Monsoon 2021

Basics of IR Systems

- Fundamental Concepts

Dr. Rajendra Prasath

Indian Institute of Information Technology Sri City, Chittoor

17 August 2021 (rajendra.2power3.com)

> Topics to be covered

- Recap:
  - IR systems
  - Classical Search Engines

- Keywords / User Information Needs
- Relevance / Irrelevance
- Personalization
- Words / Term Weighting

- Text Collection / Corpora
- Evaluation Strategy

- More topics to come up … Stay tuned …!!

Overview

Recap: Information Retrieval

- **Information Retrieval (IR)** is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

- These days we frequently think first of web search, but there are many other cases:
    - E-mail search
    - Searching your laptop
    - Corporate knowledge bases
    - Legal information retrieval
    - and so on . . .

Words and their occurrences

- Consider any city:
  - For e.g, search for **Darjeeling** (English Version)
  - Check Wikipedia Articles
  - https://en.wikipedia.org/wiki/Darjeeling
  - How many times Darjeeling comes up in the document?
    - **193 times**
  - Does it mean the following?
    - more it occurs … more it is important?

Look at Text Segment

- **Darjeeling** is a city and a municipality in the Indian state of West Bengal. It is located in the Lesser Himalayas at an elevation of 6,700 ft. It is noted for its tea industry, its views of Kangchenjunga, the world's third highest mountain, and the **Darjeeling** Himalayan Railway, a UNESCO World Heritage Site. **Darjeeling** is the headquarters of the **Darjeeling** District which has a partially autonomous status within the state of West Bengal. It is also a popular tourist destination in India.

Statistics: Text Data

- How do we extract the term statistics?

Darjeeling is a city and a municipality

in the Indian state of West Bengal.

➔ Darjeeling – 1; is – 1; a – 1; city – 1; and – 1; a – 1; municipality – 1; in - 1; the – 1; Indian – 1; state – 1; of – 1; west – 1; Bengal – 1;

➔➔ Darjeeling – 1; is – 1; **a – 2**; city – 1; and – 1; municipality – 1; in - 1; the – 1; Indian – 1; state – 1; of – 1; west – 1; Bengal – 1;

Look at 3 documents

d_1- **Darjeeling** is a city and a municipality in the Indian state of West Bengal. It is located in the Lesser Himalayas at an elevation of 6,700 feet

d_2- **Darjeeling** is noted for its tea industry, its views of Kangchenjunga, the world's third-highest mountain, and the **Darjeeling** Himalayan Railway, a UNESCO World Heritage Site

d_3- **Darjeeling** is the headquarters of the **Darjeeling** District which has a partially autonomous status within the state of West Bengal. It is also a tourist destination in India

Unique words and Counts?

d_1	d_2	d_3
2 the	2 the	3 the
2 of	2 its	2 of
2 is	2 Darjeeling	2 is
2 in	1 world's	2 a
2 a	1 views	2 Darjeeling
1 state	1 third-highest	1 within
1 municipality	1 tea	1 which
1 located	1 of	1 tourist
1 feet	1 noted	1 status
1 elevation	1 mountain	1 state
1 city	1 is	1 partially
1 at	1 industry,	1 in
1 and	1 for	1 headquarters
1 an	1 and	1 has
1 West	1 a	1 destination
1 Lesser	1 World	1 autonomous
1 It	1 UNESCO	1 also
1 Indian	1 Site	1 West
1 Himalayas	1 Railway	1 It
1 Darjeeling	1 Kangchenjunga	1 India
1 Bengal	1 Himalayan	1 District
1 6,700	1 Heritag	1 Bengal

Documents – Words / Terms*

✧ How to construct Terms - documents

Doc ID	Terms	# Words
d_1	6,700 (1), Bengal. (1), Darjeeling (1), Himalayas (1), Indian (1), It (1), Lesser (1), West (1), a (2), an (1), and (1), at (1), city (1), elevation (1), feet (1), in (2), is (2), located (1), municipality (1), of (2), state (1), the (2),	22
d_2	Darjeeling (2), Heritage (1), Himalayan (1), Kangchenjunga, (1), Railway, (1), Site (1), UNESCO (1), World (1), a (1), and (1), for (1), industry, (1), is (1), its (2), mountain, (1), noted (1), of (1), tea (1), the (2), third-highest (1), views (1), world's (1),	22
d_3	Bengal. (1), Darjeeling (2), District (1), India (1), It (1), West (1), a (2), also (1), autonomous (1), destination (1), has (1), headquarters (1), in (1), is (2), of (2), partially (1), state (1), status (1), the (3), tourist (1), which (1), within (1),	22

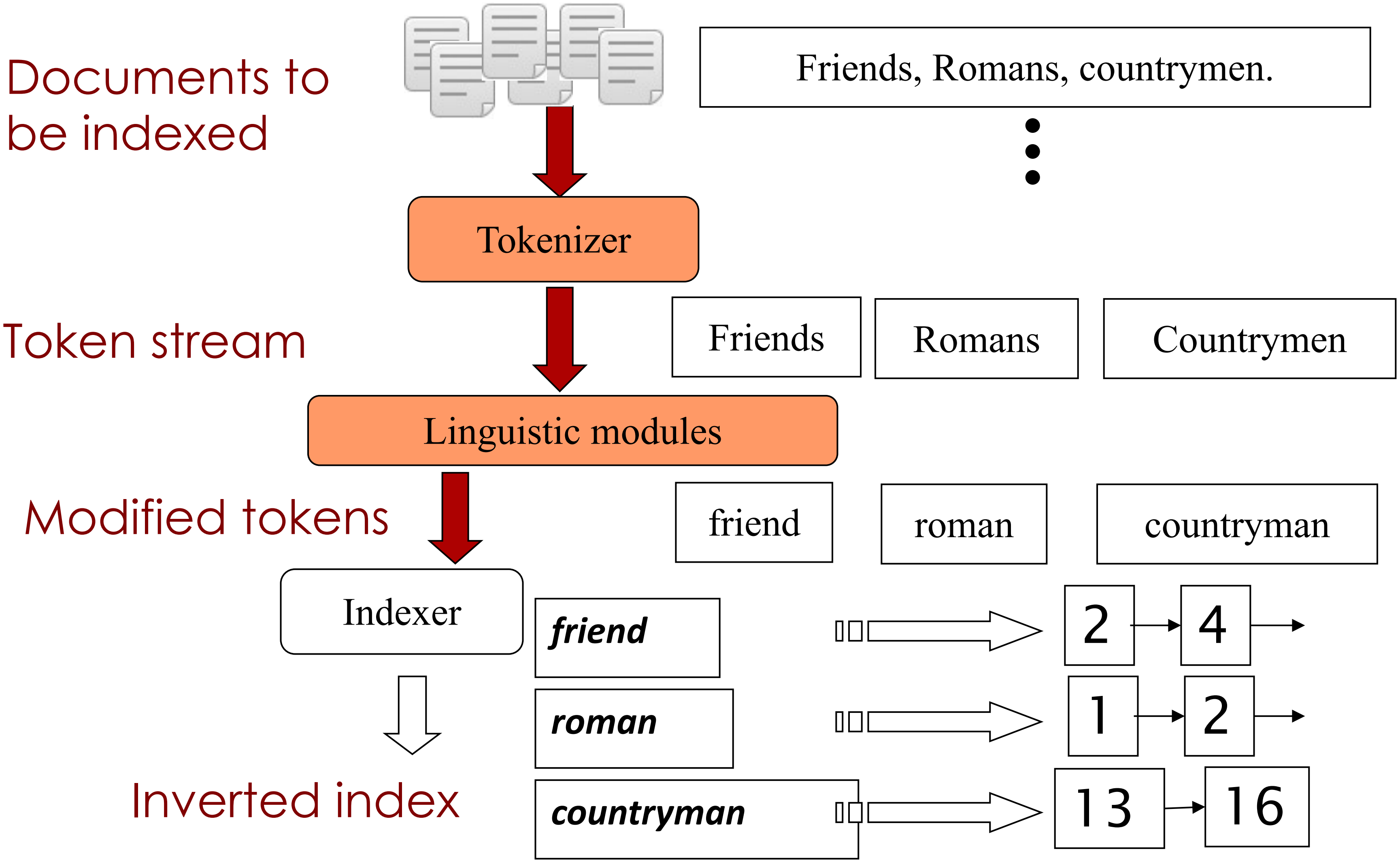
NOTE: “**Words**” and “**Terms**” are interchangeably used throughout the course

Terms - Documents

Terms	d_1	d_2	d_3	. . .	d_n
the	2	2	3	. . .	0
a	2	1	2	. . .	1
Darjeeling	1	2	2	. . .	0
is	2	1	2	. . .	0
of	2	1	2	. . .	0
in	2	0	0	. . .	1
and	1	1	0	. . .	0
Bengal	1	0	1	. . .	0
It	1	0	1	. . .	0
Its	0	2	0	. . .	2
state	1	0	1	. . .	0
West	1	0	1	. . .	1

NOTE: "**Words**" and "**Terms**" are interchangeably used throughout the course

Inverted index construction



Documents to be indexed

Friends, Romans, countrymen.

Tokenizer

Token stream

Friends

Romans

Countrymen

Linguistic modules

Modified tokens

friend

roman

countryman

Indexer

Inverted index

friend → 2 → 4

roman → 1 → 2

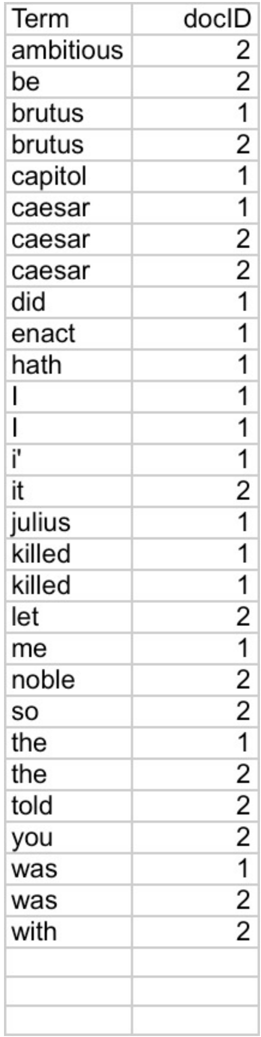
countryman → 13 → 16

Initial stages of text processing

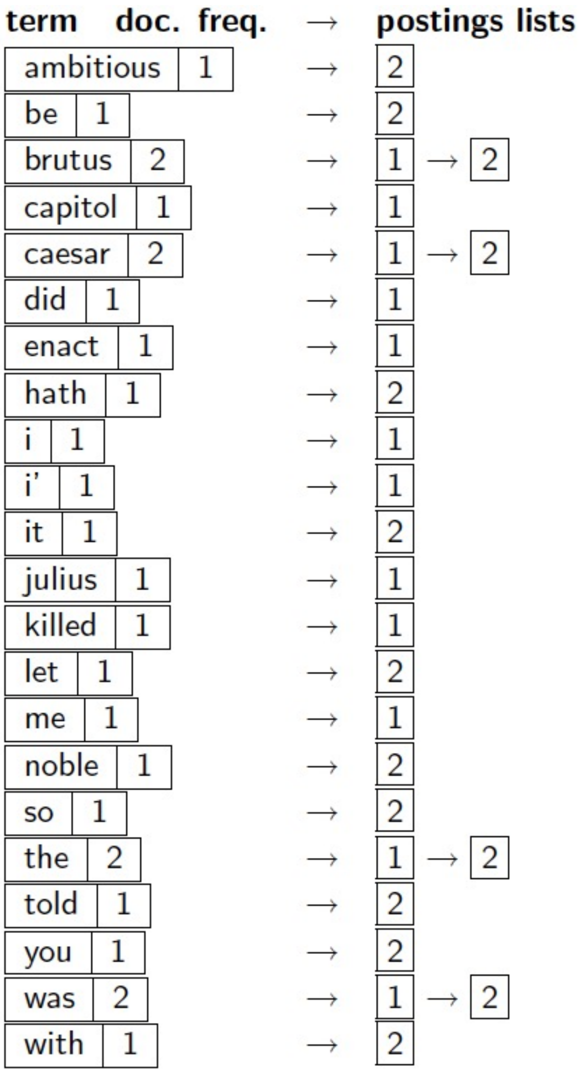
- **Tokenization**
  - Cut character sequence into word tokens
    - Deal with “John’s”, a state-of-the-art solution
- **Normalization**
  - Map text and query term to same form
    - You want U.S.A. and USA to match
- **Stemming**
  - We may wish different forms of a root to match
    - authorize, authorization
- **Stop words**
  - We may omit very common words (or not)
    - the, a, to, of

Indexer steps: Dictionary & Postings

- ✧ Multiple term entries in a single document are merged
- ✧ Split into Dictionary and Postings
- ✧ Doc. frequency information is added.



Term	docID
ambitious	2
be	2
brutus	1
brutus	2
capitol	1
caesar	1
caesar	2
caesar	2
did	1
enact	1
hath	1
I	1
I	1
i'	1
it	2
julius	1
killed	1
killed	1
let	2
me	1
noble	2
so	2
the	1
the	2
told	2
you	2
was	1
was	2
with	2



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

Summary

In this class, we focused on:

(a) Words / Terms / Lexical Units

(b) Tokenizing the terms

(c) Preparing Term – Document matrix

(d) Inverted Index Construction

i. Dictionary and Postings Lists

ii. Merging the Postings

iii. How much storage is required?

Questions
It's Your Time

Contact Information:

Dr. Rajendra Prasath
IIIT Sri City, Chittoor