Discovery

EDA: Exploratory Data Analysis

Week 09

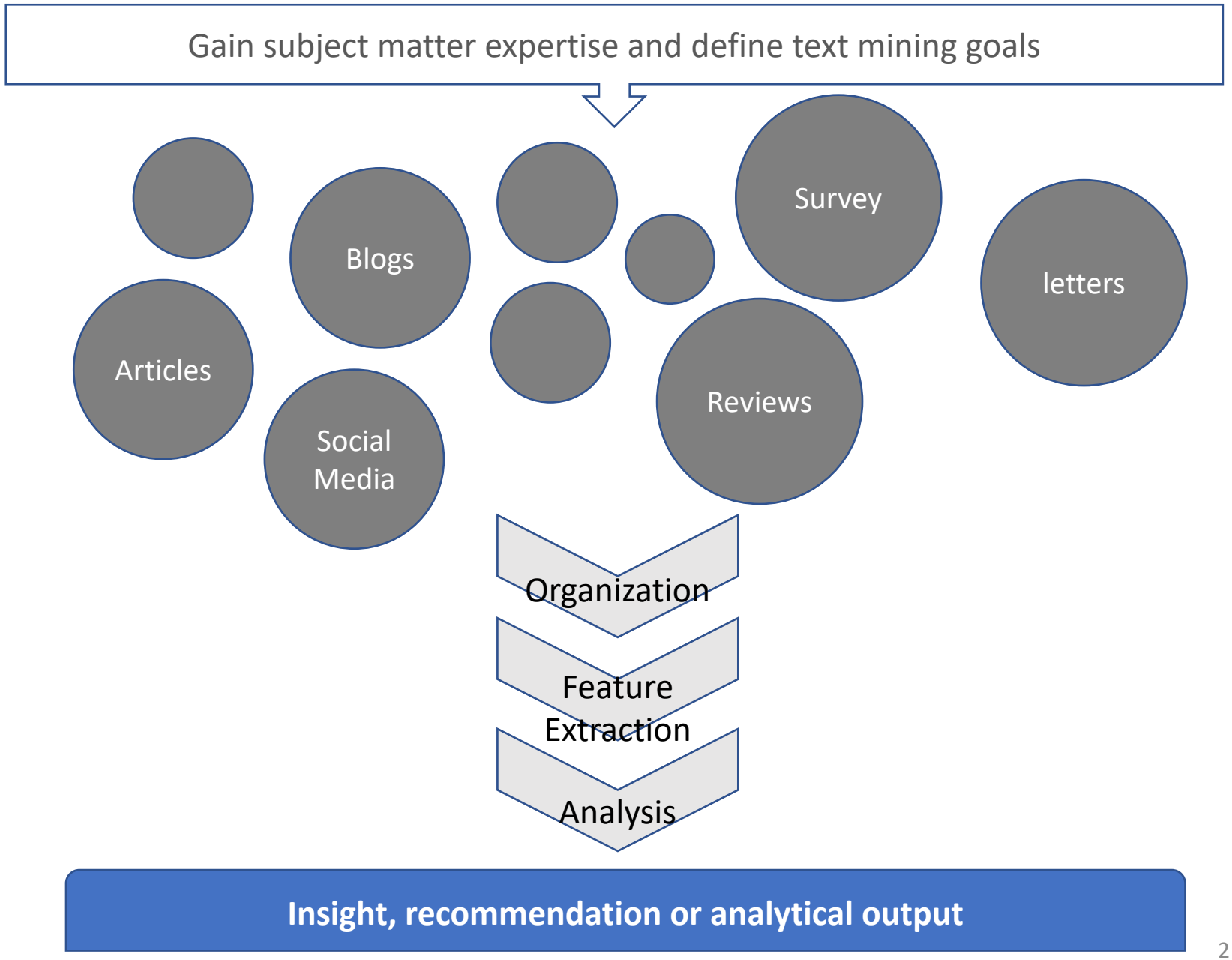
GSIAS, HUFS, Data Mining for Social Science

Basic Workflow

Problem Definition

Unorganised State

Organized State



Corpus

It may refer to any collection of linguistic data (such as written, spoken, signed, or multimodal). Furthermore, a text corpus is a large, unstructured collection of texts (typically stored and processed electronically today) used for statistical analysis and hypothesis testing, checking occurrences and validating linguistic rules within a specific language territory.

Token

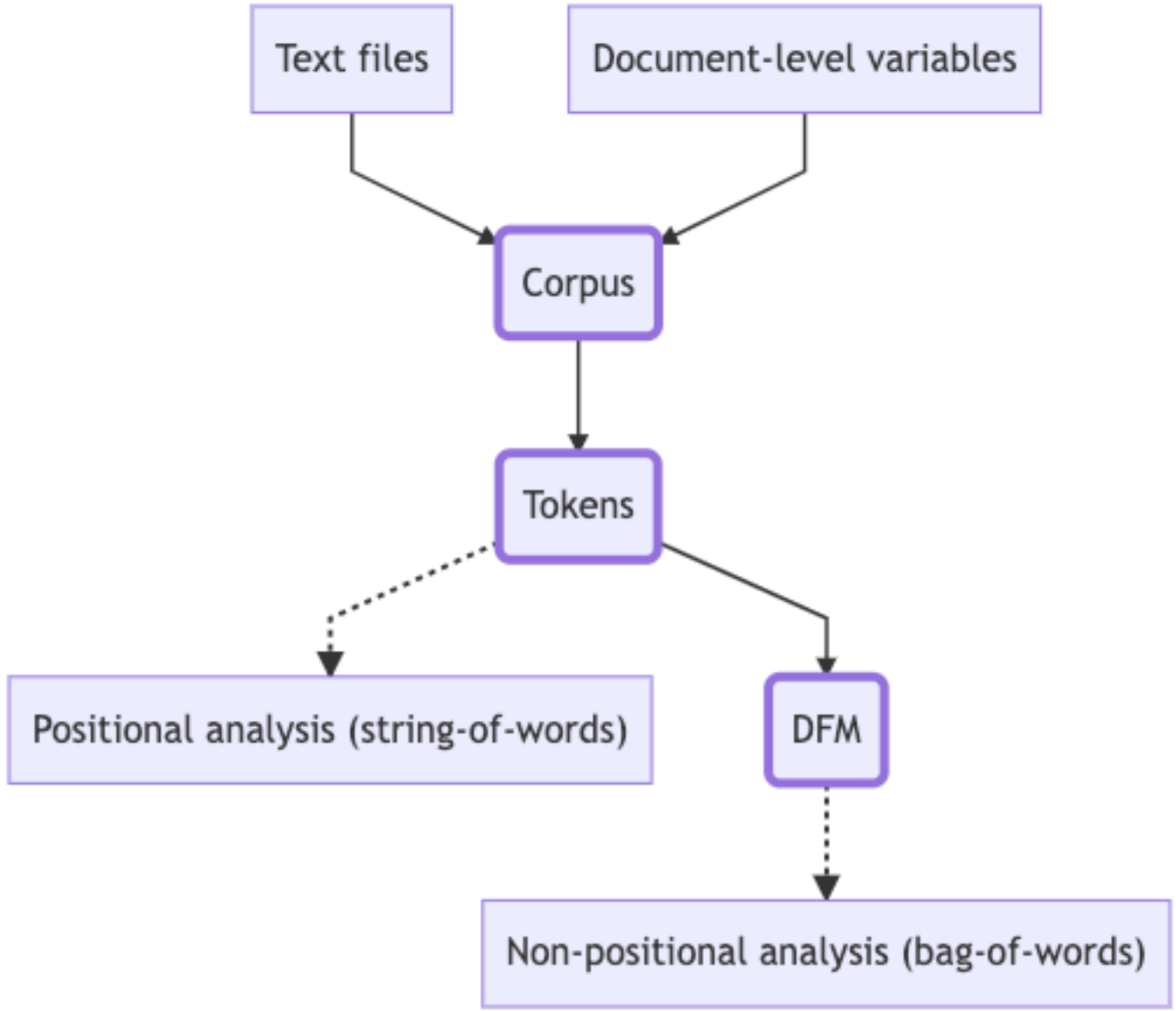
In natural language, a "token" is "an instance of a sequence of characters in a particular document that are grouped together as a useful semantic unit for processing." Like a tree's roots and branches, language is a jumble of natural growths that are split, dying, growing, and blooming. Tokenization is a part of the way we teach machines how to understand words, which is the most important part of our most important invention.

Document feature matrix (or Term-Document Matrix)

"A document feature or term-document matrix is a mathematical matrix that describes the frequency of terms that occur in a collection of documents. In a document-term matrix, rows correspond to documents in the collection, and columns correspond to terms. There are various schemes for determining the value each matrix entry should take. One such scheme is tf-idf. They are useful in the field of natural language processing."

$$\begin{bmatrix} & T_1 & T_2 & \cdots & T_t \\ D_1 & w_{11} & w_{21} & \cdots & w_{t1} \\ D_2 & w_{12} & w_{22} & \cdots & w_{t2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ D_n & w_{1n} & w_{2n} & \cdots & w_{tn} \end{bmatrix}$$

Document Term Matrix


Text files
Document-level variables
Corpus
Tokens
Positional analysis (string-of-words)
DFM
Non-positional analysis (bag-of-words)

Next Week

Please bring your laptop.

Week 10: Common data mining visualizations
Advanced text manipulation in R, word clouds, co-occurrence networks, sentiment analysis, word s core, and wordfish.

- Co-occurrence network
- Sentiment analysis
- Word score and wordfish