Text Analytics

Collaborative Learning and Development

(Session 2)

Types of Text Mining

Primary Types

Focus

Text Analysis: Given a document of free text, how can it be summarized ?

Examples:

- Information extraction
- Text summarization
- Text clustering
- Topic Modeling and classification

Text Retrieval \ Search: Given a set of documents containing text, find those closest to an inquiry.

Examples of Searching for:

- Information in documents
- Documents themselves
- Metadata describing documents
- Specific text, images or data in a database

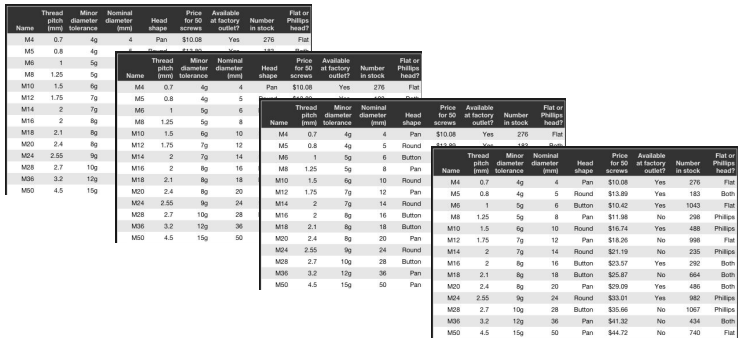
Applications of Text Analysis

Examples:

- Opinion and Sentiment analysis
- Feature and frequency summarization for modeling
- VOC summarization
- Post contact churn prediction
- Automatic and dynamic customer service interactions
- Summarization of clinician medical notes
- Entity extraction – names, companies, phone numbers, emails, URLs
- Customer journey mapping
- Fraud detection
- Language translation

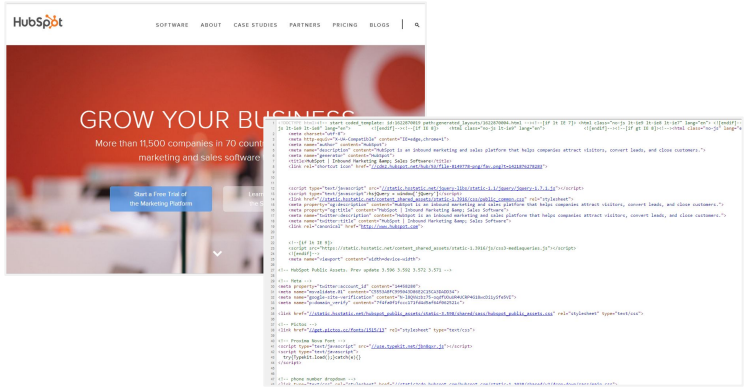
Text is Unstructured Data

Structured Data



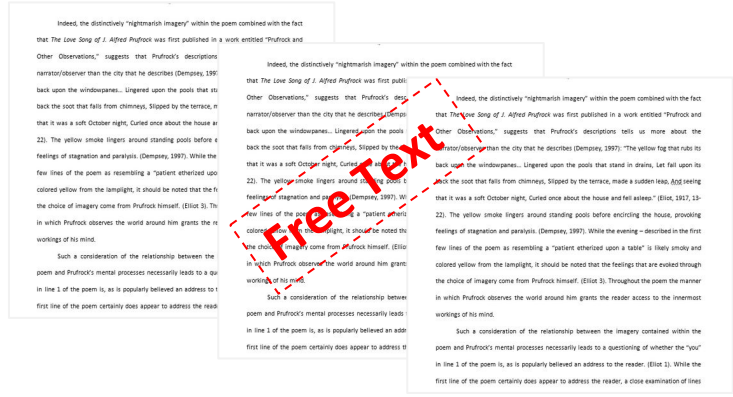
Data in **tabular** form (rows and columns) in .csv files, Excel files, relational databases – what you normally think of as consumable data.

Semi-structured Data



Data has “some” structure but is not directly query able with SQL including: a website, markup languages, **attribute / key – value pairs**.

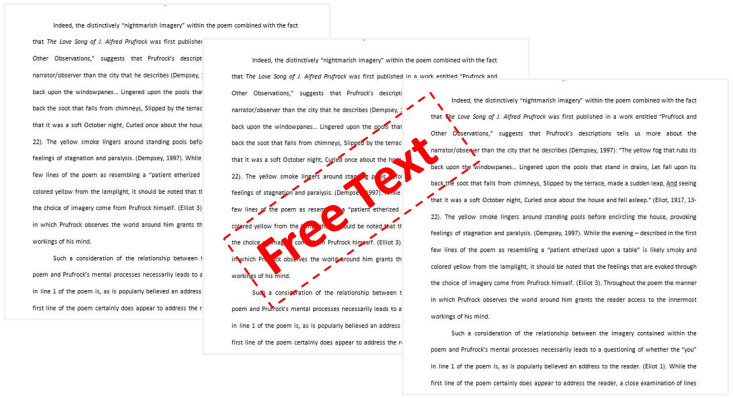
Unstructured Data



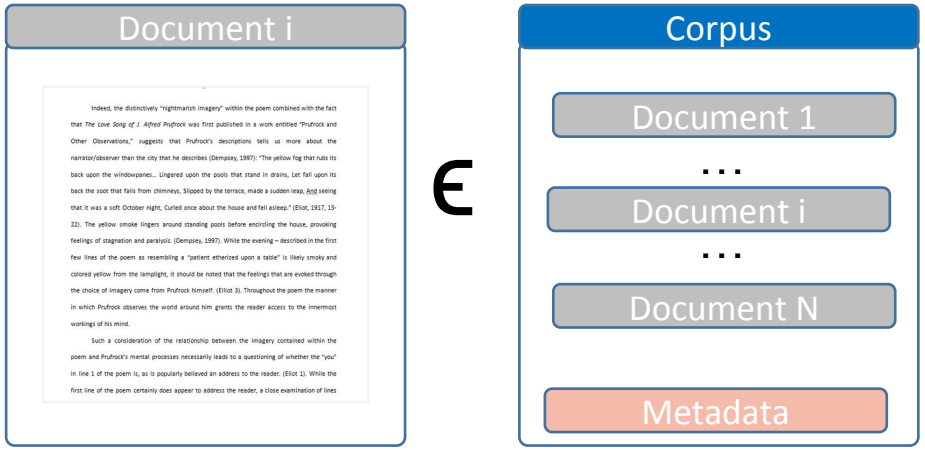
Data that is unstructured includes: **free form text** often found in survey responses, PDFs, emails, transcribed calls / chats, images, social media.

Structuring Text data

1 Unstructured Data



2 Organize



Some Structuring Methods

1 Term Document Matrix

Matrix construction of term frequencies per document (row = $document_i$, col = $term_i$, cell = count of $term_i$ in $document_i$).

	Term 1	Term i	Term N
Doc 1	#	#	#
Doc i	#	#	#
Doc N	#	#	#

2 Positions, nGrams and Spans

"indexing" words by position with hierarchies building up from a single character to a word, to combination of words to distances between words, sentences etc.

"This is a sentence?"

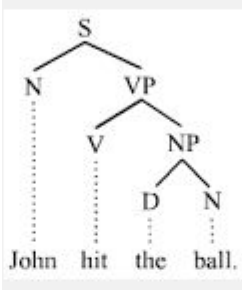
Character	
Index	Value
1	T
2	H
3	i
...	
19	?

Word	
Index	Value
1	This
2	is
3	a
4	sentence
5	?

Spans		
Start	End	Value
1	4	This
5	5	SPACE
6	7	is
8	8	SPACE
9	9	a
10	10	SPACE
11	18	sentence
19	19	?

3 Parse Trees

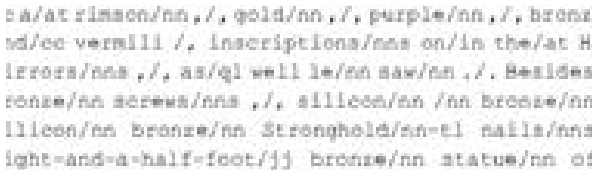
The structuring of Natural Language of text build with a set of rules and dictionaries.



4 Tagging

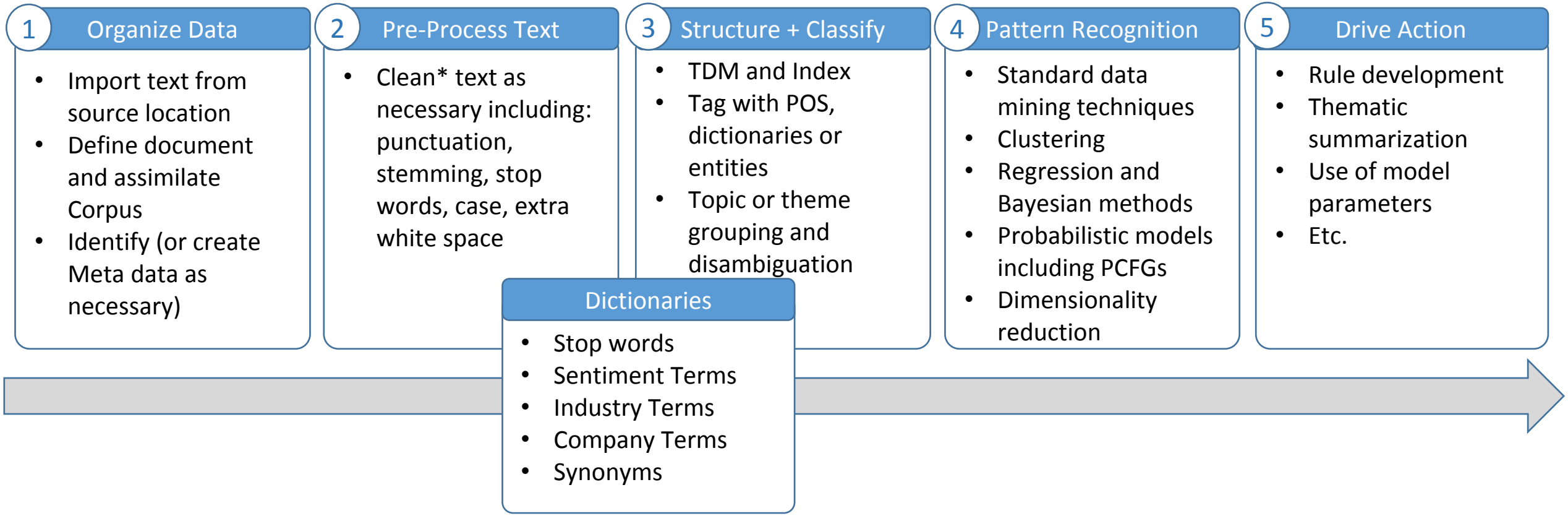
Identifying specific words and tagging their meanings in context:

- part of speech (Nouns, Verbs, etc.)
- entities (people, companies, etc.)
- belonging to dictionaries (sentiment, industry lexicons, etc.)



Text Analysis Process – High-Level

Assuming a coherent business question is formed with some potential success metrics and hypothetical courses of action, Here sare some high-level steps that occur during a text analysis process. Though importing is necessary, all other steps are optional and dependent on the analysis or business question at hand.



Text Analysis - Application Discussion

Free text responses from Survey

Scenario: NPS surveys are sent to customers with an additional free text question of: “Why did you give that score?”. How can you leverage these responses?

Data considerations: Data will be short not meant to be grammatically correct. Expect typos, abbreviations, incomplete sentences.

Potential VAIs:

- Identify VOC from extracted themes or topics provided to product teams and sr. mgmt
- Allow for predefined dictionaries of key terms and categorization of responses accordingly delivered to target departments
- Automatically route responses requesting help or claiming terrible service.
- Summarize and trend categorizations of topics to measure improvement on customer pain points

Call transcriptions

Scenario: Voice records of conversations are transcribed into text and saved in log files for support calls between customers and agents. How can you leverage these conversations?

Data considerations: Transcriptions are a computers translation of the human voice. Expect incorrect translations, “back and forths”, missed words and for terrible grammar – these are conversations, not essays.

Potential VAIs:

- Sentiment modeling for customer experience scoring and NPS proxying
- Post contact (or in contact) churn risk identification
- Agent coaching and improvement
- Enhanced call topic classification and tracking
- Purchase scoring (lead listing)
- In call support recommendations and knowledge base lookup

Web content

Scenario: Customers (or competitors) create websites as part of their online presence, though websites can be considered “semi-structured” the actual content or text within the html tags is not. How can you glean useful information assuming you can collect that text from their websites?

Data considerations: This is written text, in most cases, intention to look professional. However, it will be chunked and distributed in multiple locations on a site with little to no predefined context other than “this is a website of something”.

Potential VAIs:

- Extraction of useful data including phone number, email addresses, company name
- Classification of website (i.e. personal , business, or type of business, or tech savvy vs. novice)
- Needs of website not met – (i.e.) collects payments online, but does not have an SSL (i.e.)
- Product, price, feature and promotion extraction of competitors (of ours or of target customers)

Code Bank (R)