus for positive (or negative) **'polarity'**

4. Plot the distribution of sentiment scores for the corpus.

5. Standardize the matrix; examine individual docs.

E.g., the *qdap* library follows this pattern to the dot.

6

# Sentiment Dictionaries in Tidytext

Open the markdown 'sentiment-an.Rmd'

7

## Three Tidytext Sentiment Dictionaries

- Bing
  - Simple dictionary with positive-negative classifications in a wordlist. Created by Bing Liu et al.
- AFINN
  - AFINN is a list of English words rated for valence with an integer between minus five (negative) and plus five (positive). The words have been manually labeled by Finn Arup Nielsen in 2009-2011.
- nrc
  - The NRC Emotion Lexicon is a list of English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). The annotations were manually done by crowdsourcing.
- Loughran (finance sentiment dict, in development)

8

8

## Sentiment-An Recap: Some quick Qs

- What is Sentiment-An?
- What main libraries did we explore in Sentiment-An?
- What main functions did we use in Sentiment-An?
- What are the advantages and limitations of wordlist based sentiment-An?

9

9

## Valence Shifters in *sentimentR*

10

## Sentiment-An with Valence Shifters: Preliminaries

- Here're two review excerpts on a popular TV series. Examine their 'sentiment content'.

- "This is one of the best TV series I have seen in a long time. I am yet to read the books but if the TV series is anything to go by then the books will be outstanding."
- "After seeing today in Game of Trones, I realized that author of this serial, as well as HBO may need professional (medical) assistance. Actually, since the beginning, it is hard to find anything that would make anyone feeling good. ..."

- Precisely *what tokens* in the reviews made you decide on positive/negative?
- Note how *valence shifters* like 'is hard to' modify the valence of 'good'.

11

11

## Valence Shifters and SentimentR

- Valence shifters must be assessed minimum at the _____ level of analysis. (words, sentence, para, doc or something else?)
- '
- ?

| Text | Negator | Amplifier | Deamplifier | Adversative |
|---|---|---|---|---|
| Cannon reviews | 21% | 23% | 8% | 12% |
| 2012 presidential debate | 23% | 18% | 1% | 11% |
| Trump speeches | 12% | 14% | 3% | 10% |
| Trump tweets | 19% | 18% | 4% | 4% |
| Dylan songs | 4% | 10% | 0% | 4% |
| Austen books | 21% | 18% | 6% | 11% |
| Hamlet | 26% | 17% | 2% | 16% |

- Open the RMD 'valence shifters' and follow the code flow.

12

12

Some useful text-an shinyapps

to look at

13

## 2 Shinyapps in text-an

- Shiny is a set of workflow tools to render interactive user experience in R.
- I built some shiny apps for text-an mainly for PGP classroom use, two of which I show here.
  - Basic txt-an app
  - Basic sentiment-an app
- Recall our use of 'source()' to make available a whole host of under defined funcs we'd written?
- With shiny, the aim is to appreciate another useful way of making your cool work available to nontechnical people in your org.

14

14

## Running the shinyapps

- To activate shiny, copy-paste the following code into your R console

- # Basic Text Analysis shiny App
- source("https://raw.githubusercontent.com/sudhir-voleti/basic-text-analysis-shinyapp/master/dependency-basic-text-analysis-shinyapp.R")
- runGitHub("basic-text-analysis-shinyapp", "sudhir-voleti")

- # Basic Sentiment Analysis App
- source("https://raw.githubusercontent.com/sudhir-voleti/tidy-sentiment-analysis/master/dependency-tidy-sentiment-analysis.R")
- runGitHub("tidy-sentiment-analysis", "sudhir-voleti")

15

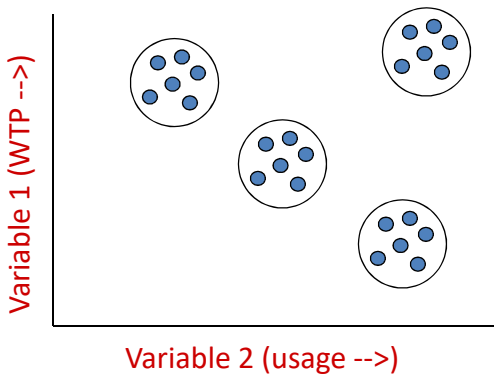
15

## Clustering Text Data

16

# An Ideal Clustering Situation

Consider the location of 24 people on a usage versus WTP map in the cooking oil category …

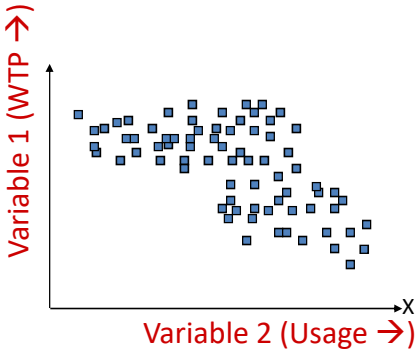



Note **closeness** within a *cluster*, **distance** between clusters.

17

# Segmentation: A Practical Clustering Situation

We saw the *ideal* case. Reality is seldom that clear-cut…

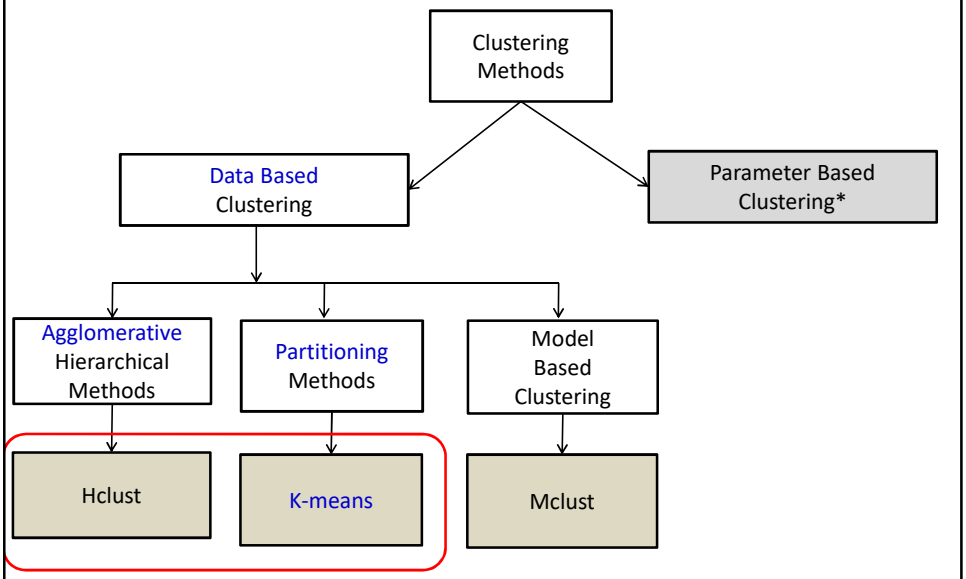



**Some Questions of interest:**

(i) How many clusters are there? Just **one**? Many? How many?
(ii) What's their size? Their other **characteristics**?
(iii) Are we **missing** some variables? How would we know?
(iv) Suppose there were **10** *basis* variables for clustering, then how to proceed?
(v) Etc.

18

## Cluster-An: Some Broad Approaches




19

19

## Cluster An with k-means on Text Data

- Recall the fundamental units of analysis we spoke of in text an – tokens, DTMs, n-grams etc.
- Which among them would be most amenable to clustering? Why?
- What might clustering text data in a corpus give us in terms of output? Interpretation? Insight?
- How about we try this on a data set and see? We'll use the Ice cream dataset.

**Open file 'cluster an applied to text'**

20

20

## Cluster An on Text: Recap and Review

- What happens when we run cluster-an on a DTM?
- What gets clustered?
- What are the basis variables?
- What insights emerged from the clustering of ice-cream preference data?
- Point to ponder: We just used a well-established matrix analysis method (kmeans clustering) on a text matrix.

21

21

## Hierarchical clustering with Stringdist

- Recall stringdist? It also yields distance metrics between text units (units).
- And we know that cluster-an uses distance metrics between units of analysis to cluster them into groups of 'similar' units.
- So why not apply cluster-an on stringdist output? Aim is to cluster strings (names, brands etc) into groups of similar strings.
- Open file '*cluster an with stringdist.Rmd*' and follow me.
- **Recap** - What did we just do? With what funcs? Useful in what applications? Etc.

22

22

Sentiment-An

In-class Exercise
(time permitting)

23

## Sentiment-An Exercise on Review Data

- Read-in the file 'Iron man reviews.txt'
- Now try these Qs:
- 1. What was the size of the DTM?
- 2. What sentiment-words (positive and negative) occurred the most in your data?
- 3. What are the most frequent 3 words corresponding to the 'anticipation' emotion that come from your data?
- 4. What if any display aids did you use to answer the above questions?

24

24

## Sentiment Analysis: Recap

- Q: What could we accomplish with elementary Sentiment-An?
- Able to rapidly, scale-ably, cheaply crunch through raw text input,
- … and locate sentiment-laden words in the corpus.
- Able to display broad contours of which sentiment arise most,
- And view the *distribution* of sentiment across the corpus
- Ability to make one's own *custom built sentiment lexica* and use with tidytext.

25

Q & A

26