

## Machine Learning

# Text Mining

greatlearning

- Nature of Corporate Data
  - 20-75-5 Rule
  - Structure of Data
- Text as Data
  - Loose Structure
  - Poor Spellings
  - Non traditional
  - Multi-lingual



hannah @lawlorr

MY ELECTRIC HAS WENT OUT AND A GIANT SPIDER IS COMING  
ME AND MY ONLY SOURCE OF LIGHT IS THE FLASHLIGHT ON MY  
PHONE GOD BLESS @Apple

iPhone 7: no headphones

iPhone 8: no home button

iPhone 9: no camera

iPhone 10: no phone #AppleEvent #iPhoneX



— Michael WizMin (@MichaelWizmin) September 12, 2017

# Why is it Hard? **greatlearning**

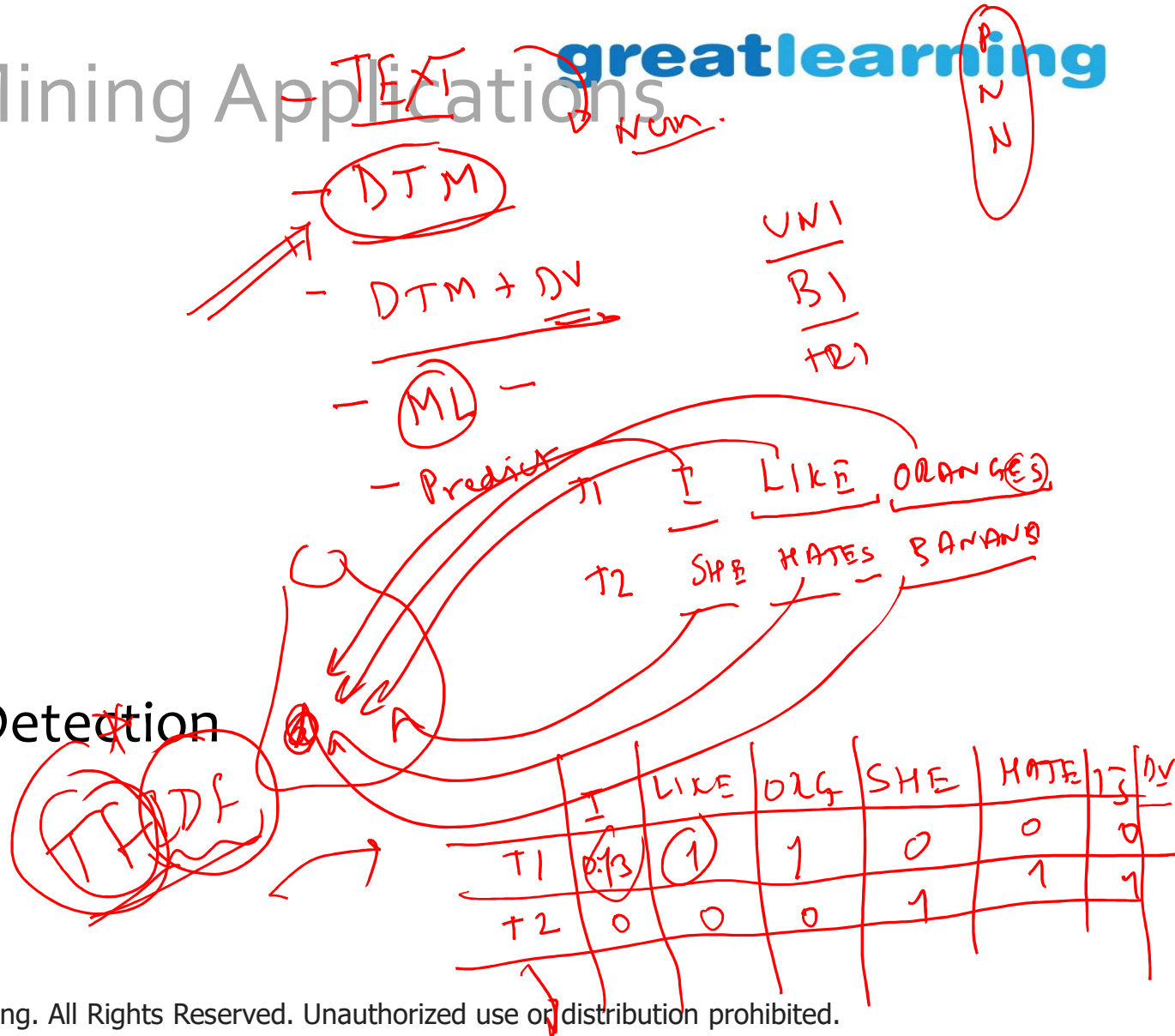
- Ambiguity:
  - “I put my **bag** in the **car**. **It** is large and blue”
  - “**It**” = **bag**? “**It**” = **car**?
- Context:
  - Homonyms, metaphors
  - Sarcasm

# Text Mining Applications

greatlearning

P  
2  
N

- Web and Social Media
  - Discovery of:
    - Sentiments
    - Opinions
    - Emotions
    - Topics
- Fraud and Irregularity Detection
- Spam identification
- Business Intelligence
- Content Enrichment



# Process of TM

**greatlearning**

- Generic Process
  - Problem or Challenges
  - Collecting data
  - Human input- Average
  - Text mining packages
  - Creating Corpus
  - Processing Text- upper, lower, stop words, stem
  - Document Term Matrix
- Using Methods of Analytics



Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.