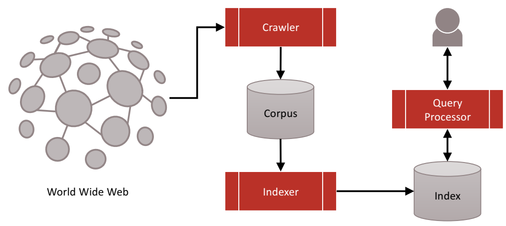
UFMG UNIVERSIDADE FEDERAL DE MINAS GERAIS

Ranking Models

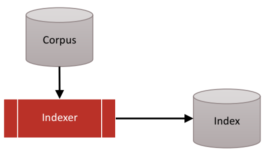
# Document Understanding

Rodrygo L. T. Santos
rodrygo@dcc.ufmg.br

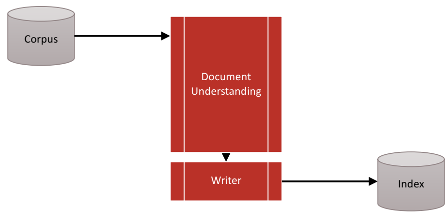
## Search components




## Search components




## Indexing overview

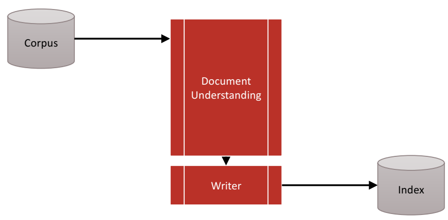



## Document understanding

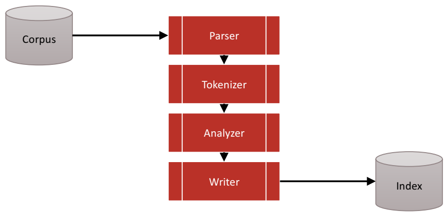
Making sense of text is a challenging task
- Not always clear what a document is
- Not always clear what a term is

Matching exact strings is too restrictive
- Not all words are of equal value in a search

## Indexing overview



## Indexing overview




## Document parsing

We previously assumed

- We know what a document is
- We can “machine-read” each document

This can be complex in reality…

## What is a document?

Or, in IR parlance, what is our retrieval unit?

- A single file?
- How about an email with 5 attachments?
- Or a book with 15 chapters?

What content types will be accepted?

- text/html? application/pdf? application/msword?

## How to read a document?

Must handle structure

- Text vs. binary, plain text vs. markup

It ain’t always beautiful

```
<div id="foo">
  <div id="bar">
    <span>Test</span>
</div>
```

## How to read a document?

Must handle encoding

- Translate between bits and characters

Sometimes, multiple, ill-specified ones

ノーベル平和賞を受賞したワンガリ・マータイさんが名
誉会長を務めるMOTTA I NA Iキャンペーンの一環として、
毎日新聞社とマガジンハウスは「私の、もったいない」
を募集します。皆様が日ごろ「もったいない」と感じて
実践していることや、それにまつわるエピソードを800字
以内の文章にまとめ、簡単な写真、イラスト、図などを
添えて10月20日までにお送りください。大賞受賞者には5
0万円相当の旅行券とエコ製品2点の副賞が贈られます

## Document tokenization

All along the watchtower
Princes kept the view
While all the women came and went
Barefoot servants, too
Outside in the cold distance
A wildcat did growl
Two riders were approaching
And the wind began to howl

| | |
|---|---|
| all | while |
| along | all |
| the | the |
| watchtower | women |
| princes | came |
| kept | and |
| the | went |
| view | … |

## How to tokenize?

One simple strategy (early IR systems)
- Any sequence of 3+ alphanumeric characters
- Terminated by a space or other special character
- Upper-case changed to lower-case

## What could go wrong?

## What could go wrong?

*Bigcorp's 2007 bi-annual report showed profits of 10%.*
↳ bigcorp s 2007 bi annual report showed profits of 10
↳ bigcorp 2007 annual report showed profits

Too much information lost
- Small tokenization decisions can have a major impact on the effectiveness of some queries

## Token length

Small words tend to be poorly discriminative
- a, an, be, of, to…

But they can also aid disambiguation
- ben e king, el paso, master p, world war ii

And even be crucial for matching
- xp, ma, pm, gm, j lo, c

## Special characters

**Apostrophes** can be a part of a word, a part of a possessive, or just a mistake
- rosie o'donnell, can't, don't, 80's, master's degree

**Accents and diacritics** can change meaning
- résumé vs. resume, cocô vs. coco

## Special characters

**Periods** can occur in numbers, abbreviations, URLs, ends of sentences, and other situations
- I.B.M., Ph.D., cs.umass.edu, F.E.A.R.

**Hyphens** are often not needed
- e-bay, wal-mart, active-x, cd-rom, t-shirts

## Numbers and lowercasing

**Numbers** can be important, including decimals

- nokia 3250, top 10 courses, united 93, quicktime 6.5

**Lowercasing** can change meaning

- Bush vs. bush, Apple vs. apple

## Non-delimited tokens

How to tokenize this?

White House aides wrestle with Trump's comments

How about these?

whitehouse.gov, #ImpeachTrump

And this?!?!

莎拉波娃现在居住在美国东南部的佛罗里达

## Token analysis

Discriminative power

Equivalence classing

Phrasing

Scoping

## Discriminative power

*Tropical fish are generally those fish found in aquatic tropical environments around the world, including both freshwater and saltwater species.*

**Document frequency (in millions)**

| | | | |
|---|---|---|---|
| saltwater | 46 | and | 25,270 |
| freshwater | 95 | in | 25,270 |
| aquatic | 118 | the | 25,270 |
| species | 377 | are | 15,830 |

## Stopping

Discard poorly discriminative words (aka ***stopwords***)

- a, an, and, are, as, at, be, by, for, from, has, he, in, is, it, its, not, of, on, or, that, the, to, was, were, will, with

Can be standardized or automatically derived

- Can be domain-specific (e.g. "click" for anchor text)

## Stopping

Reduce index space and response time

- May improve effectiveness

Discouraged in modern search engines

- Stopwords can be important in combinations

*To be, or not to be: that is the question.*

↳ question

## Equivalence classing

Reduce words to a canonical form
- Lexical equivalence
- Phonetic equivalence
- Semantic equivalence

## Lexical equivalence

Many morphological variations of words
- Inflectional (e.g., plurals, tenses)
- Derivational (e.g., making verbs into nouns)

In most cases, these have very similar meanings
- swimming, swam → swim

## Stemming

Reduce morphological variations to a stem
- Usually involves removing suffixes

Crude approximation of a principled lemmatization
- Ignores grammatical category
- Ignores surrounding context

Runs ***much*** faster!

## Porter's stemmer

A set of sequentially applied rules

| **Rule** | **Example** |
|---|---|
| SSES → SS | caresses → caress |
| IES → I | ponies → poni |
| SS → SS | caress → caress |
| S → | cats → cat |

## Stemming effectiveness

Stemming usually improves recall
- But can potentially hurt precision

False positive equivalence
- universal, university, universe → univers

False negative equivalence
- alumnus → alumnu, alumni → alumni

## Phonetic and semantic equivalence

Phonetic equivalence
- Reduce similar-sounding words to same form (e.g., Hermann ↔ Herman)

Semantic equivalence
- Reduce multiple surface forms to same entity (e.g., car ↔ automobile)

## Phrasing

Many queries are 2-3 word phrases
- [bob dylan lyrics]

More precise than single words
- Documents with “bob dylan” vs. “bob” and “dylan”

Less ambiguous
- “big apple” vs. “apple”

## Phrasing

Two broad strategies
- Syntactic phrasing
- Statistical phrasing

## Syntactic phrasing

Part-of-speech (POS) taggers can label words according to their syntactic role in natural language
- e.g., NN (singular noun), NNS (plural noun), VB (verb), VBD (verb, past tense), VBN (verb, past participle)

Phrases can be identified as simple noun groups
- Sequences of nouns, adjectives followed by nouns…

## POS tagging example

Document will describe marketing strategies carried out by U.S. companies for their agricultural chemicals.

↳ 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 ./.

## POS tagging example

Document will describe marketing strategies carried out by U.S. companies for their agricultural chemicals.

↳ Document/NN will/MD describe/VB <u>marketing/NN</u> <u>strategies/NNS</u> carried/VBD out/IN by/IN <u>U.S./NNP</u> <u>companies/NNS</u> for/IN their/PRP <u>agricultural/JJ</u> <u>chemicals/NNS</u> ./.

## Statistical phrasing

POS tagging is too slow for large collections
- Do we need a full syntactic analysis?

Simpler definition: phrases as n-grams
- Unigram: single words
- Bigram: 2-word sequence
- Trigram: 3-word sequence

## Statistical phrasing

N-gram frequencies form a Zipf distribution
- Some very frequent, lots less frequent
- Frequent n-grams tend to be meaningful phrases

Could index all n-grams up to a specified length
- Much faster than POS tagging
- Uses a lot of storage

## Statistical phrasing

Google n-grams
[Franz and Brants, 2006]

| | |
|---|---|
| # tokens | 1,024,908,267,229 |
| # sentences | 95,119,665,584 |
| **# 1-grams** | **13,588,391** |
| # 2-grams | 314,843,401 |
| # 3-grams | 977,069,902 |
| # 4-grams | 1,313,818,354 |
| # 5-grams | 1,176,470,663 |

## Scoping

Documents often have structure
- HTML tags (e.g., h1, h2, p, a)

Not all parts are equally important
- Document title, URL, metadata, body sections

Can record the scope of word occurrences
- Enable scoped queries and structural ranking models

## Summary

Document understanding improves representation
- Matching things rather than strings

Lots of important decisions
- May not know what's best at indexing time

Keep it simple, but keep it all!
- Index everything, defer complexity to querying time

## Summary

Indexing vs. querying

| | |
|---|---|
| Stopping | Query relaxation |
| Equivalence classing | Query expansion |
| Phrasing | Query segmentation |
| Scoping | Query scoping |

## References

Search Engines: Information Retrieval in Practice, Ch. 4
Croft et al., 2009

Introduction to Information Retrieval, Ch. 2
Manning et al., 2008

UFMG UNIVERSIDADE FEDERAL DE MINAS GERAIS

Coming next…

## Query Understanding

Rodrygo L. T. Santos
rodrygo@dcc.ufmg.br