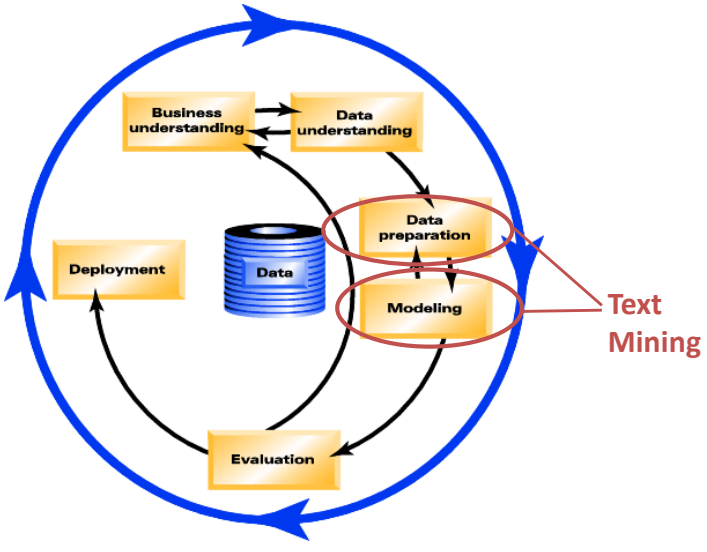
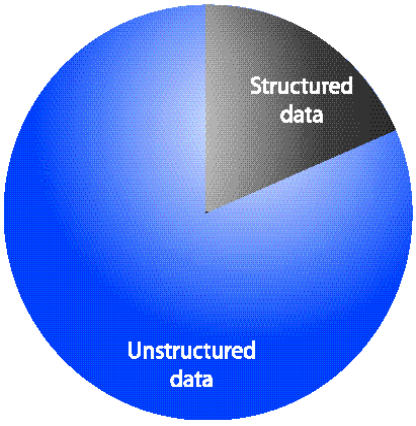
Data Mining

Unit # 10

Back to the Process



80% of Data is Unstructured



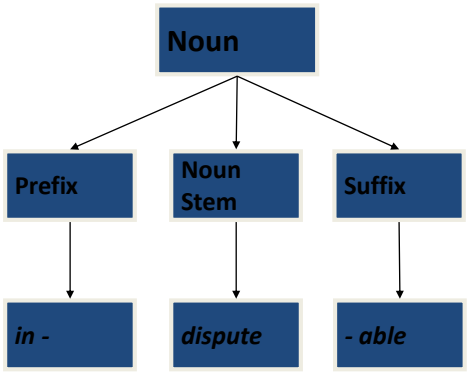
- Database notes:
  - Call center transcripts
  - Other CRM
- Email
- Open-ended survey responses
- Web pages
- NewsGroups
- Social Media

Major Steps in a Text Analytics Process

- Tagging
  - Name Entity Tagging
  - Part-of-Speech Tagging
- Pre-Processing
  - Punctuation Eraser
  - Number Filter
  - N-Char Filter
  - Stop world Filter
- Frequency
- Tag Cloud

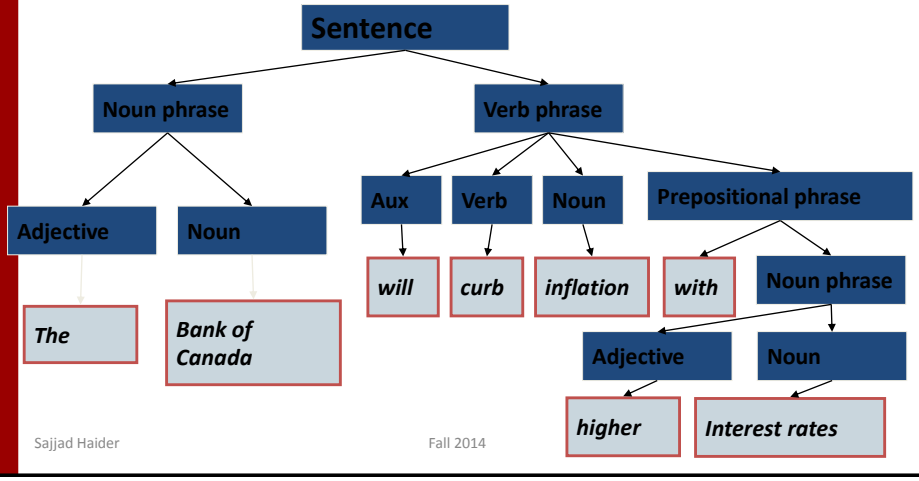
Morphology

- Understanding words
  - Stems
  - Affixes
    - Prefix
    - Suffix
  - Inflectional elements
- Reducing complexity of analysis
- Reduces complexity of representation
- Supports text mining



Syntax

- *The Bank of Canada will curb inflation with higher interest rates*



Sajjad Haider Fall 2014

Part-of-Speech Tagging

a: adjective	b: adverb	c: preposition
d: determiner	n: noun	v: verb
o: coordination	p: participle	s: stop word

How is a Concept Extracted?

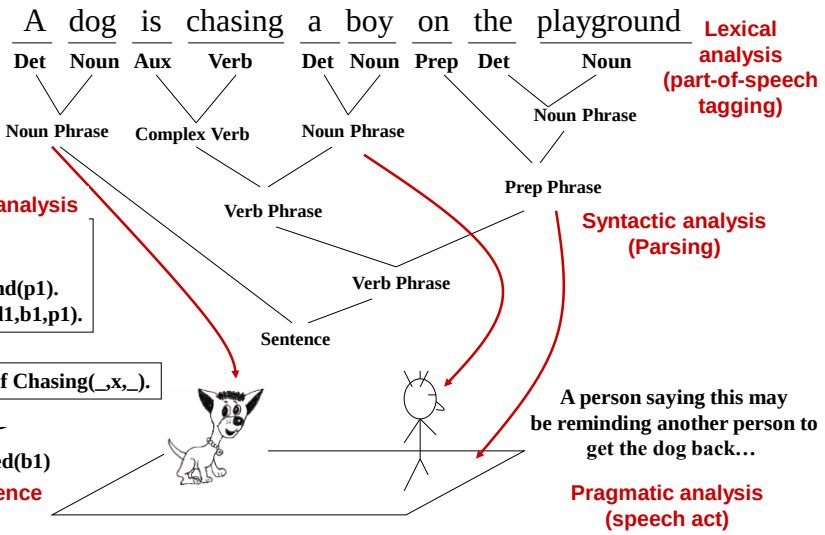
Step 1: Part-of-Speech Tagging

Using	a	tool	like	LexiQuest	Mine	is	a	great
V	P	N	A	N	N	V	P	A

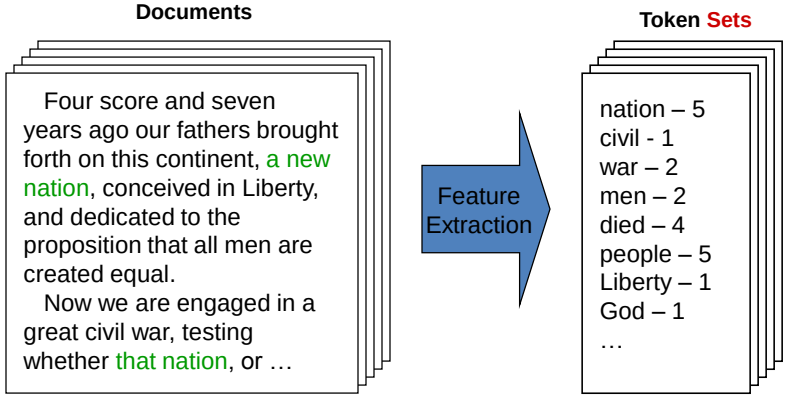
idea	for	any	organization	that	is	interested	in	maintaining
N	P	A	N	P	V	V	P	V

information	on	competitive	intelligence.
N	P	N	N

Natural Language Processing



Bag-of-Tokens Approaches



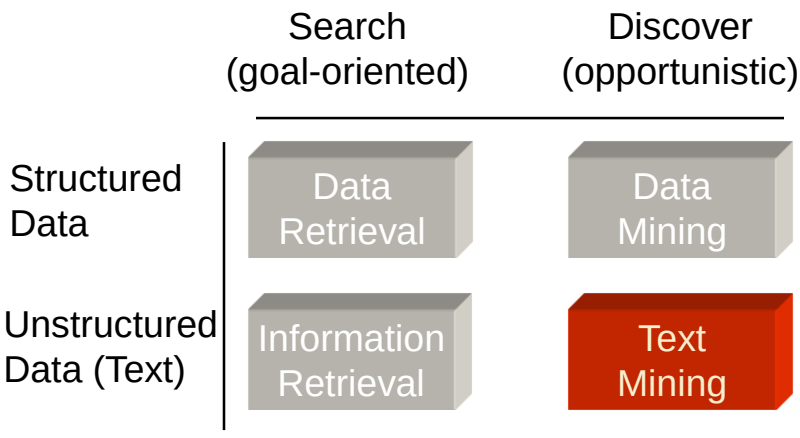
Loses all order-specific information!
Severely limits context!

General NLP—Too Difficult!

- Word-level ambiguity
  - **"design" can be a noun or a verb** (Ambiguous POS)
  - **"root" has multiple meanings** (Ambiguous sense)
- Syntactic ambiguity
  - **"natural language processing"** (Modification)
  - **"A man saw a boy *with a telescope*."** (PP Attachment)
- Anaphora resolution
  - **"John persuaded Bill to buy a TV for *himself*."**
    (*himself* = John or Bill?)
- Presupposition
  - **"He has quit smoking." implies that he smoked before.**

Humans rely on context to interpret (when possible).
This context may extend beyond a given document!

"Search" versus "Discover"



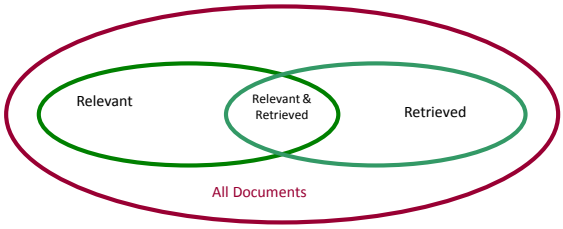
Text Databases and IR

- Text databases (document databases)
  - Large collections of documents from various sources: news articles, research papers, books, digital libraries, e-mail messages, and Web pages, library database, etc.
  - Data stored is usually *semi-structured*
- Information retrieval
  - A field developed in parallel with database systems
  - Information is organized into (a large number of) documents
  - Information retrieval problem: locating relevant documents based on user input, such as keywords or example documents

Information Retrieval

- Typical IR systems
  - Online library catalogs
  - Online document management systems
- Information retrieval vs. database systems
  - Some DB problems are not present in IR, e.g., update, transaction management, complex objects
  - Some IR problems are not addressed well in DBMS, e.g., unstructured documents, approximate search using keywords and relevance

Basic Measures for Text Retrieval



- Precision: the percentage of retrieved documents that are in fact relevant to the query (i.e., "correct" responses)

$$precision = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Retrieved\}|}$$

- Recall: the percentage of documents that are relevant to the query and were, in fact, retrieved

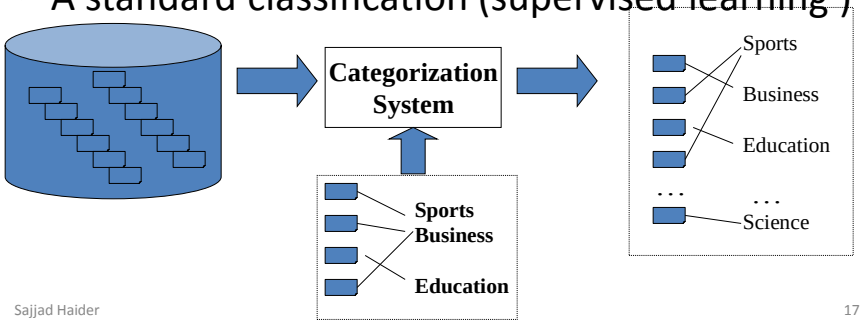
$$recall = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Relevant\}|}$$

Document Clustering

- Motivation
  - Automatically group related documents based on their contents
  - No predetermined training sets or taxonomies
  - Generate a taxonomy at runtime
- Clustering Process
  - Data preprocessing: remove stop words, stem, feature extraction, lexical analysis, etc.
  - Hierarchical clustering: compute similarities applying clustering algorithms.
  - Model-Based clustering (Neural Network Approach): clusters are represented by "exemplars". (e.g.: SOM)

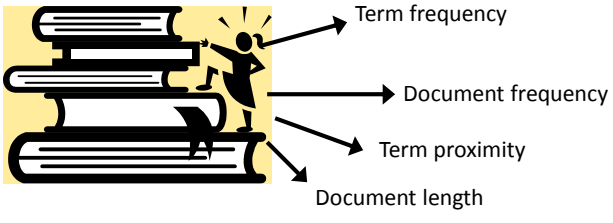
Document Categorization

- Pre-given categories and labeled document examples (Categories may form hierarchy)
- Classify new documents
- A standard classification (supervised learning)



Statistical Analysis

- Use statistics to add a numerical dimension to unstructured text



Computing Relevance

- Call up all the documents that have any of the terms from the query, and count how many times each term occurs:

$$\text{Relevance}_{document} = \sum_{q_i} tf_{q_i}$$

Inverse Document Frequency (idf)

$$idf_i = \log (N/tf_i)$$

- N: Number of documents in corpus
- tf_i: Number of documents in which term occurs in the corpus

- Measures term uniqueness in corpus
  - "phone" vs. "brick"
- Indicates the importance of the term
  - Search (relevance)
  - Classification (discriminatory power)

TF-IDF and Modified Retrieval Algorithm

- Term frequency – inverse document frequency (tf-idf)

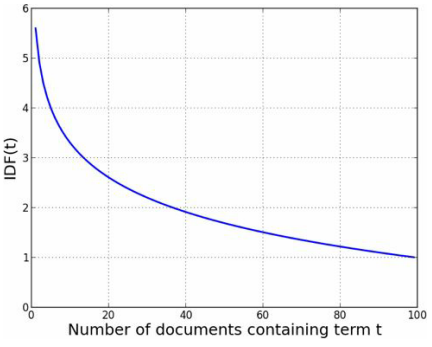
  $tf_{document}$(term) * idf(term)

  query: *"unbrick phone"*
- Document with "unbrick" a few times more relevant than document with "phone" many times
- Measure of Relevance with tf-idf
- Call up all the documents that have any of the terms from the query, and sum up the tf-idf of each term:

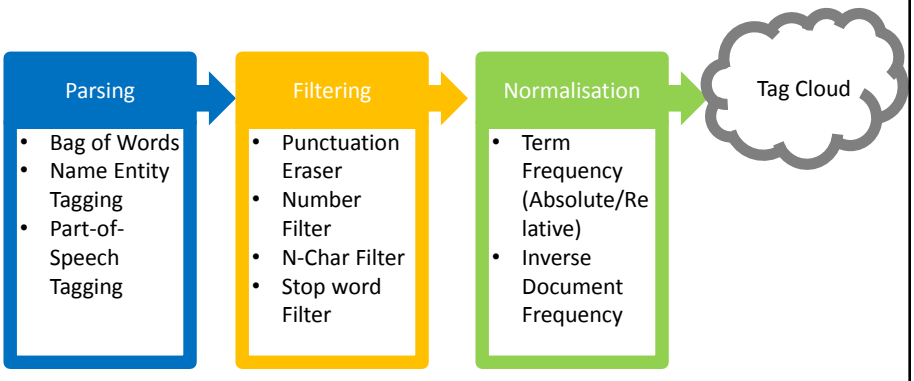
$$\text{Relevance}_{document} = \sum_{q_i} tfidf_{q_i}$$

TF-IDF

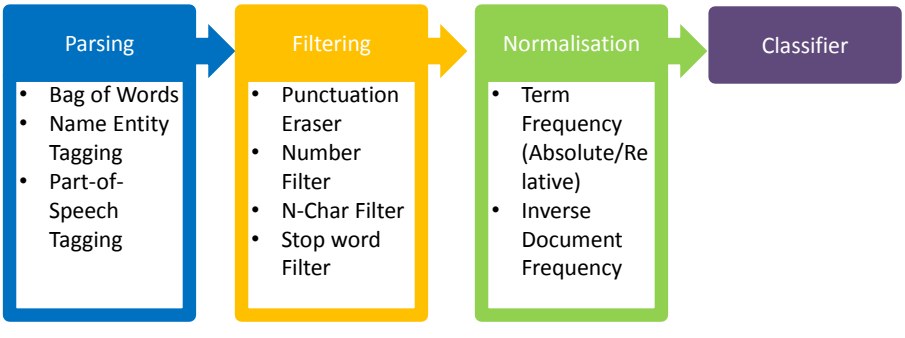
- LESS 'important' words occur MORE often
- Words can be weighted by "their inverse document frequency (IDF)"



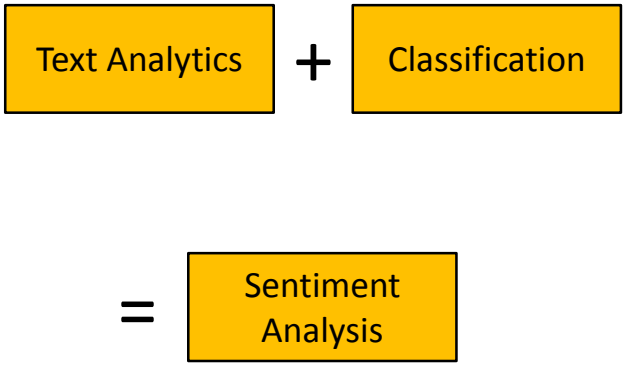
Major Steps



Major Steps



Sentiment Analysis



Tag Clouds

- A **tag cloud** (**word cloud**) is a visual representation for text data, typically used to depict keyword metadata (tags) on websites, or to visualize free form text.
- Tags are usually single words, and the importance of each tag is shown with font size or color.
- This format is useful for quickly perceiving the most prominent terms and for locating a term alphabetically to determine its relative prominence.

TagCrowd.com

action agents algorithm annotation based bayesian best bipedal blend building collaboration comparison coordinated data dynamic evolving extraction files finding fourier frameworks imitation influence information intelligence kick modelling nets networks ontology ontology-based opponent optimization particle portfolio reference-set-based reporting robocup robot rule-based semantic simulation skills soccer sources strategies study testing unstructured via

Wordle.com (1)

Wordle.com (2)

Application of Text Analytics

- Document Classification
  - E-mail Filtering
- Sentiment Analysis
  - Twitter
  - Reviews
- Visualization
- Document Clustering

TEXT ANALYTICS HANDS-ON

Twitter Analysis in KNIME

- Sign up for a Twitter account.
- Log in to Twitter Developers
- Go to My Applications
- Create a new application
- On the application page, click the 'Create my access token' button, wait some seconds and refresh the page
- Enter the following data in the KNIME settings: Consumer key, Consumer secret, Access token, Access token secret