Boolean Retrieval.
Postings list

Term Vocabulary / Tokenization / Normalization in IR.

Documents ← units of retrieval.

Process these documents ← 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?
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 appear in running text (corpus).

Term — A "normalized" word.

Token — An instance of a word.

Type — An instance of distict occurrence of a word (≈ term).

Inverted index construction.

Input.

Friends, Romans, countrymen | So let it be with Caesar

↳ terms → for which we could build postings lists...

Output: (tokens).

friend ✓ | roman ✓ | so ✓ | let ✓

Tokenization & Normalization

↳ Splitting of string of characters into tokens / words.

(whitespace) & delimiters → split to get the tokens.

Issues

One word or many?

~~Hewlett-Packard~~ "Hewlett-Packard" ←

State-of-the-art

co-education

San Francisco

cheap San Francisco - Los Angeles fares"

frequent queries. an IR system encounters

Chinese.

Ideogram.

German.

Computerlinguistik → Computer + Linguistik.

tusaatsiarunnangittualuujunga. (Inuit).

(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

Université vs Universitaet

↓

umlaots.

→ 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 (mit).

Fed vs. fed.

Normalization.

↳ identify query terms that have the same form.

Equivalence classes of terms: [{ U.S.A , USA ... } { cars , automobile ... }]

→ hand-crafted rules.

Assymetric expansion:

window → window, windows

windows → windows, Windows

Windows → ✗ restricted to expand.

↓

named entity