

Boolean Retrieval.
[Postings list

Term Vocabulary / Tokenization /

Normalization in IR.

Documents \leftarrow units of retrieval.

Process these documents \leftarrow Passing of documents

- Converting byte sequences into a linear sequence of characters.

Issues:

- ① Different format / different language.
- ② Different encoding (UTF-8).
- ③ Character set in use.

- Use heuristics to deal with these problems rather than using full-fledged classification.

Document — ill defined.

html, pdf, pptx, docx etc.

Is a document a single file?
5 files?

Is a document n files?

Is a document a single email?

a chain of emails?
an email with a

pdf attachment

a Spanish email with
French pdf

Design ←
Choices
↓
Down-stream
application

Definitions:

Word — A delimited string of characters as it appears in running text (corpus).

Term — A "normalized" word.

Token — An instance of a word.

Type — An instance of distinct occurrence of a word (\approx term).

Inverted index construction.

Input.

Friends, Romans, countrymen

So let it be with Caesar

Output: (tokens).

→ terms → for which we could build postings lists.

friend

roman

so

let

Tokenization & Normalization

→ Splitting of string of characters into tokens/words.

(white space) / delimiters → split to get the tokens.

Issues

One word or many?

~~He~~ Hewlett-Packard

State-of-the-art

Co-eduction

San Francisco

Cheap San Francisco - Los Angeles fares

frequent

queries.

an IR

system

encounter

Chinese

monk.



char 1
and

char 2
skill

Ideogram

German

Computational Linguistics

Computer +
Linguistics.

tusaatsiarunnaungittualununga. (Intuit).
(I can't hear you very well).

"New York University"

York University

Numbers in their different avatars.

Date: 3/20/91
20/3/91

Mar 20, 1991

Phone numbers-

100. 2. 86. 144.

800. 234. 3323.

(800) 234-2323.

CamelCase

iPhone

eBay

Accents.

résumé

Universität or Universitaet

↓
umlauts.



Careful observation of the
user of an IR system.

→ drop accents

Resume.

[Did you
mean " "]

Case folding.

↓
Blindly convert everything to lowercase.

Caveats.

MIT and mit

← German (with).

Fed vs. fed.

Normalization.

↳ identify query terms that have the same form.

Equivalence
classes
of words

{ U.S.A

, USA - - - }

{ cars

, automobile - - - }

→ hand-crafted rules.

Asymmetric expansion:

~~❌~~ window → window, windows

windows → windows, Windows

Windows → X restricted to expand.

↓
named entity