Topic Mining and Analysis: Motivation and Task Definition

ChengXiang “Cheng” Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

Topic Mining and Analysis: Motivation and Task Definition

5. Text-based prediction

3. Topic mining and analysis

Real World

Perceive

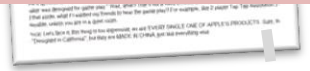
(Perspective)

Observed World

Express

(English)

Text Data



1. Natural language processing and text representation

4. Opinion mining and sentiment analysis

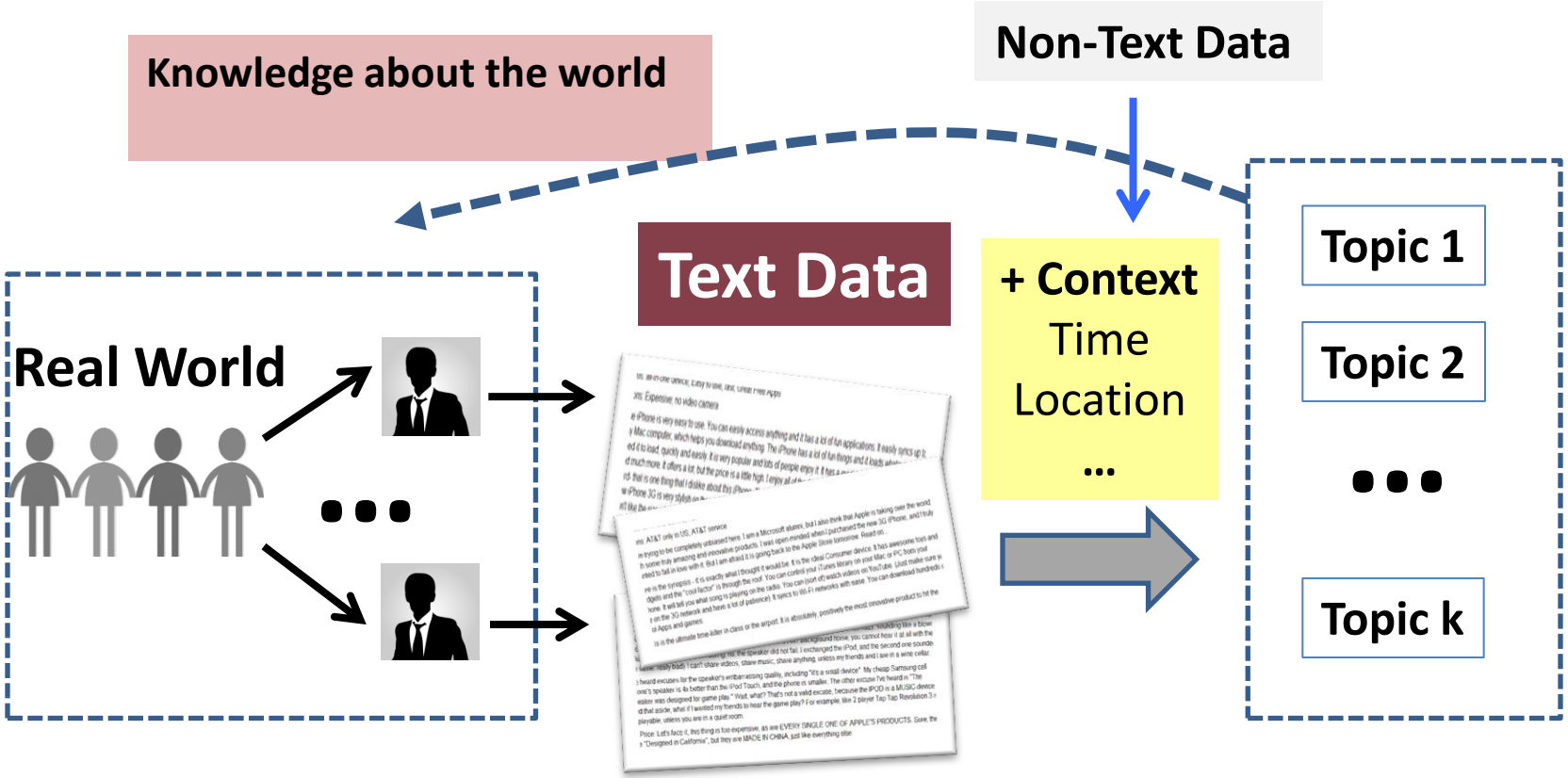
2. Word association mining and analysis

Topic Mining and Analysis: Motivation

根据 content 去辨别 content 的 topic.

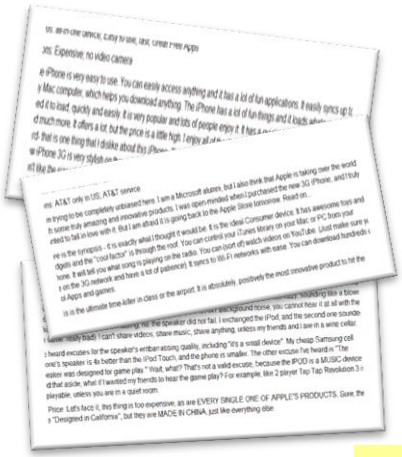
- Topic ≈ main idea discussed in text data
  - Theme/subject of a discussion or conversation
  - Different granularities (e.g., topic of a sentence, an article, etc.)
- Many applications require discovery of topics in text
  - What are Twitter users talking about today?
  - What are the current research topics in data mining? How are they different from those 5 years ago?
  - What do people like about the iPhone 6? What do they dislike?
  - What were the major topics debated in 2012 presidential election?

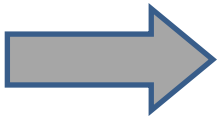
Topics As Knowledge About the World

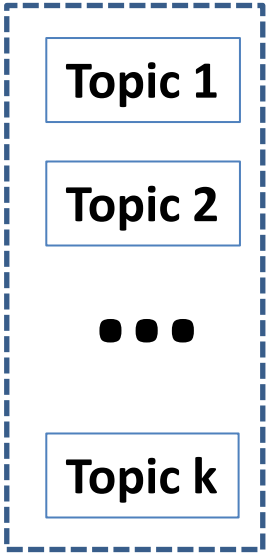


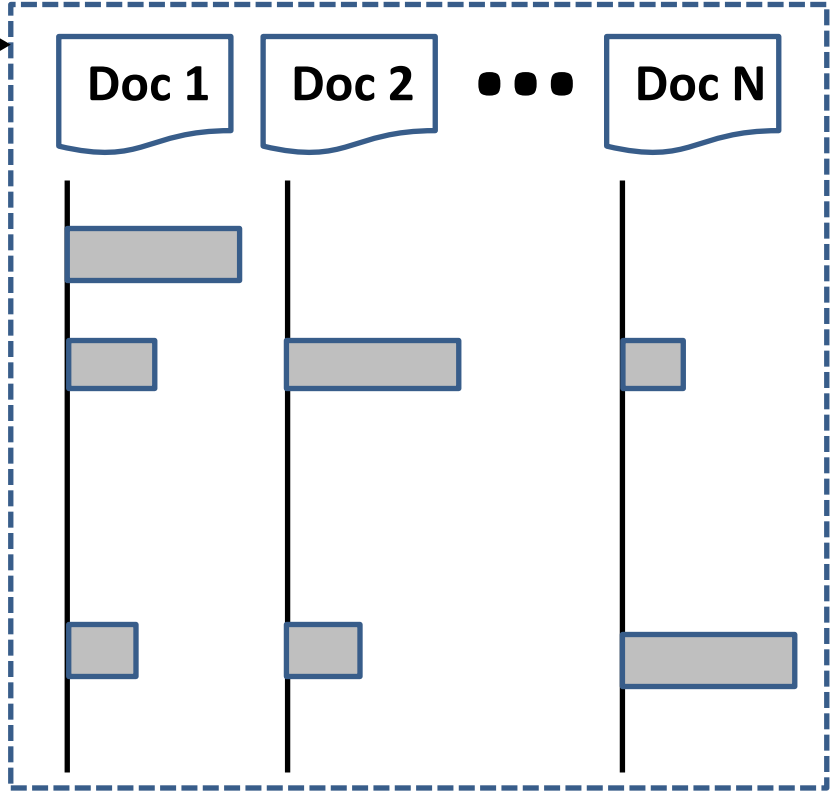
Tasks of Topic Mining and Analysis

Task 2: Figure out which documents cover which topics









Task 1: Discover k topics

Formal Definition of Topic Mining and Analysis

- Input
  - A **collection** of **N** text documents **C={d_1, …, d_N}**
  - **Number of topics**: **k**
- Output
  - **k topics**: **{ θ_1, …, θ_k }**
  - **Coverage of topics in each d_i**: **{ π_{i1}, …, π_{ik} }**
  - π_{ij} = prob. of d_i covering topic θ_j

probabl

$$\sum_{j=1}^{k} \pi_{ij} = 1$$

How to define θ_i ?