## **Information processing**

**Questions:**

* **Distinguish between data, metadata, and information.**
  + - Data:
      * Measurement of something on a scale
      * A fact known by direct observation
    - Metadata:
      * Data about data
      * Not the content of data but data providing information about one or more aspects of the data, such as:
        + Description – data, time, author
        + Structure – format, version
        + Administrative – permissions
        + Legal
        + …
    - Information:
      * data with a context/ meaning thus enabling decision making
      * data that has been processed, organized, and structured
* **Identify and describe the phases of a typical information lifecycle.**
  + - Progress and welfare is increasingly dependent on the successful and efficient management of the life cycle of information.
      * Occurrence – discover, design, author, **…**
      * Processing and Management – collecting, validating, modifying, indexing, classifying, filtering, sorting, storing, …
      * Transmission – networking, accessing, retrieving, transmitting, …
      * Usage – monitoring, explaining, planning, forecasting, decision-making, educating, learning, …
    - Evolved from being mainly recording systems to being communication systems, to also currently being processing and producing systems.
* **Describe typical data processing patterns, pipelines and frameworks, e.g. ETL, EtlT, OSEMN.**
  + - Data Stages:
      * Raw – Focus on data discovery
        + The primary goals are ingestion, understanding and meta data creation.
        + Common questions include: what of records are in the data? How are record fields encoded?
      * Refined – Focus is on data preparation for further exploration
        + Removing unwanted parts
        + Reshaping poorly formatted elements
        + Establishing relationships between datasets
        + Assessing data quality issues
      * Production – focus is on integrating the data into production processes or products
    - **ELT Pattern** - Used in centralized IT driven operations
      * Extract - extracting data from the source system
      * Transform – a series of operations or transformations are applied to the extracted data
      * Load – publishing data to the target system (either simple flat files or other infrastructures)
    - **ELT (extract-load-transform) and EtLT framework** – Well-suited to the division of responsibilities in multidisciplinary teams
      * Recent evolution of ETL framework
      * Increasingly common access to data storage infrastructures capable of handling large volumes of data has led to a more flexible pattern
      * Column-oriented data structures are particularly well-suited to typical data processing tasks (organizing operations per field or property)
      * The sub-patter EtLT introduces a transformation step before the loading, typically associated with data cleaning tasks
    - **OSEMN Framework**
      * Obtain – gathering data
      * Scrub – clear, arrange, prepare data
      * Explore – observe, experiment, visualize
      * Model – create a statistical model of data
      * Interpret – draw conclusions, evaluating and communicating results
      * Although presented as a series of steps, real-world processes are typically non-linear, and this involves an iterative process
* **Describe the challenges associated with data processing.**
  + - **Data selection – Things to consider:**
      * Is the author a trustworthy source that can be contacted?
      * Is the data regularly updated?
      * Does the data include information about how and when it was acquired?
      * Does it seem plausible through observation?
    - **Diversity of Data Sources:**
      * Ownership – either owned or from third parties (know what data you have access to and what you can do with it)
      * Ingestion interface and structure – how do you get the data and in what form is it.
        + **Ingestion Interfaces examples:**

Relational database

Layer of abstraction (REST API)

Endpoint to a message queue system (RabbitMQ)

Shared network file system (CSV files or other flat files)

* + - * Volume - in each step of the pipeline volume needs to be considered
      * Cleanliness and validity
        + Missing data
        + Inconsistent values
        + Precision problems
        + Duplicate values
        + Others – text encoding issues, mislabled data, incomplete, outdated, …
      * Latency and bandwidth of the source – need to consider internal update requirements and source system limits, speed, timeouts, …
    - Outliers – Items that differ significantly from others
      * Necessary to understand if these values are exceptions but valid cases of errors that need to be removed (expert domain-knowledge is necessary)
      * Errors may be the result of:
        + Problems in data collection
        + Hardware or software issues
        + Human error
      * Outliers may significantly distort descriptive statistics or visualizations
    - Missing data – important aspect of data quality that always needs a detailed investigation to determine its origin and impact on following steps
      * Isolated instances of aren’t usually a problem, however if the missing data is not randomly distributed or occurs in large numbers globally or in specific variables the data set will be biased and thus not appropriate for a valid analysis
      * Can be an indicator of flaws with the data collection process
* Identify and describe challenges and techniques related to: data cleaning, data preparation, and data presentation.
  + - Data Preparation – Make data suitable for the follow-up phases
      * Data cleaning – identify and fix data quality issues
      * Data transformation – transform data to improve analysis or manipulation
      * Synthesis of data – create new attributes derived from existing data
      * Data integration – combine data from different sources
      * Data reduction or selection – eliminate data from the collection
* Describe the importance of data pipelines and how Makefiles can be used to implement them.

*Ref: PRI slides.*

## **IR tasks and systems**

Information retrieval vs data retrieval, modules in an IR system.

**Questions:**

* What is the difference between information retrieval and data retrieval?
* Give examples of IR and data retrieval systems.
* Give some examples of retrieval tasks evaluated in TREC.
* What are the modules of an IR system?

*Ref: BY, Chap. 1 (Intro).*

*Ref:*[*TREC tracks*](https://www.google.com/url?q=http://trec.nist.gov/tracks.html&sa=D&source=editors&ust=1673102615092405&usg=AOvVaw0CYngSgl6Xgptfr2Avrp1y)*.*

*Ref:*[*The Anatomy...*](https://www.google.com/url?q=http://infolab.stanford.edu/~backrub/google.html&sa=D&source=editors&ust=1673102615093141&usg=AOvVaw3ED0u0TPiDOKlfd1jx-BR_)*, Brin & Page.*

## **IR concepts**

Concepts: document, information need, relevance, bag of words, inverted index, postings list, term pre-processing.

**Questions:**

* What is… a document, a collection, a term, a bag of words?
* Define stemming.
* What is… an inverted index, a vocabulary, a postings list?
* What is… an information need, a query, a results list?
* What is a relevant result in a results list?

*Ref: Manning, Chap. 1 (Boolean Retrieval)*

## **Vector model**

Term weighting, tf, df, cf, idf, vector model, ranking in the vector model.

**Questions:**

* What is the bag of words model for a document?
* What is… term frequency, collection frequency, document frequency, inverse document frequency?
* How do you calculate tf-idf weights?
* How do you rank documents in the vector model?

Exercises: look at Exercises 6.8, 6.9, 6.10, 6.11, 6.15, 6.16, 6.17 and Examples 6.2, 6.3, 6.4.

*Ref: Manning, Chap. 2 (The term vocabulary and postings lists) (2.2) and Chap. 6 (Scoring, term weighting and the vector space model) (6.2, 6.3)*

## **Evaluation**

Precision, recall, P-R curves, MAP, reference collections, relevance judgements.

**Questions:**

* What is… precision, recall, interpolated precision?
* What is… precision at k, R-precision?
* Name the components of a test collection.
* Why is a set of relevance judgements considered a “ground truth” for IR?
* Draw a precision-recall curve for capturing the evolution of precision in the ranked list of results for a query.
* What is an average 11-point precision-recall graph for a set of queries?
* What is MAP, and do you calculate it for a set of queries in a test collection?

Exercises: look at Exercises 8.1, 8.4, 8.8, 8.9.

*Ref: Manning, Chap. 8.*

*Ref: TREC pages:*[*http://trec.nist.gov/*](https://www.google.com/url?q=http://trec.nist.gov/&sa=D&source=editors&ust=1673102615097384&usg=AOvVaw1_56nbimkoKIF3ligfvF-j)

## **Web search**

Web information needs, the bow-tie model, web search vs enterprise search, multimedia content, ranking functions and ranking signals

**Questions:**

* What are informational, transactional and navigational information needs?
* Name some differences between web search and enterprise search.
* How do you index images?
* Give examples of ranking signals used by search engines.
* What are the SCC, IN and OUT components in the view of the web as a bowtie?

Examples: look at Manning.

*Ref: Manning, Chap. 19.*

## **Link analysis**

Web ranking, anchor text, PageRank, hubs and authorities.

Questions:

* What are in-links and out-links for a web page?
* How is anchor text used in web search?
* Calculate PageRank values for a set of linked documents.
* Calculate Hub and Authority values for a set of linked documents.

*Ref: Manning, Chap. 21.*

## **Query processing**

Query processing techniques and relevance feedback strategies.

**Questions:**

* Describe and distinguish between the two query processing techniques — document-at-a-time and term-at-a-time.
* In what contexts is query transformation / expansion advantageous?
* What techniques can be used to apply transformations / expansions to user queries?
* Identify and describe query expansions techniques, such as relevance feedback or pseudo-relevance feedback.

*Ref: PRI slides.*

*Ref: Manning, Chap. 9.*

*Ref: Croft, Chap. 5.*

## **Entity-oriented search**

Entity-oriented search concepts, building entity representations, and data sources.

**Questions:**

* What is entity-oriented search? What is necessary to implement it?
* Describe the challenges and techniques associated with… building entity descriptions, entity ranking, entity linking.
* Describe the data sources typically required for entity oriented search and its characteristics.

*Ref: PRI slides.*

*Ref: Balog.*

## **Search user interfaces**

Search user interface design for search services, design principles and heuristics, and application examples.

**Questions:**

* Identify and describe user interface techniques and elements that can be used to improve user experience in using search systems.
* Describe how user interaction innovations and experiments can be evaluated.
* What are design principles and heuristics?

*Ref: PRI slides.*

## **Learning to Rank and Neural Information Retrieval**

Learning to rank definition, concepts, and the main approaches for modelling ranking as a machine learning problem.

Neural information retrieval definition, concepts, and application to the retrieval process.

**Questions:**

* What is Learning to Rank?
* Which are the main approaches in LTR? How do they differ in terms of input and output data?
* What is Neural Information Retrieval?
* How can neural models be used in the retrieval process?
* What are word embeddings?
* What is the difference between Learning to Rank and Neural Information Retrieval?

*Ref: PRI slides.*