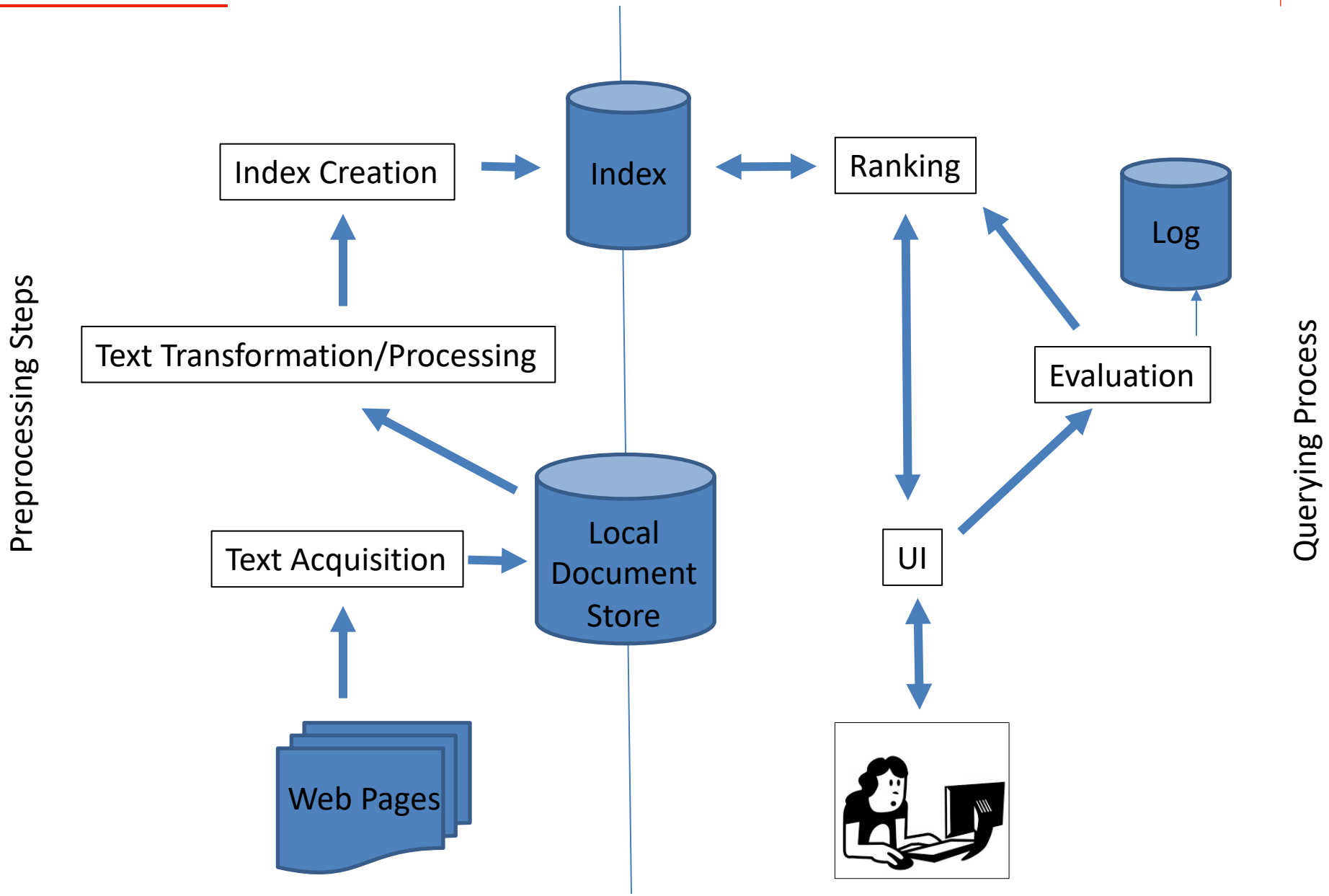
Informatics 225
Computer Science 221

Information Retrieval

Lecture 14

These course materials borrow, with permission, from those of Prof. Cristina Videira Lopes, Prof. Alberto Krone-Martins, Addison Wesley 2008, Chris Manning, Pandu Nayak, Hinrich Schütze, Heike Adel, Sascha Rothe, Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie. Powerpoint theme by Prof. André van der Hoek.

Architecture



Text processing & tokenization++

Information Retrieval

Phrases

- Phrases are important in search applications
- Many web search queries are 2-3 word phrases

Phrases

- Phrases are important in search applications
- Many web search queries are 2-3 word phrases
- Phrases are
  - More precise than single words
    - e.g., documents containing “black sea” vs. two words “black” and “sea”
    - Order is important (e.g. do you want pages with “the sea turned black”?)
  - Less ambiguous
    - e.g., “big apple” vs. “apple”

Phrases

- Phrases are important in search applications
- Many web search queries are 2-3 word phrases
- Phrases are
  - More precise than single words
    - e.g., documents containing “black sea” vs. two words “black” and “sea”
    - Order is important (e.g. do you want pages with “the sea turned black”?)
  - Less ambiguous
    - e.g., “big apple” vs. “apple”
- Can be difficult for ranking : details depend on our model
    - e.g., Given query “fishing supplies”, how do we score documents with
      - exact phrase many times, exact phrase just once, individual words in same sentence, same paragraph, whole document, variations on words?

Phrases

- Text processing issue – what is a phrase?
- Related issue – how are phrases recognized?

Phrases

- Text processing issue – what is a phrase?
- Related issue – how are phrases recognized?
- Three possible approaches:
  - Identify syntactic phrases using a *part-of-speech* (POS) tagger
  - Use word *n-grams*
  - Store word positions in indexes and use *proximity operators* in queries

Part-of-speech (POS) Tagging

- POS taggers use statistical / machine learning models of text to predict syntactic tags of words
  - Example tags:
    - NN (singular noun)
    - NNS (plural noun)
    - VB (verb)
    - VBD (verb, past tense)
    - VBN (verb, past participle)
    - IN (preposition)
    - JJ (adjective),
    - CC (conjunction, e.g., “and”, “or”)
    - PRP (pronoun)
    - MD (modal auxiliary, e.g., “can”, “will”).

- Phrases can then be defined as simple noun groups

Part-of-speech (POS) Tagging Example

Original text:

Document will describe marketing strategies carried out by U.S. companies for their agricultural chemicals, report predictions for market share of such chemicals, or report market statistics for agrochemicals, pesticide, herbicide, fungicide, insecticide, fertilizer, predicted sales, market share, stimulate demand, price cut, volume of sales.

Brill tagger:

Document/NN will/MD describe/VB marketing/NN strategies/NNS carried/VBD out/IN by/IN U.S./NNP companies/NNS for/IN their/PRP agricultural/JJ chemicals/NNS ,/, report/NN predictions/NNS for/IN market/NN share/NN of/IN such/JJ chemicals/NNS ,/, or/CC report/NN market/NN statistics/NNS for/IN agrochemicals/NNS ,/, pesticide/NN ,/, herbicide/NN ,/, fungicide/NN ,/, insecticide/NN ,/, fertilizer/NN ,/, predicted/VBN sales/NNS ,/, market/NN share/NN ,/, stimulate/VB demand/NN ,/, price/NN cut/NN ,/, volume/NN of/IN sales/NNS ./.

Example Noun Phrases

TREC data		Patent data	
Frequency	*Phrase*	*Frequency*	*Phrase*
65824	united states	975362	present invention
61327	article type	191625	u.s. pat
33864	los angeles	147352	preferred embodiment
18062	hong kong	95097	carbon atoms
17788	north korea	87903	group consisting
17308	new york	81809	room temperature
15513	san diego	78458	seq id
15009	orange county	75850	brief description
12869	prime minister	66407	prior art
12799	first time	59828	perspective view
12067	soviet union	58724	first embodiment
10811	russian federation	56715	reaction mixture
9912	united nations	54619	detailed description
8127	southern california	54117	ethyl acetate
7640	south korea	52195	example 1
7620	end recording	52003	block diagram
7524	european union	46299	second embodiment
7436	south africa	41694	accompanying drawings
7362	san francisco	40554	output signal
7086	news conference	37911	first end
6792	city council	35827	second end
6348	middle east	34881	appended claims
6157	peace process	33947	distal end
5955	human rights	32338	cross-sectional view
5837	white house	30193	outer surface

Word N-Grams

- Part-of-speech (POS) tagging too slow for large collections

Word N-Grams

- Part-of-speech (POS) tagging too slow for large collections
- Simpler definition – phrase is any sequence of *n* words – known as *n-grams (or shingles)*
  - *bigram*: 2 word sequence, *trigram*: 3 word sequence, etc.
  - N-grams also used at the character level for applications such as OCR
- N-grams typically formed from *overlapping* word sequences
  - i.e. move n-word "window" one word at a time in document

N-Grams

- Frequent n-grams are more likely to be meaningful phrases
- N-grams form a Zipf distribution
  - *Better fit than the words alone!*
- Could index all n-grams up to specified length
  - Advantage: Much faster than POS tagging

N-Grams

- Frequent n-grams are more likely to be meaningful phrases
- N-grams form a Zipf distribution
  - *Better fit than the words alone!*
- Could index all n-grams up to specified length
  - Advantage: Much faster than POS tagging
  - Drawback : Uses a lot of storage
    - e.g., document containing 1,000 words would contain 3,990 instances of word n-grams of length $2 \leq n \leq 5$

Google N-Grams

- **Even using a lot of storage, many web search engines do index n-grams**

- Google sample:

Number of tokens:	1,024,908,267,229
Number of sentences:	95,119,665,584
Number of unigrams:	13,588,391
Number of bigrams:	314,843,401
Number of trigrams:	977,069,902
Number of fourgrams:	1,313,818,354
Number of fivegrams:	1,176,470,663

Google N-Grams

- Even using a lot of storage, many web search engines do index n-grams

- Google sample:

Number of tokens:	1,024,908,267,229
Number of sentences:	95,119,665,584
Number of unigrams:	13,588,391
Number of bigrams:	314,843,401
Number of trigrams:	977,069,902
Number of fourgrams:	1,313,818,354
Number of fivegrams:	1,176,470,663

- Most frequent trigram in English (on the web): “all rights reserved”
  - In Chinese, “limited liability corporation”

Phrases

- Text processing issue – what is a phrase?
- Related issue – how are phrases recognized?
- Three possible approaches:
  - Identify syntactic phrases using a *part-of-speech* (POS) tagger
  - Use word *n-grams*
  - Store word positions in indexes and use *proximity operators* in queries

Document Structure and Markup and Links

Information Retrieval

Document Structure and Markup

- Some parts of documents are more important than others

Document Structure and Markup

- Some parts of documents are more important than others
- Document parser recognizes structure using markup, such as HTML tags
  - Headers, anchor text, bolded text all likely to contain important text!
  - Metadata can also be important
  - Links are used for *link analysis* (more about this later in the course)

Example Web Page

Tropical fish

From Wikipedia, the free encyclopedia

Tropical fish include fish found in tropical environments around the world, including both freshwater and salt water species. Fishkeepers often use the term *tropical fish* to refer only those requiring fresh water, with saltwater tropical fish referred to as *marine fish*.

Tropical fish are popular aquarium fish , due to their often bright coloration. In freshwater fish, this coloration typically derives from iridescence, while salt water fish are generally pigmented.

Example Web Page

```
<html>
<head>
<meta name="keywords" content="Tropical fish, Airstone, Albinism, Algae eater,
Aquarium, Aquarium fish feeder, Aquarium furniture, Aquascaping, Bath treatment
(fishkeeping),Berlin Method, Biotope" />
…
<title>Tropical fish - Wikipedia, the free encyclopedia</title>
</head>
<body>
 …
<h1 class="firstHeading">Tropical fish</h1>
…
<p><b>Tropical fish</b> include <a href="/wiki/Fish" title="Fish">fish</a> found in <a
href="/wiki/Tropics" title="Tropics">tropical</a> environments around the world,
including both <a href="/wiki/Fresh_water" title="Fresh water">freshwater</a> and <a
href="/wiki/Sea_water" title="Sea water">salt water</a> species. <a
href="/wiki/Fishkeeping" title="Fishkeeping">Fishkeepers</a> often use the term
<i>tropical fish</i> to refer only those requiring fresh water, with saltwater tropical fish
referred to as <i><a href="/wiki/List_of_marine_aquarium_fish_species" title="List of
marine aquarium fish species">marine fish</a></i>.</p>
<p>Tropical fish are popular <a href="/wiki/Aquarium" title="Aquarium">aquarium</a>
fish , due to their often bright coloration. In freshwater fish, this coloration typically
derives from <a href="/wiki/Iridescence" title="Iridescence">iridescence</a>, while salt
water fish are generally <a href="/wiki/Pigment" title="Pigment">pigmented</a>.</p>
…
</body></html>
```

Example Web Page

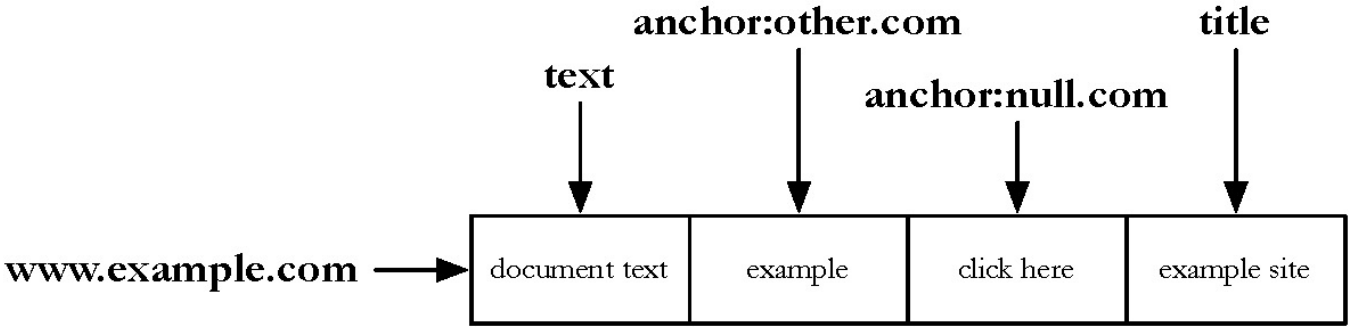
```
<html>
<head>
<meta name="keywords" content="Tropical fish, Airstone, Albinism, Algae eater,
Aquarium, Aquarium fish feeder, Aquarium furniture, Aquascaping, Bath treatment
(fishkeeping),Berlin Method, Biotope" />
…
<title>Tropical fish - Wikipedia, the free encyclopedia</title>
</head>
<body>
 …
<h1 class="firstHeading">Tropical fish</h1>
…
<p><b>Tropical fish</b> include <a href="/wiki/Fish" title="Fish">fish</a> found in <a
href="/wiki/Tropics" title="Tropics">tropical</a> environments around the world,
including both <a href="/wiki/Fresh_water" title="Fresh water">freshwater</a> and <a
href="/wiki/Sea_water" title="Sea water">salt water</a> species. <a
href="/wiki/Fishkeeping" title="Fishkeeping">Fishkeepers</a> often use the term
<i>tropical fish</i> to refer only those requiring fresh water, with saltwater tropical fish
referred to as <i><a href="/wiki/List_of_marine_aquarium_fish_species" title="List of
marine aquarium fish species">marine fish</a></i>.</p>
<p>Tropical fish are popular <a href="/wiki/Aquarium" title="Aquarium">aquarium</a>
fish , due to their often bright coloration. In freshwater fish, this coloration typically
derives from <a href="/wiki/Iridescence" title="Iridescence">iridescence</a>, while salt
water fish are generally <a href="/wiki/Pigment" title="Pigment">pigmented</a></p>
…
</body></html>
```

Links

- Links are a key component of the Web
- Important for navigation, but also for search
  - e.g., Example website
  - "*Example website*" is the anchor text
  - "*http://example.com*" is the destination link
  - both are used by search engines

Links : Anchor Text

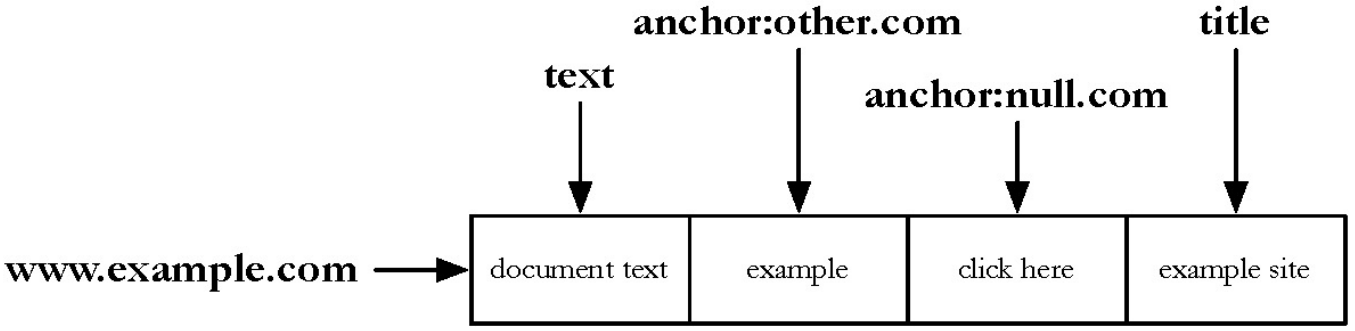
- Used as a description of the content of the *destination page*
  - i.e., collection of anchor text in all links pointing to a page used as an additional text field



Links : Anchor Text

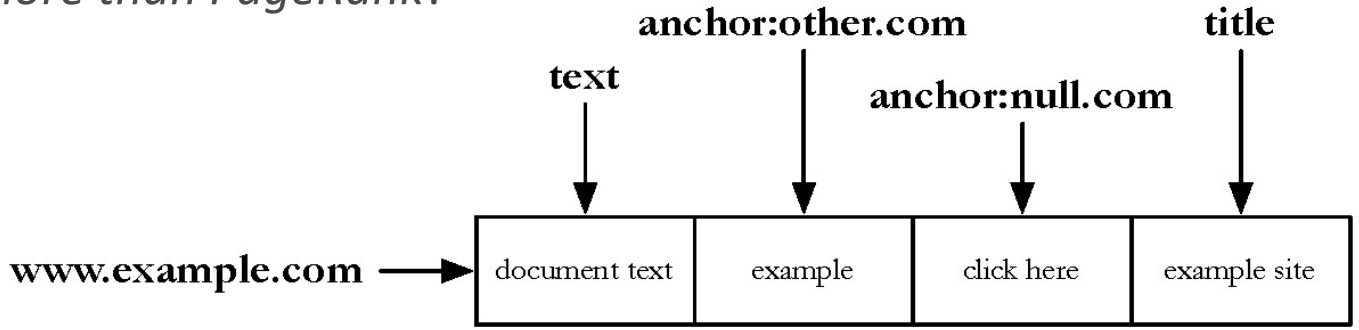
- Used as a description of the content of the *destination page*
  - i.e., collection of anchor text in all links pointing to a page used as an additional text field

- Anchor text tends to be short, descriptive, and similar to query text

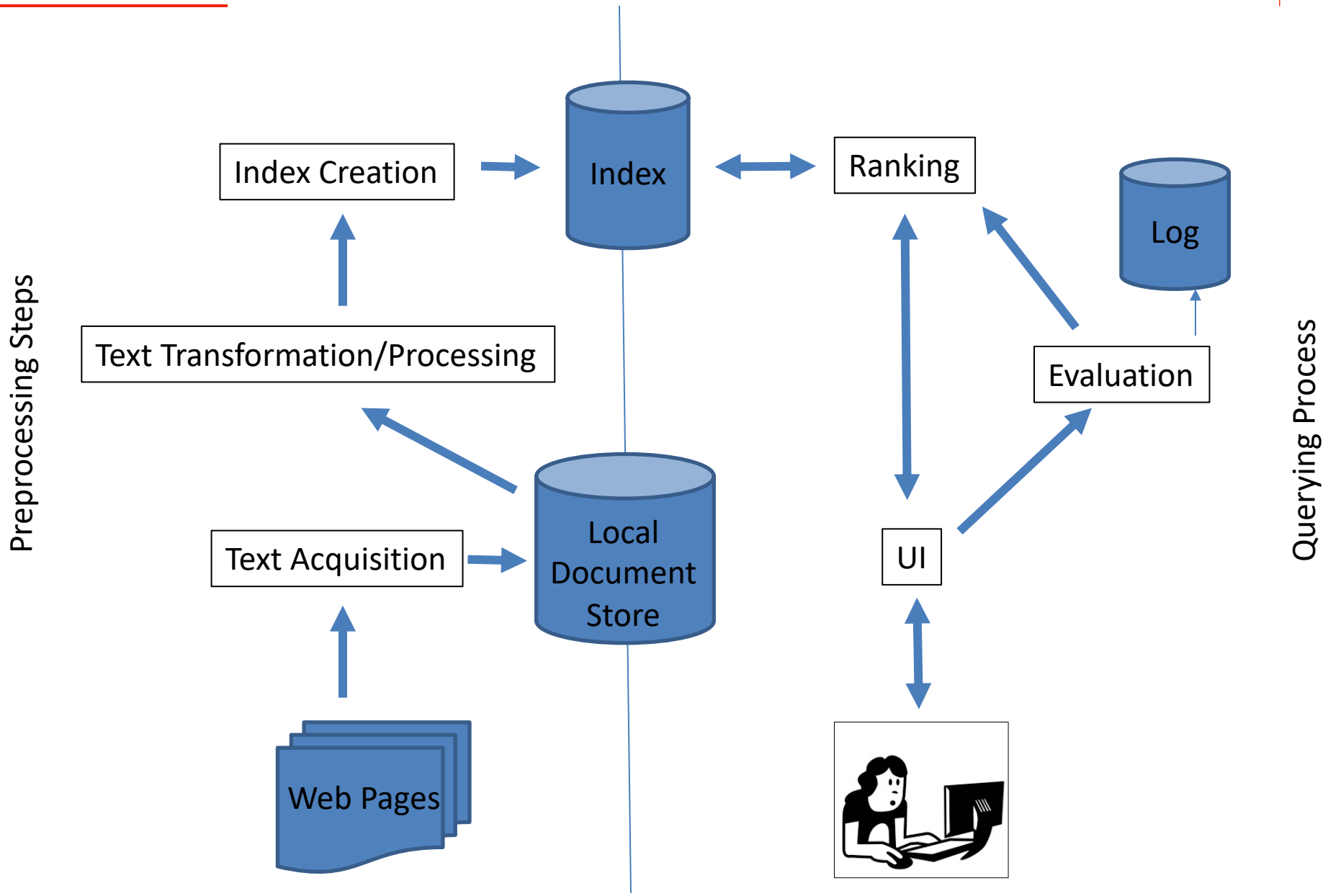


Links : Anchor Text

- Used as a description of the content of the *destination page*
  - i.e., collection of anchor text in all links pointing to a page used as an additional text field
- Anchor text tends to be short, descriptive, and similar to query text
- Retrieval experiments have shown that anchor text has significant impact on effectiveness for *some types of queries*
  - i.e., *more than PageRank*!



Architecture



Term-document incidence matrix

Information Retrieval

Unstructured data in 1620

- Which plays of Shakespeare contain the words *Brutus AND Caesar* but *NOT Calpurnia*?

Unstructured data in 1620

- Which plays of Shakespeare contain the words *Brutus AND Caesar* but *NOT Calpurnia*?
- Could you grep all of Shakespeare's plays for *Brutus* and *Caesar,* then strip out lines containing *Calpurnia*?
- Why is that not the answer for search?
  - Slow for large corpora
  - *NOT Calpurnia* is non-trivial
  - Other operations (e.g., find the word *Romans* near *countrymen*) is not feasible
  - Ranked retrieval is not possible
    - Users want ordered documents

Term-document incidence matrices

- Term-document incidence matrices : Boolean matrix indicating if a term (rows) exists in a certain document (columns).

Term-document incidence matrices

- Term-document incidence matrices : Boolean matrix indicating if a term (rows) exists in a certain document (columns).

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

1 if play contains word, 0 otherwise

Term-document incidence matrices

- Term-document incidence matrices : Boolean matrix indicating if a term (rows) exists in a certain document (columns).

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

1 if play contains word, 0 otherwise

How to answer the query:

Brutus** AND **Caesar** BUT NOT **Calpurnia

Incidence vectors

- Each term has a 0/1 vector (rows of the matrix)

- To answer query: take the rows for ***Brutus, Caesar*** and ***Calpurnia*** (complemented) ➔ bitwise *AND*.

$$\mathcal{R} = \mathcal{I}_{Bru} \wedge_{bitwise} \mathcal{I}_{Cae} \wedge_{bitwise} (\neg \mathcal{I}_{Cal})$$

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

Incidence vectors

- Each term has a 0/1 vector (rows of the matrix)

- To answer query: take the rows for ***Brutus, Caesar*** and ***Calpurnia*** (complemented) ➔ bitwise *AND*.

  - (Brutus I_{Bru}) 1 1 0 1 0 0 *AND*
  - (Caesar I_{Cae}) 1 1 0 1 1 1 *AND*
  - (NOT Calpurnia I_{Cal}) 1 0 1 1 1 1 =
  - **1 0 0 1 0 0**

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

Incidence vectors

- Each term has a 0/1 vector (rows of the matrix)

- To answer query: take the rows for ***Brutus, Caesar*** and ***Calpurnia*** (complemented) ➔ bitwise *AND*.

  - (Brutus I_{Bru}) 1 1 0 1 0 0 *AND*
  - (Caesar I_{Cae}) 1 1 0 1 1 1 *AND*
  - (NOT Calpurnia I_{Cal}) 1 0 1 1 1 1 =
  - **1 0 0 1 0 0**

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

Answers to query

- **Antony and Cleopatra, Act III, Scene ii**

  Agrippa [Aside to DOMITIUS ENOBARBUS]:

  Why, Enobarbus,
  When Antony found Julius ***Caesar*** dead,
  He cried almost to roaring; and he wept
  When at Philippi he found ***Brutus*** slain.

- **Hamlet, Act III, Scene ii**

  Lord Polonius:

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

Can you use this in bigger collections?

- Consider $N = 10^6$ documents, with ~10^3 words/document.
- Avg ~6 bytes/word, including spaces/punctuation
  - This results in ~6GB of data.
- Suppose there are M = 500K distinct terms among these.

Can you use this in bigger collections?

- Consider $N = 10^6$ documents, with ~10^3 words/document.
- Avg ~6 bytes/word, including spaces/punctuation
  - This results in ~6GB of data.
- Suppose there are M = 500K distinct terms among these.
- What is the size of the Term-document matrix?

The matrix cannot be built in memory

- 500K x 1M matrix has half-a-trillion 0's and 1's.
- **But it has no more than one billion 1's!**
  - **The matrix is extremely sparse !!!**

The matrix cannot be built in memory

- 500K x 1M matrix has half-a-trillion 0's and 1's.
- But it has no more than one billion 1's!
  - The matrix is extremely sparse.
- What's a better representation?
  - **We only record the 1 positions.**