Practical Text Analytics: Introduction

Fan Dai, Subrata Pal

Iowa State University

September 23, 2019

Introduction: Text data

- Text data
  - Product reviews, social media posts, customer feedback, etc.

Source: **https://www.wordstream.com/social-media-marketing**

* Concepts & topics
* Sentiment & Emotions
* Behaviors

Introduction: Text data

- Text data
  - Product reviews, social media posts, customer feedback, etc.
  - **Unstructured data**
    * Structured data: stored and organized in typical databases like Excel, Google Sheets, and SQL.
    * Unstructured data: everything else..

Introduction: Text data

- Text data
  - Product reviews, social media posts, customer feedback, etc.
  - **Unstructured data**
    - Structured data: stored and organized in typical databases like Excel, Google Sheets, and SQL.
    - Unstructured data: everything else..

Traveller Id	Traveller star	location	cleanliness	Service	Sleep Quality	Value	Rooms	Trip Type	Text reviews
Filippo S	5		5	5	5			business	I had the most incredible experience in this hotel. I
Explorer078	3	1	3	2	2	3	3	business	Getting in and out of hotel is extremely hazardous I
hazimismail	4	3	3	4	4	3	4	business	Wverything was fine with the hotel. Rooms were a
Barranquita	5	5	5	5	5	5	5	family	Free airport shuttle on time both at arrival and dep
Josephine B	5								HGI Miami APW was a very nice place to unwind af Cruise. They were very accommodating (able to che free water bottles). It is not too far from the airport free shuttle) and has many fast-food restaurants (M Starbucks, Dunkin Donuts) within walking distance. Also appreciate the onsite washing machines and cl a small fee) and morning breakfast (another fee). Highly recommend.
Isabelle W	5	5	5	5	5	5	5	couple	Just check-out of my favorite Miami hotel - again, th
mikeslady	4								Great breakfast buffet, nice seating area in lobby fo
Frankie_Suri	4	3	4	4	5	4	5	family	Very Nice Hotel , Clean rooms, free shuttle to Airpo

Figure: Customer reviews for Hilton Garden Inn-Miami Airport (Dec Jan, 2019)

Source: **https://www.tripadvisor.com/**

Introduction: Text data

- Text data
  - Product reviews, social media posts, customer feedback, etc.
  - **Unstructured data**
    * Structured data: stored and organized in typical databases like Excel, Google Sheets, and SQL.
    * Unstructured data: everything else..
  - **Big data**
    * Massive volumes
    * Many different file types
    * High-speed data creation

Introduction: Text analytics

- Text analytics
  - Information retrieval
  - Information extraction
  - Data mining
    * Quantitative results
    * Graphs, tables and other sorts of visual reports.

Introduction: Text analytics

Figure: Word cloud of text analytics article titles and abstracts

Source: **Practical text analytics: maximizing the value of text data**

Introduction: Text analytics

Figure: Word cloud of text analytics job titles

Source: **Practical text analytics: maximizing the value of text data**

Introduction: Text analytics

Figure: Word cloud of the skills required for text analytics jobs

Source: **Practical text analytics: maximizing the value of text data**

Introduction: Text analytics process

Text analytics process

- Planning the text analytics projects
- Text preparing and preprocessing
- Unsupervised and supervised analysis
- Result interpretations

Introduction: Text analytics process

Planning the text analytics projects

- Initial consideration
- Problem framing
- Data generation
- Method and implementation selection
- Inference and decision-making

Introduction: Text analytics process

Initial considerations

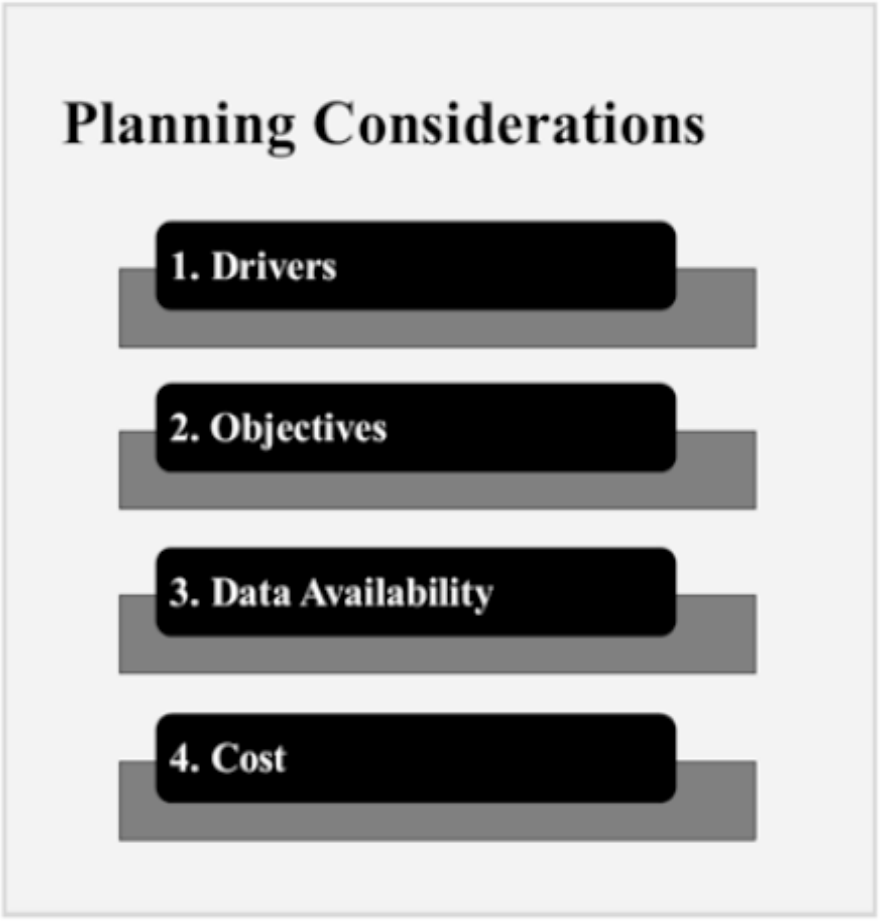

Figure: Initial planning considerations

Source: **Practical text analytics: maximizing the value of text data**

Introduction: Text analytics process

Problem framing

- Identify analysis problems
- Decide types of inference
  - **Deductive inference**
    - Research questions ⇒ Theory ⇒ Hypothese ⇒ Conceptualization
    - Theory-driven coding schemes

Introduction: Text analytics process

Problem framing

- Identify analysis problems
- Decide types of inference
  - **Deductive inference**
    - Research questions ⇒ Theory ⇒ Hypothese ⇒ Conceptualization
    - Theory-driven coding schemes
  - **Inductive inference**
    - Data-driven coding schemes
    - Create a generalizable theory

Introduction: Text analytics process

Data generation

- Define research scope and purpose
  - Unitization
    - Sampling unit: each row of data or text document
    - Recording unit: piece of text that is categorized in the coding step
    - Context unit: define the amount of text that can be used

Introduction: Text analytics process

Data generation

- Define research scope and purpose
  - Unitization
    - Sampling unit: each row of data or text document
    - Recording unit: piece of text that is categorized in the coding step
    - Context unit: define the amount of text that can be used
- Collect text data
  - Sampling techniques
    - Non-probability sampling: convenience sampling, relevance sampling
    - Probability sampling: systematic sampling, stratified sampling

Introduction: Text analytics process

- Data quality

Figure: Quality dimensions of text data

Source: **Practical text analytics: maximizing the value of text data**

Example: Business policies on human rights (Preuss and Brown, 2012)

- Inductive inference
- Research questions: frequency and content of human rights policies in large corporations
- Relevance sampling: all companies on the Financial Times Stock Exchange 100 (FTSE 100) list
- Sampling unit: each firm's website
- Context unit: corporate codes of conduct ans supporting documents
- Recording unit:
  whether a company had a human rights policy: corporate website

Introduction: Text analytics process

Different methods based on our objective(s)

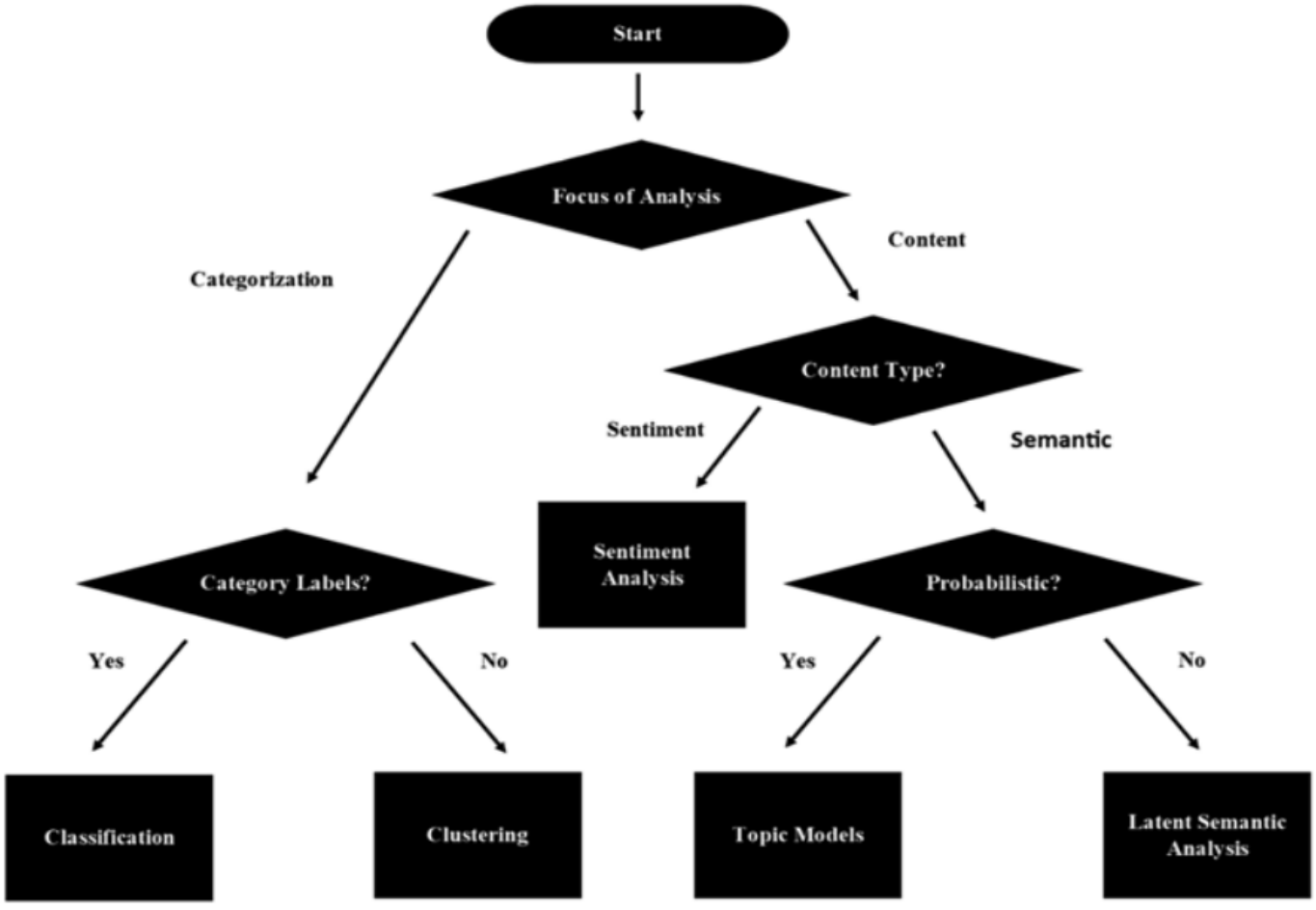

Figure: Flow chart of what-to-do based on our objective

Source: **Practical text analytics: maximizing the value of text data**

Different methods based on our objective(s):(contd.)

- If our objective is to categorize or classify the documents, go for **Classification** or **Clustering** depending on levels exists or not.
- If our objective is not to categorize or classify the documents, rather we're concerned with the content of the documents, we should go for the other type:
  - Identify the sentiment of the documents in the document collection → sentiment analysis.
    If the data have sentiment labels, classification methods can be used also.
  - To identify semantic ("linguistic and philosophical study of meaning in language..." - Wikipedia) information in the documents, there are two classical analysis methods available: **latent semantic analysis (LSA)** and **topic models**. Topic models are probabilistic, generative models, while LSA deals with dimension reduction methods.