Information Retrieval
CS 6900

Lecture 01

Razvan C. Bunescu

School of Electrical Engineering and Computer Science

bunescu@ohio.edu

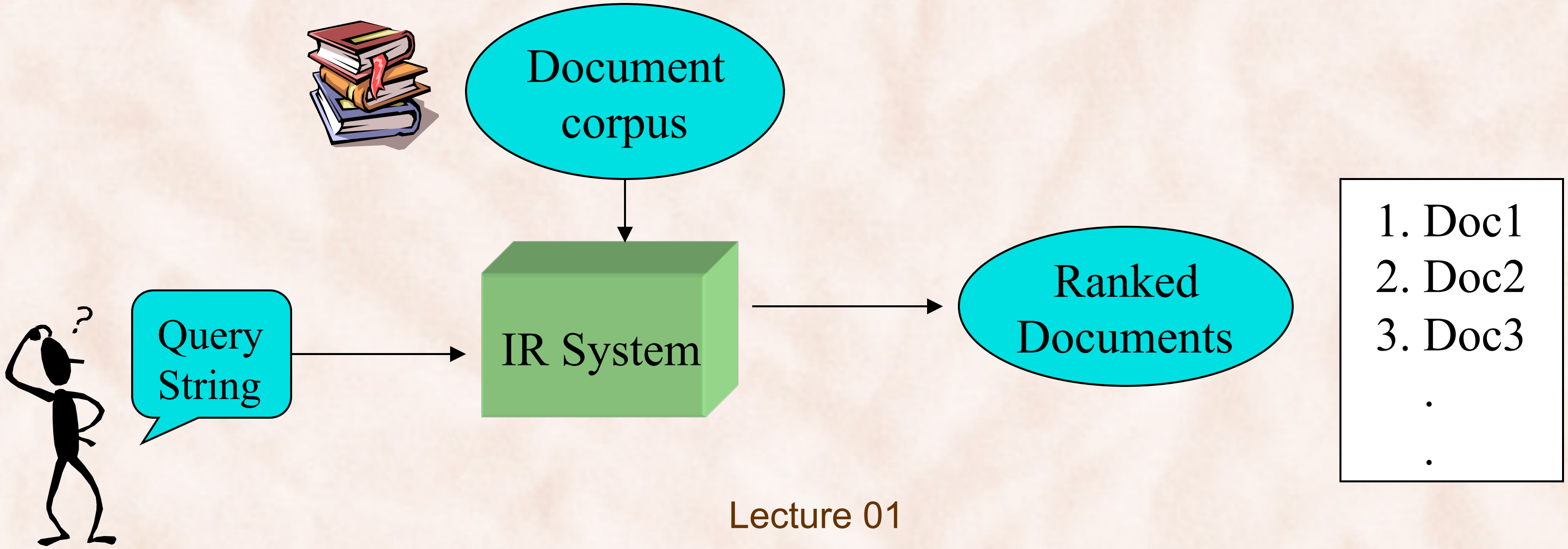
Information Retrieval

- Information Retrieval (IR) is finding material of an unstructured nature that satisfies an information need from within large collections.

- Examples of large collections and informations needs:
  1) Large corpus of literary texts:
     - Find Shakespeare plays that talk about the meaning of life.
  2) World Wide Web:
     - Find affordable hotels on the beach in Destin, Florida.
  3) My computer:
     - Find files that contain the words "information retrieval".

Typical IR task

- Input:
  - A large collection of unstructured text documents.
  - A user query expressed as text.
- Output:
  - A ranked list of documents that are relevant to the query



Document corpus

Query String

IR System

Ranked Documents

1. Doc1
2. Doc2
3. Doc3

.

.

IR on a Large Text Corpus

1. *“Find Shakespeare plays that talk about the meaning of life”:*
   - Information Need expressed as a string Query:
     - Boolean:
       - Naïve: meaning AND life
       - Better: (meaning OR signify) AND life
     - Phrase: “the meaning of life”
     - Proximity: meaning NEAR life
     - Keywords: meaning life
   - Material of an unstructured nature:
     - text documents (plays).

IR on the Web (Web Search)

- *"Find affordable hotels on the beach in Destin, Florida"*:
  - Information Need, typically expressed as a keyword query:
    - Keywords: 3 star hotel on the beach in Destin FL.
  - Material of an unstructured nature:
    - Text (unstructured)
    - HTML (semistructured).
      - Exploit the HTML structure.
      - Exploit the link structure of the Web (PageRank, HITS).

IR on My Computer (Personal IR)

- *“Find files that contain the words Information Retrieval”*:
  - Information Need, typically expressed as a keyword query:
    - Keywords: information retrieval
      - Interpreted as a conjunctive Boolean query in MS Vista Instant Search and Mac OS X Spotlight:
        - » Boolean: information AND retrieval
  - Material of an unstructured nature:
    - Need to handle a broad range of documents types:
      - Text, HTML, XML, PDF, ODT, DOCX, PPTX, …

Information Retrieval vs. Database Search

- Information Retrieval:
  - Finding information in unstructured repositories (text).
  - Queries: Boolean, keyword, phrase, proximity, …
    - 3 star hotel on the beach in Destin FL
- Database Search:
  - Finding information stored in structured repositories (relational databases, graph databases, etc.).
  - Queries: SQL, SPARQL, RPQ, Cypher, …
    - SELECT * FROM Book WHERE price > 100

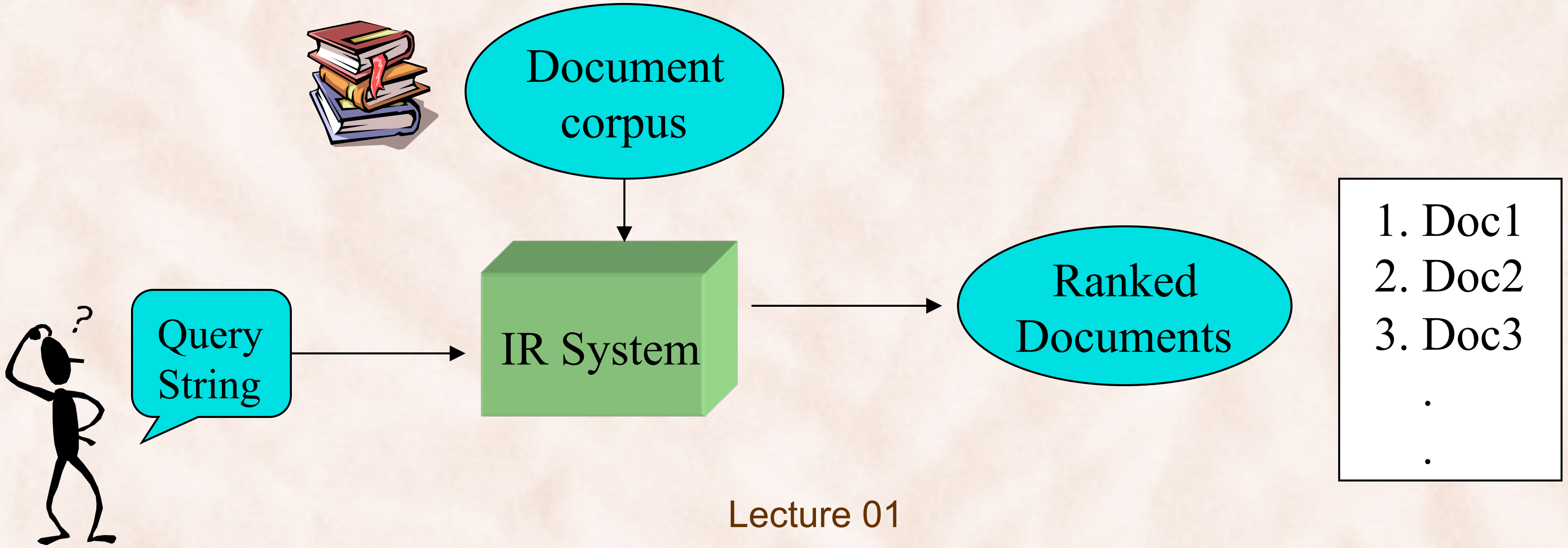
      ORDER BY title;

(Semi)Structured Information Retrieval

- (Semi)Structured IR: find information in text with markup:
  - Queries combine textual criteria with structural criteria:
    - Digital libraries: give me a full-length article on fast fourier transforms
    - Patent DBs: give me patents whose claims mention RSA public key encryption and that cite US patent 4,405,829.
    - Entity-tagged text: give me articles about sightseeing tours of the Vatican and the Coliseum.
  - Markup languages: HTML, XML, ODT (OpenOffice), …

Typical IR task

- Input:
  - A large collection of unstructured text documents.
  - A user query expressed as text.
- Output:
  - A ranked list of documents that are **relevant** to the query



Relevance

- Relevance is a subjective judgment and may include:
  - Being on the subject.
  - Being timely (recent information).
  - Being authoritative (from a trusted source).
  - Satisfying the user’s information need i.e. his/her goals and intended use of the information.
    - “Find Shakespeare plays that talk about the meaning of life”.
    - Typically expressed as a Query String:
      - meaning of life

From Queries to Relevant Documents

- **Phrase Queries**:
  - Simplest notion of relevance is that the query string appears verbatim in the document.
  - “meaning of life”

- **Keyword Queries**:
  - Slightly less strict notion is that the words in the query appear frequently in the document, in any order (bag-of-words).
  - meaning life

“Find Shakespeare plays that talk about the meaning of life”
Keyword Query: meaning life

Tomorrow, and tomorrow, and tomorrow,
Creeps in this petty pace from day to day,
To the last syllable of recorded time;
And all our yesterdays have lighted fools
The way to dusty death. Out, out, brief candle!
Life's but a walking shadow, a poor player
That struts and frets his hour upon the stage
And then is heard no more. It is a tale
Told by an idiot, full of sound and fury
Signifying nothing.

— Skakespeare’s Macbeth (Act 5, Scene 5, lines 17-28)

“Find Shakespeare plays that talk about the meaning of life”
Keyword Query: meaning life

Tomorrow, and tomorrow, and tomorrow,
Creeps in this petty pace from day to day,
To the last syllable of recorded time;
And all our yesterdays have lighted fools
The way to dusty death. Out, out, brief candle!
***Life**'s but a walking shadow, a poor player*
That struts and frets his ho
And then is heard no more. It is a tale
Told by an idiot, full of sound and fury
Signifying nothing.

— Skakespeare’s Macbeth (Act 5, Scene 5, lines 17-28)

=> need to bridge the **Lexical Gap**

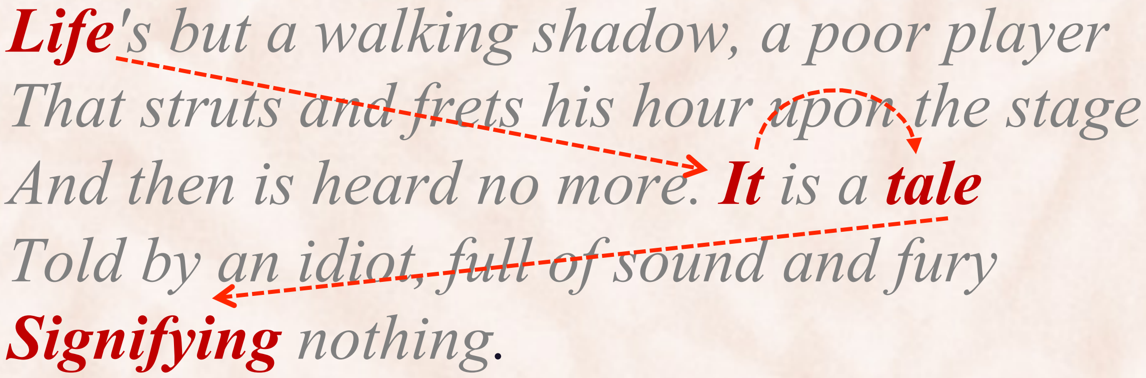
"Find Shakespeare plays that talk about the meaning of life"
Boolean Query: (meaning OR signify) AND life

Tomorrow, and tomorrow, and tomorrow,
Creeps in this petty pace from day to day,
To the last syllable of recorded time;
And all our yesterdays have lighted fools
The way to dusty death. Out, out, brief candle!
Life's *but a walking shadow, a poor player*
That struts and frets his hour upon the stage
And then is heard no more. It is a tale
Told by an idiot, full of sound and fury
Signifying *nothing.*

— Skakespeare's Macbeth (Act 5, Scene 5, lines 17-28)

From Information Retrieval (IR) to Question Answering (QA)

Tomorrow, and tomorrow, and tomorrow,
Creeps in this petty pace from day to day,
To the last syllable of recorded time;
And all our yesterdays have lighted fools
The way to dusty death. Out, out, brief candle!



Life*'s but a walking shadow, a poor player*
That struts and frets his hour upon the stage
And then is heard no more. ***It*** *is a* ***tale***
Told by an idiot, full of sound and fury
Signifying *nothing.*

— Skakespeare’s Macbeth (Act 5, Scene 5, lines 17-28)

From Information Retrieval (IR) to Question Answering (QA)

Tomorrow, and tomorrow, and tomorrow,
Creeps in this petty pace from day to day,
To the last syllable of rec
And all our yesterdays ha
The way to dusty death. Out, out, brief candle!
Life*'s but a walking shadow, a poor player*
That struts and frets his hour upon the stage
And then is heard no more. ***It*** is a ***tale***
Told by an idiot, full of sound and fury
Signifying ***nothing***.

— Skakespeare's Macbeth (Act 5, Scene 5, lines 17-28)

Q: What is the meaning of life?
A: Nothing!

Question Answering vs. Information Retrieval

- QA enables users to express information needs through questions in natural language.
  - Answer in QA is focused, typically a noun phrase for factual QA.
  - Answer in IR is a ranked list of relevant documents.
- QA needs deeper linguistic processing of the text ⇒ more difficult than classical keyword–based IR:
  - Coreference Resolution.
  - Syntactic/Dependency Parsing.
  - Word Sense Disambiguation.

Problems with Simple Keyword-based IR

- May not retrieve relevant documents that include synonymous terms.
  - meaning vs. signifying
  - FL v
- May re
  polyse
  - Python (baseball vs. mammal)
  - Apple (company vs. fruit)
  - play (theater play vs. act of playing)

In this course:

- We will cover the basics of keyword-based IR.
- Also address more complex techniques for “intelligent” IR.

Intelligent IR

- Take into account the *meaning* of the words used.
- Take into account the *order* of words in the query.
- Adapt to the user based on automatic or semi-automatic *feedback.*
- *Expand* search query with related terms.
- Perform automatic *spell checking* / *diacritics restoration.*
- Take into account the *authority* of the source.

Classic IR Models

- Each document represented by a set of representative keywords or **index terms**.
- An **index term** is a document word useful for remembering the document main themes.
- Index terms may be selected to be only nouns, since nouns have meaning by themselves:
  - Should reduce the size of the index.
  - ... But it requires the identification of nouns ⇒ Part of Speech tagger
- However, search engines assume that all words are index terms (full text representation).

Classic IR Models

- Not all terms are equally useful for representing the document contents:
  - less frequent terms allow identifying a narrower set of documents
- The importance of the index terms is represented by weights associated to them.
- Let:
  - k_i be an index term
  - d_j be a document
  - w_{ij} is a weight associated with (k_i,d_j)
- The weight w_{ij} quantifies the importance of the index term for describing the document contents.

IR System Components

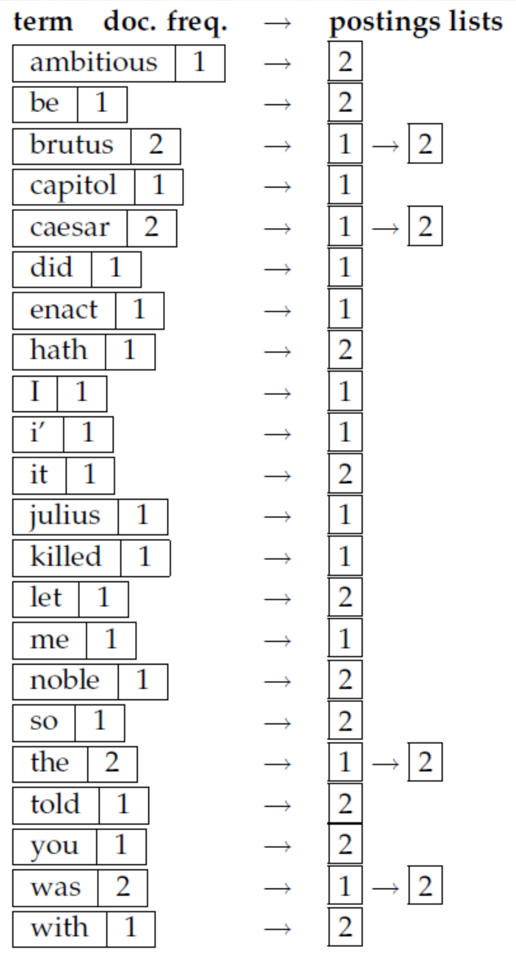
- **Text Operations** form index words (tokens)
  - Tokenization.
  - Stopword removal.
  - Stemming.
- **Indexing** constructs an *inverted index* of word to document pointers.
  - Mapping from tokens to document IDs.

Doc 1

I did enact Julius Caesar I was killed i' the Capitol; Brutus killed me.

Doc 2

So let it be with Caesar. The noble Brutus hath told you Caesar was ambitious



term	doc. freq.	→	postings lists
ambitious	1	→	2
be	1	→	2
brutus	2	→	1 → 2
capitol	1	→	1
caesar	2	→	1 → 2
did	1	→	1
enact	1	→	1
hath	1	→	2
I	1	→	1
i'	1	→	1
it	1	→	2
julius	1	→	1
killed	1	→	1
let	1	→	2
me	1	→	1
noble	1	→	2
so	1	→	2
the	2	→	1 → 2
told	1	→	2
you	1	→	2
was	2	→	1 → 2
with	1	→	2

IR System Components

- **Searching** retrieves documents that contain a given query token from the inverted index.
- **Ranking** scores all retrieved documents according to a relevance metric.
- **User Interface** manages interaction with the user:
  - Query input and document output.
  - Relevance feedback.
  - Visualization of results.
- **Query Operations** transform the query to improve retrieval:
  - Query expansion using a thesaurus.
  - Query transformation using relevance feedback.

Relevant Disciplines

- **Natural Language Processing**:
  - Tokenization & Stemming.
  - Part-Of-Speech (POS) tagging.
  - Syntactic Parsing, Word Sense Disambiguation, Information Extraction, …
- **Artificial Intelligence**:
  - Focused on the representation of knowledge, reasoning, and intelligent action.
  - Formalisms for representing knowledge and queries:
    - First-order Predicate Logic.
    - Bayesian Networks.

Relevant Disciplines

- **Machine Learning**:
  - Text Categorization:
    - Automatic hierarchical classification (Yahoo).
    - Adaptive filtering/routing/recommending.
    - Automated spam filtering.
  - Text Clustering:
    - Clustering of IR query results.
    - Automatic formation of hierarchies (Yahoo).
  - Learning to rank relevant documents.
  - Learning models for basically any relevant NLP task:
    - Tokenization, POS tagging, syntactic parsing, WSD, …

Relevant Disciplines

- **Linear Algebra**:
  - Vector Space Models.
  - Latent Semantic Indexing.
  - Link Analysis.

- **Probability and Statistics**:
  - Probabilistic IR.
  - Language Models for IR.
  - Link Analysis.

Course Topics (Tentative)

1. Classical IR models:
   - Boolean & Vector Space Models.
   - Text operations & Indexing
2. Probabilistic IR.
3. Language Models for IR.
4. Evaluation of IR performance.
5. Relevance feedback and query expansion.
6. Web Search:
   - Web crawling.
   - Link analysis (PageRank, Hubs and Authorities).

Course Topics (Tentative)

7. Text Classification and Clustering.
8. Personalized IR.
9. Question Answering.

- Tutorials: Python & NLTK.
- Background: Linear Algebra, Probability and Statistics.