Information Retrieval

BITS Pilani
Pilani Campus

Abhishek
January 2020

CS F469, Information Retrieval
Lecture No. 3

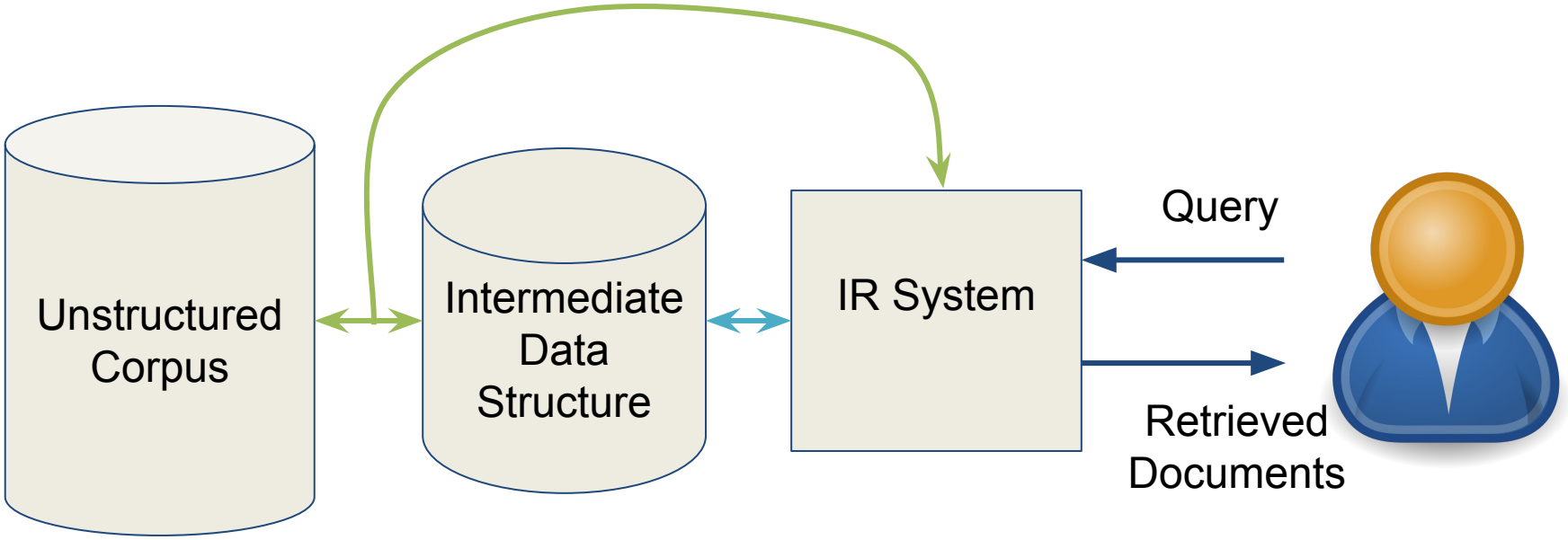
Recap of Lecture 2

- A simple IR task
- Boolean Retrieval models
- Indexing

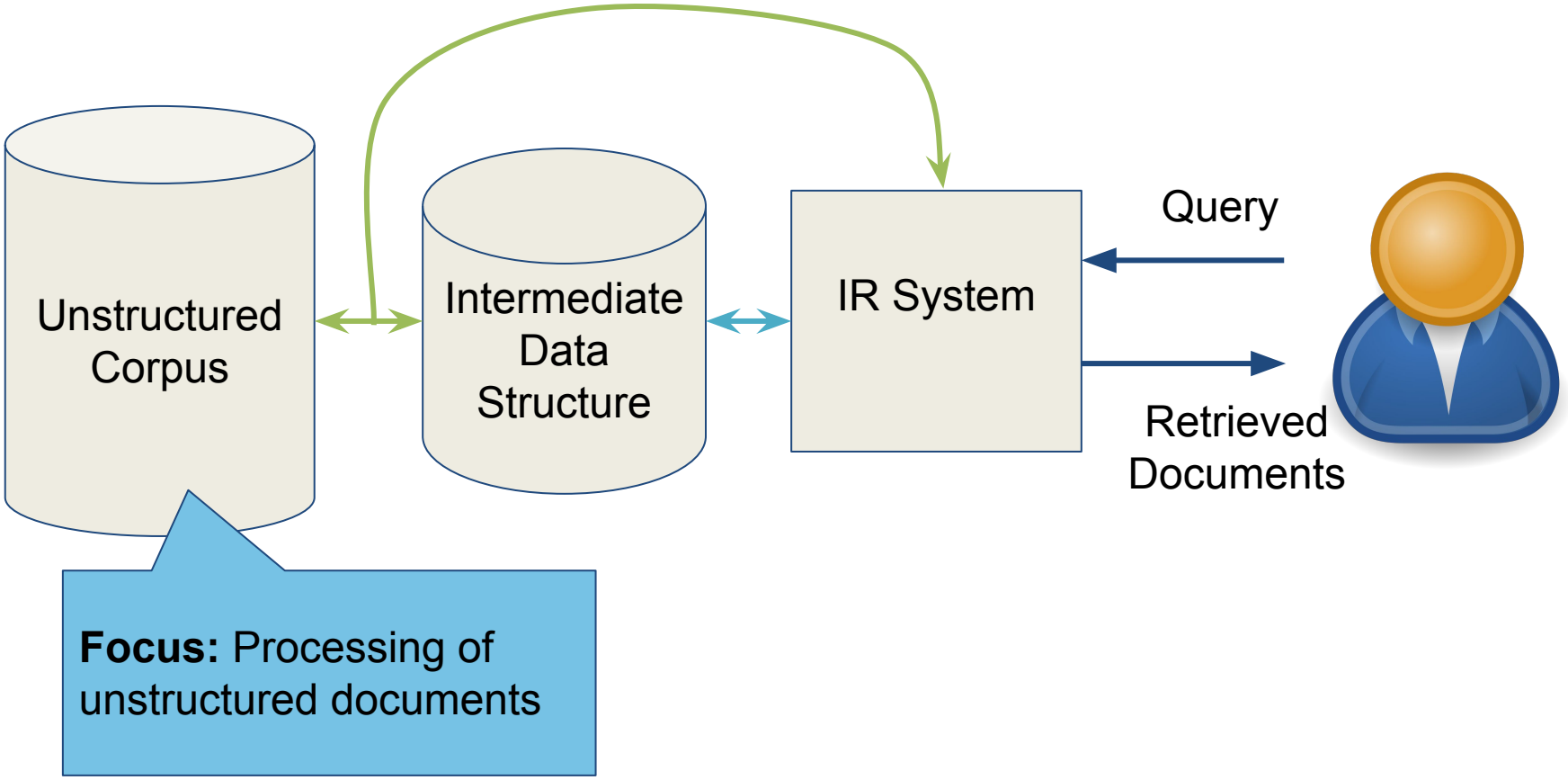
Today's Lecture

- Character encodings
- Document unit
- Tokenization
- Normalization
- Stemming and Lemmatization

Overview of IR system



Overview of IR system



Document file-format decoding

- Documents can be stored in several different file format such as .docx, .pdf, .xml.
- These file-format stores additional data such as color and sections.
- These documents can further also be in compressed file-formats such as .pdf.zip, .gz

Text needs to be extracted from these file-format as per the format specification.

Character encodings

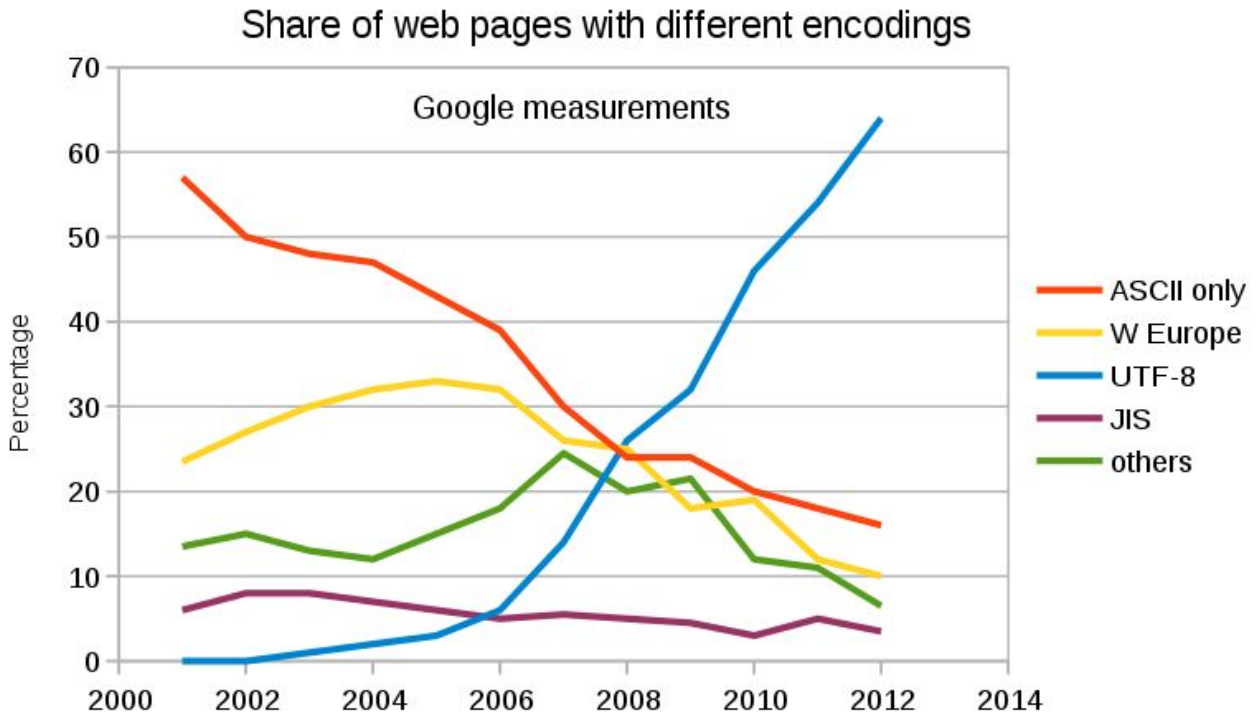
ASCII encoding

- Uses 7 bits: 0-127
- Example:

Decimal	Character	Description
65	A	uppercase A
90	Z	uppercase Z
94	^	caret
97	a	lowercase a
127	DEL	delete

List of popular encodings

- UTF-8 : Multi bytes (1 to 4 bytes)
- ISO-8859 : 1 byte
- Extended ASCII : 1 byte
- JIS : 2 bytes



UTF-8 Encoding

- Capable of encoding 1,112,064 valid character unicodes, including several languages characters and emojis.
  - 😃 = U+1F603 (\xF0\x9F\x98\x83)
  - 🚲 =U+1F6B2 (\xF0\x9F\x9A\xB2)
- Compatible with standard ASCII code (first 128 characters).

How to find which text encoding is used?

- User decided
- Present in metadata of the file.
- Heuristic

How to find which text encoding is used?

- User decided
- Present in metadata of the file.
- Heuristic

```
'ascii' codec can't decode byte 0xe3 in position 6: ordinal
not in range(128)
```

Document Unit

- Till now we assumed that the documents are fixed units for the purpose of indexing. For example, files present in a folder/directory.
- There are several cases where we might want to do something different.
    - Considering a single file as multiple documents. (Eg. mbox format for storing emails.)
    - Considering multiple files as a single document. (Eg. multiple .tex files can generate a single document.)

Indexing Granularity

- The document length should neither be very long nor very short.
- If document length is long, it can lead to retrieval of non-relevant documents. (recall high)
- If document length is very short, it can lead to non-retrieval of relevant documents. (precision low)

Determining Vocabulary of terms

Tokenization

Token: A token is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit of processing.

Example:

Input: Friends, Romans, Countrymen, lend me your ears;

Output:

Friends	Romans	Countrymen	lend	me	your	ears

Token vs Type vs Term

Eg. Input: “to sleep perchance to dream”

Tokens:	to	sleep	perchance	to	dream

Type: A type is a class of all tokens containing the same character sequence. There are only four types in the above example.

Term: A term is a (perhaps normalized) type that is included in the IR system’s dictionary. If to is omitted from index, then there are three terms.

Tokenization issues

- What sequence of characters constitutes as a useful semantic unit?
- Example sentence: O'Neill thinks that the boys' stories about Chile's capital aren't amusing.
  - For O'Neill, which of the following will be desired tokenization?
    - neill
    - oneill
    - o'neill
    - o' neill
    - o neill

If no preprocessing of the query is done, the query can only match with the 3rd token here.

Tokenization issues

- What sequence of characters constitutes as a useful semantic unit?
- Example sentence: O'Neill thinks that the boys' stories about Chile's capital aren't amusing.
  - For aren't, which of the following will be desired tokenization?
    - aren't
    - arent
    - are n't
    - aren t

Tokenization issues

- Emails? (abhishek@pilani.bits-pilani.ac.in)
- IP addresses: 172.16.16.93
- URLS: https://en.wikipedia.org/wiki/UTF-8
- Currency: $10, $ 10,000
- Dates: 12/01/2020
- **Hyphens:** co-education, New Delhi-Mumbai

Stop words

Stop words: Extremely common words that would appear to be of little value.

- Can be easily found by sorting the terms based on their frequencies.
- Eg. for english language:

  a, an, and, are, as, at, be, by, for, from, has, he, in, is, it, its, of, on, that, the, to, was, were, will, with

Stop words

- In keyword searches with terms like the and by don't seems very useful.
- However, stop words removal in phrasal searches can impact system's performance.
-

Eg: "President of the United States", "flights to London"

Several songs, movies names include words that are only stop words.

Eg. Let it be, To be or not to be

Normalization

- In several cases, the tokens in documents and query can be different but their meaning might be same.

- Eg. If query is USA, we might hope to also match document containing U.S.A. (another example: M.Tech., MTech)

- Token normalization is the process of canonicalizing so that matches occurs despite superficial differences in the character sequences of the tokens.

Normalization approaches

- Equivalence classes:

  Example: anti-discriminatory and antidiscriminatory are equivalent and both are mapped into the term antidiscriminatory.

Advantage: Some matching rules can be easily automatically defined. Eg. removal of hyphens or full stops in words (U.S.A.)

Disadvantage: Its symmetric and sometime asymmetric is better. On some cases, it can completely change the meaning, eg., C.A.T., New Delhi-Mumbai.

Normalization approaches

- relations between unnormalized tokens

  Eg. Hand constructed lists of synonyms such as car and automobile.

  - Usual way: Index unnormalized tokens and maintain a query expansion list.
    - Eg. query: car
    - Expanded query: car OR automobile
  - Other way: Perform token expansion during index construction.

Usefulness of asymmetric expansion of query

Query Term	Terms in documents that should be matched
Windows	Windows
windows	Windows, windows, window
window	window, windows

Commonly used normalization

- **Accents and Diacritics:** In English, diacritics on characters have a fairly marginal status. Eg. naive vs naïve, cliché vs cliche.
  - In spanish meaning of the word can change based on the diacritics. Eg. peña (a cliff) vs pena (sorrow)
  - Also, the important question is how the user enters the queries? How easy it is for the user to type non-ascii characters?

Commonly used normalization

- **Case-folding:**
  - CASE -> case
  - Information Retrieval -> information retrieval
  - CAT -> cat

- common strategies:
  - All words appearing in the beginning of a sentence.
  - All words appearing in the title.
  - Mid-sentence capitalized words are left as capitalized.

Other issued in normalization

- **British spellings vs American spellings:** colour vs color
- **Dates:** 13.01.2020, 13/01/2020, 13th Jan 2020, 13 Jan 2020, Jan 13, 2020
- **Currencies:** INR10, INR 10, INR 10,000, INR 10000, ₹ 1000, 1000 rupees.

Stemming and Lemmatization

- For gramatical reasons, documents are going to use different forms of a word. Eg. organize, organizes, organizing.

  Eg. car, cars, car's, cars' $\longrightarrow$ car

  am, are, is $\longrightarrow$ be

- Stemming is a crude heuristic process that chops off the ends of the word.
- Lemmatization uses language vocabulary and morphological analysis of words in order to retrieve the base form of the word.

Reference

https://nlp.stanford.edu/IR-book/

Chapter 2

Thank You!