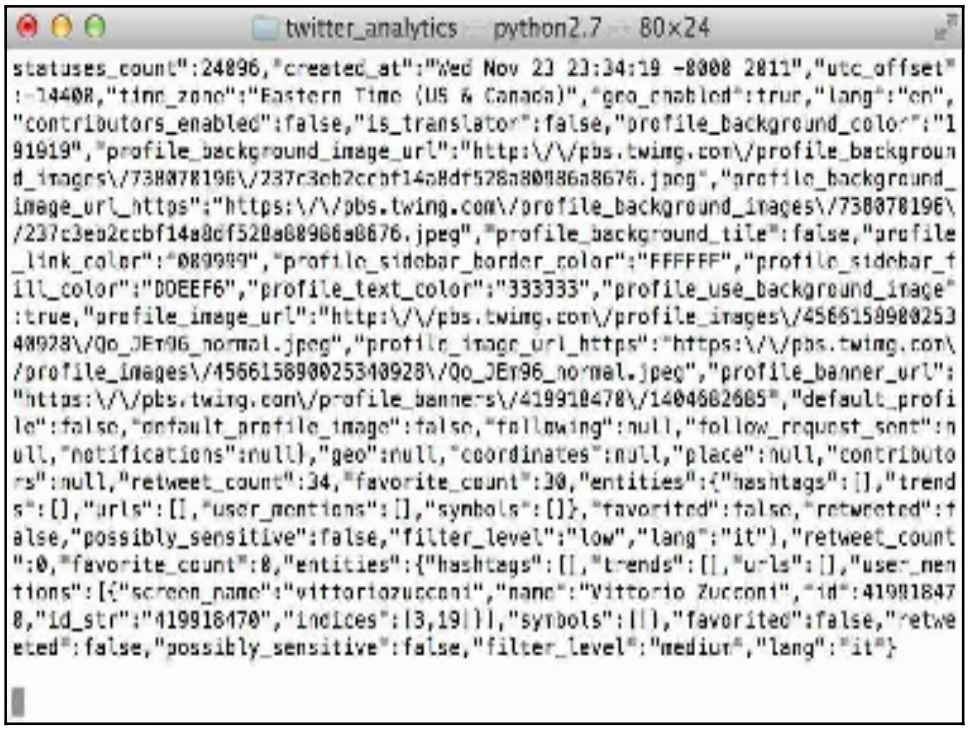

text analysis
text mining


twitter_analytics — python2.7 — 80×24
statuses_count":24896,"created_at":"Wed Nov 23 23:34:19 +0000 2011","utc_offset"
:-14400,"time_zone":"Eastern Time (US & Canada)","geo_enabled":true,"lang":"en",
"contributors_enabled":false,"is_translator":false,"profile_background_color":"1
91919","profile_background_image_url":"http:\/\/pbs.twimg.com\/profile_backgroun
d_images\/738078196\/237c3eb2ccbf14a8df528a80986a8676.jpeg","profile_background_
image_url_https":"https:\/\/pbs.twimg.com\/profile_background_images\/738078196\
/237c3eb2ccbf14a8df528a80986a8676.jpeg","profile_background_tile":false,"profile
_link_color":"089999","profile_sidebar_border_color":"FFFFFF","profile_sidebar_f
ill_color":"DDEEF6","profile_text_color":"333333","profile_use_background_image"
:true,"profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/4566150980253
40928\/Qo_JEr96_normal.jpeg","profile_image_url_https":"https:\/\/pbs.twimg.com\
/profile_images\/456615098025340928\/Qo_JEr96_normal.jpeg","profile_banner_url":
"https:\/\/pbs.twimg.com\/profile_banners\/419918478\/1404682685","default_profi
le":false,"default_profile_image":false,"following":null,"follow_request_sent":n
ull,"notifications":null},"geo":null,"coordinates":null,"place":null,"contributo
rs":null,"retweet_count":34,"favorite_count":30,"entities":{"hashtags":[],"trend
s":[],"urls":[],"user_mentions":[],"symbols":[]},"favorited":false,"retweeted":f
alse,"possibly_sensitive":false,"filter_level":"low","lang":"it"},"retweet_count
":0,"favorite_count":0,"entities":{"hashtags":[],"trends":[],"urls":[],"user_men
tions":[{"screen_name":"vittoriozucconi","name":"Vittorio Zucconi","id":41981847
8,"id_str":"419918470","indices":[3,19]}],"symbols":[]},"favorited":false,"retwe
eted":false,"possibly_sensitive":false,"filter_level":"medium","lang":"it"}

Text Mining

" The objective of Text Mining is to exploit information contained in textual documents in various ways, including …discovery of patterns and trends in data, associations among entities, predictive rules, etc." (Grobelnik et al., 2001)

" Another way to view text data mining is as a process of exploratory data analysis that leads to heretofore unknown information, or to answers for questions for which the answer is not currently known." (Hearst, 1999)

Text Mining Definition

"The non trivial extraction of implicit, previously unknown, and potentially useful information from (large amount of) textual data".

- An exploration and analysis of textual (natural-language) data by automatic and semi automatic means to discover new knowledge.

- *Previously unknown Means*
  - Strict definition
    - Information that not even the writer knows.
  - Lenient definition
    - Rediscover the information that the author encoded in the text

Data Retrieval

- Find records within a structured database.

Database Type	Structured
Search Mode	Goal-driven
Atomic entity	Data Record
Example Information Need	"Find a Japanese restaurant in Boston that serves vegetarian food."
Example Query	"SELECT * FROM restaurants WHERE city = boston AND type = japanese AND has_veg = true"

Information Retrieval

Find relevant information in an unstructured information source (usually text)

Database Type	Unstructured
Search Mode	Goal-driven
Atomic entity	Document
Example Information Need	"Find a Japanese restaurant in Boston that serves vegetarian food."
Example Query	"Japanese restaurant Boston" or Boston->Restaurants->Japanese

Data Mining

Discover new knowledge through analysis of data

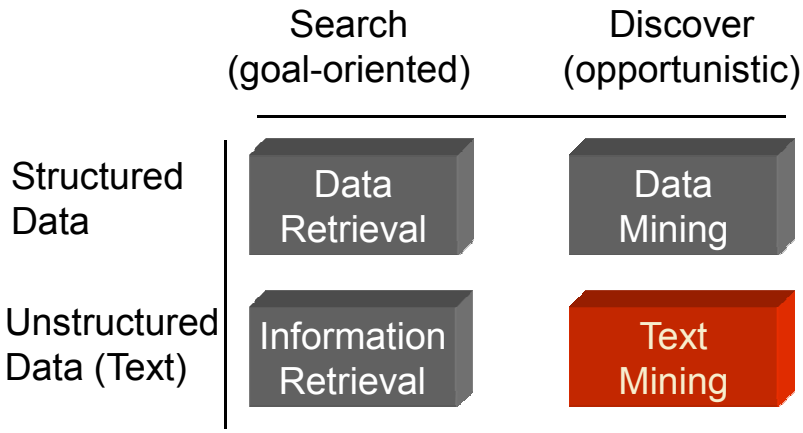
Database Type	Structured
Search Mode	Opportunistic
Atomic entity	Numbers and Dimensions
Example Information Need	"Show trend over time in # of visits to Japanese restaurants in Boston "
Example Query	"SELECT SUM(visits) FROM restaurants WHERE city = boston AND type = japanese ORDER BY date"

Text Mining

Discover new knowledge through analysis of text

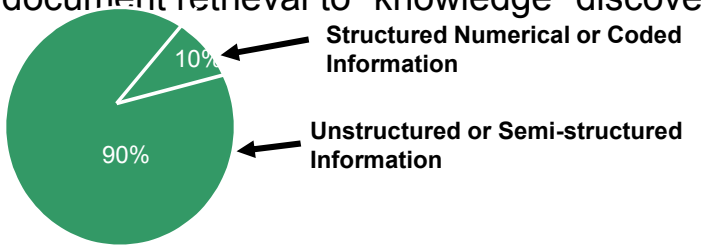
Database Type	Unstructured
Search Mode	Opportunistic
Atomic entity	Language feature or concept
Example Information Need	"Find the types of food poisoning most often associated with Japanese restaurants"
Example Query	Rank **diseases** found associated with "Japanese restaurants"

"Search" versus "Discover"



Motivation for Text Mining

- Approximately **90%** of the world's data is held in unstructured formats (source: Oracle Corporation)
- Information intensive business processes demand that we transcend from simple document retrieval to "knowledge" discovery.



Text characteristics

- Dependency
  - relevant information is a complex conjunction of words/phrases
    - e.g., Document categorization.
      Pronoun disambiguation.
- Ambiguity
  - Word ambiguity
    - Pronouns (he, she …)
    - "buy", "purchase"
  - Semantic ambiguity
    - The king saw the rabbit with his glasses. (8 meanings)

Text characteristics

- Large textual data base
  - Efficiency consideration
    - over 2,000,000,000 web pages
    - almost all publications are also in electronic form
- High dimensionality (Sparse input)
  - Consider each word/phrase as a dimension
- Several input modes
  - e.g., Web mining: information about user is generated by semantics, browse pattern and outside knowledgebase.

Text characteristics

- Noisy data
  - Example: Spelling mistakes
- Not well structured text
  - Chat rooms
    - "r u available ?"
    - "Hey whazzzzzz up"
  - Speech

Text Mining Methods

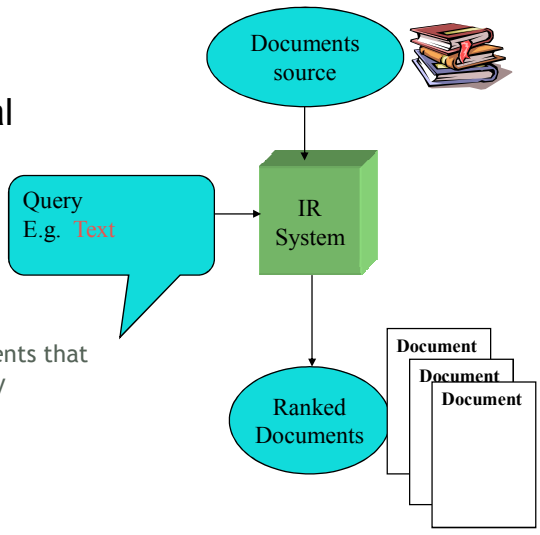
- Information Retrieval
  - Indexing and retrieval of textual documents
  - Include online library catalog
- Information Extraction
  - Extraction of partial knowledge in the text

Information retrieval (IR)

- **Information retrieval** (**IR**) is the area of study concerned with searching for documents, for information within documents, and for metadata about documents, as well as that of searching relational databases and the World Wide Web.
- Web search engines are the most visible IR applications

Information Retrieval

- **Given:**
  - A source of textual documents
  - A user query (text based)
- **Find:**
  - A set (ranked) of documents that are relevant to the query



Intelligent Information Retrieval

- ***meaning*** of words
  - Synonyms "buy" / "purchase"
  - Ambiguity "bat" (baseball vs. cricket)
- ***order*** of words in the query
- ***user dependency*** for the data
  - direct feedback
  - indirect feedback
- ***authority*** of the source
  - IBM is more likely to be an authorized source then my second far cousin

Information Extraction

- **Information extraction** (IE) is a type of information retrieval whose goal is to automatically extract structured information from unstructured and/or semi-structured machine-readable documents.
- In most of the case this activity concern processing human language texts by means of natural language processing (NLP).
- Goal is to allow logical reasoning to draw inferences based on the logical content of the input data

Subtask of IE

1. Named entity extraction which could include:
   a) Named entity recognition: recognition of known entity names (for people and organizations) PERSON located in LOCATION (extracted from the sentence "Bill is in France.")
   b) Relationship extraction: identification of relations between entities, such as: PERSON works for ORGANIZATION (extracted from the sentence "Bill works for IBM.")

2. Semi-structured information extraction : Table extraction: finding and extracting tables from documents. Comments extraction : extracting comments from actual content of article

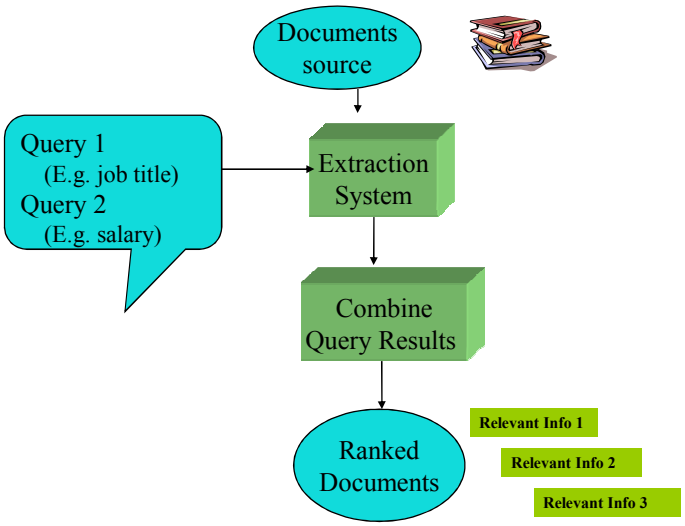
3. Terminology extraction

Terminology extraction: finding the relevant terms for a given corpus

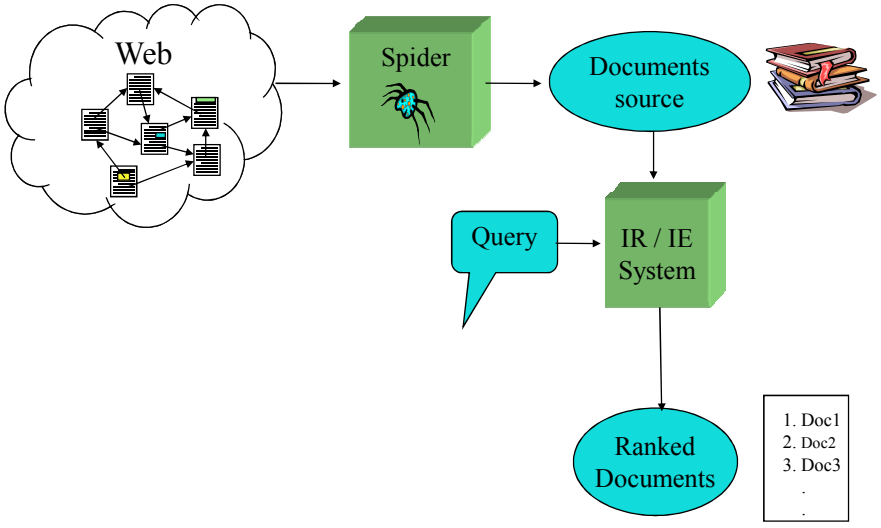
What is Information Extraction?

- **Given:**
  - A source of textual documents
  - A well defined limited query (text based)
- **Find:**
  - Sentences with relevant information
  - Extract the relevant information and ignore non-relevant information (important!)
  - Link related information and output in a predetermined format

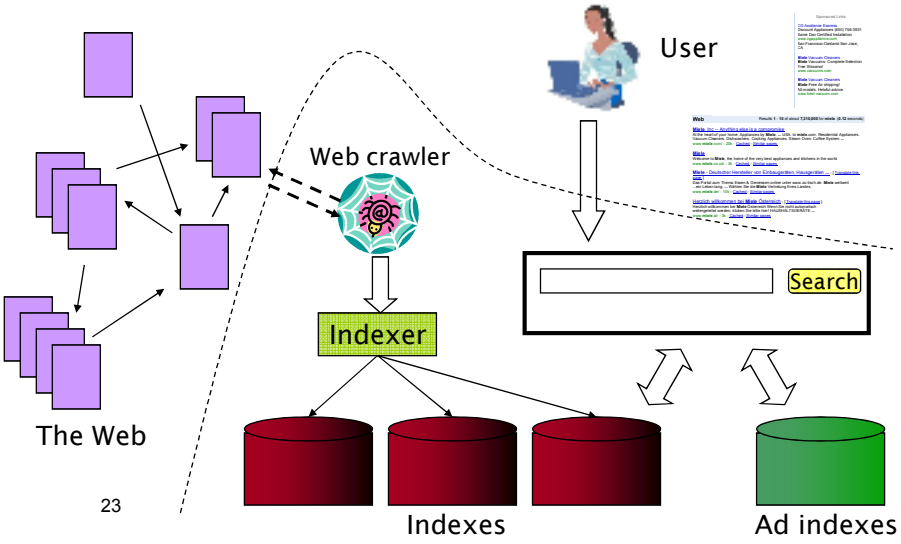
What is Information Extraction?



Mining the Web



Web search basics



23

Crawlers

- Robot (spider) is a program that traverses the hypertext sructure in the Web.
  - Collect information from visited pages
  - The Page(or pages) that the crawler starts with are referred to as the seed URL
  - Used to construct indexes for search engines
- Periodic Crawler – may visit a certain number of pages and then stop, built an index, and replace the existing index.
  - It is activated periodically and serach portions of the Web. (Notices and Circular)

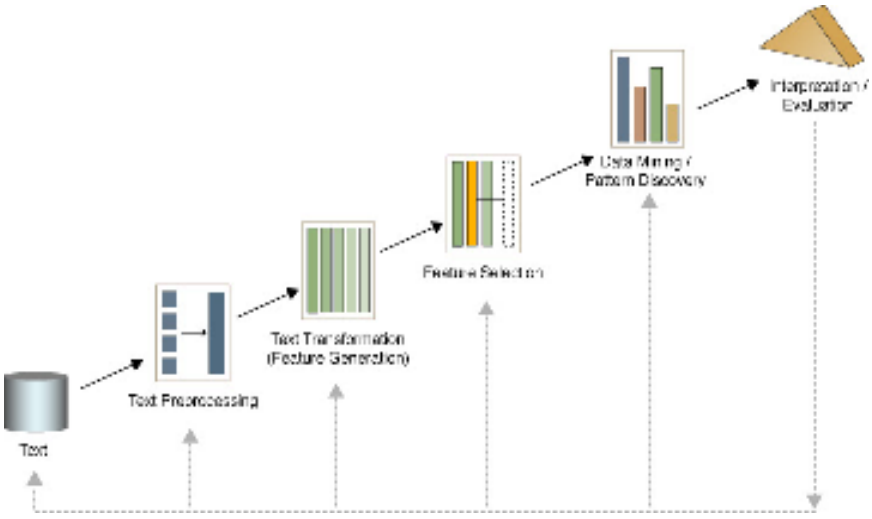
24

Crawlers

- **Incremental Crawler – selectively searches the Web and only updates the index incrementally as oppose to replacing them. incrementally modifies index**
- **Focused Crawler – visits pages related to a particular subject (Topic of Interest)**

25

Text mining process



Text mining process

- Text preprocessing
  - Syntactic/Semantic text analysis
- Features Generation
  - Bag of words
- Features Selection
  - Simple counting
  - Statistics
- Text/Data Mining
  - Classification- Supervised learning
  - Clustering- Unsupervised learning
- Analyzing results

Text Preprocessing

Text Cleanup

e.g., remove ads from web pages, normalize text converted from binary formats, deal with tables, figures and formulas, …

Tokenization

– Splitting up a string of characters into a set of tokens.

– Need to deal with issues like:

• Apostrophes, e.g., “John’s sick”, is it 1 or 2 tokens?

• Hyphens, e.g., database vs. data-base vs. data base.

• How should we deal with “C++”, “A/C”, “:-)”, “…”?

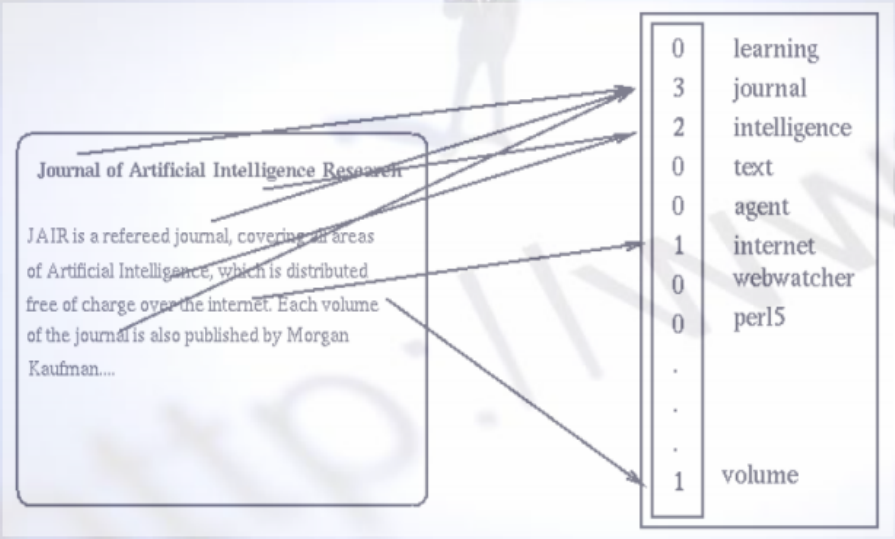
Syntactic / Semantic text analysis

- Part Of Speech (pos) tagging
  - Find the corresponding pos for each word
    e.g., John *(noun)* gave *(verb)* the *(det)* ball *(noun)*
  - *~98% accurate.*
- Parsing
  - Generates a parse tree (graph) for each sentence
  - Each sentence is a stand alone graph

Feature Generation: Bag of words

- Text document is represented by the words it contains (and their occurrences)
  - e.g., "Lord of the rings" → {"the", "Lord", "rings", "of"}
  - Highly efficient
  - Makes learning far simpler and easier
  - Order of words is not that important for certain applications
- Stemming: identifies a word by its root
  - e.g., flying, flew → fly
  - Reduce dimensionality
- Stop words: The most common words are unlikely to help text mining
  - e.g., "the", "a", "an", "you" …

"Bag of words" Document Representation



Word Weighting

- In "Bag of words" representation each word is represented as a separate variable having numeric weight.
- The most popular weighting schema is normalized word frequency *tfidf*:

$$tfidf(w) = tf \cdot \log\left(\frac{N}{df(w)}\right)$$

- *tf(w)* –term frequency (number of word occurrences in a document)
- *df(w)* –document frequency (number of documents containing the word)
- *N* –number of all documents
- *tfidf(w)* –relative importance of the word in the document

The word is more important if it appears several times in a target document

The word is more important if it appears in less documents

Text Mining Applications

- **Marketing:** Discover distinct groups of potential buyers according to a user text based profile
  - e.g. amazon
- **Industry:** Identifying groups of competitors web pages
  - e.g., competing products and their prices
- **Job seeking:** Identify parameters in searching for jobs
  - e.g., www.flipdog.com