Ghislain Fourny

Information Retrieval Spring 2019

3. Term vocabulary
What we have seen so far
Warm up
Boolean retrieval

Input
Set of documents
Boolean retrieval

Input
Set of documents

query

lawyer AND Penang AND NOT silver
Boolean retrieval

Input
Set of documents

Output
Subset of documents

query

lawyer AND Penang AND NOT silver
Simple boolean language (EBNF)

PrimaryExpr ::= Term | "(" Expr ")"

NotExpr ::= "NOT"? PrimaryExpr

AndExpr ::= NotExpr ("AND" NotExpr)*

Expr ::= NotExpr ("OR" AndExpr)*
Model and abstraction

Document as a list of words (with duplicates)
Model and abstraction

Document as a list of words (with duplicates)

Simplification

Document as a set of words
Model and abstraction

Document as a list of words (with duplicates)

Simplification

Document as a vector of booleans

(0 1 0 1 0 1 1 1 0 0 1 1 0 1 0 1 0 1 0 1 0)

Document as a set of words

Linearization
Term-document model

Documents as lists of words (with duplicates)

Simplification

Term-document bipartite graph
Term-document model

Term-document bipartite graph
Term-document model

Term-document bipartite graph

Adjacency matrix

\[
\begin{pmatrix}
1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\
0 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\
\end{pmatrix}
\]
**Term-document model**

- **Term-document bipartite graph**
- **Adjacency matrix**
  \[
  \begin{pmatrix}
  1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 \\
  0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\
  0 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\
  0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
  1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 \\
  0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 
  \end{pmatrix}
  \]
- **Postings**
Document

Term-document bipartite graph

Adjacency matrix

Postings
Term

- Term-document bipartite graph

Adjacency matrix

Postings
Posting

Term-document bipartite graph

Adjacency matrix

Postings
Index construction
Last time...

Plenty of simplifying assumptions + white magic
Index construction in reality...

Collect documents
Index construction in reality...

Collect documents

Tokenizing
Index construction in reality...

- Collect documents
- Tokenizing
- Linguistic preprocessing
Index construction in reality...

1. Collect documents
2. Tokenizing
3. Linguistic preprocessing
4. Build the index (postings list)
Documents
Collecting documents
Collecting documents: first challenge
Collecting documents: first challenge

Possess it merely. That it should come to this!
Collecting documents: sequence of characters

Possess it merely. That it should come to this!

Possess it merely. That it should come to this!
Character set

Possess it merely. That it should come to this!
Collecting documents: encoding

Possess it merely. That it should come to this!

Here: ASCII
# ASCII Table

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>0.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>1.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>2.</strong></td>
<td><strong>SP</strong></td>
<td>!</td>
<td>&quot;</td>
<td>#</td>
<td>$</td>
<td>%</td>
<td>&amp;</td>
<td>'</td>
<td>(</td>
<td>)</td>
<td>*</td>
<td>+</td>
<td>,</td>
<td>-</td>
<td>.</td>
</tr>
<tr>
<td><strong>3.</strong></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>:</td>
<td>:</td>
<td>&lt;</td>
<td>=</td>
<td>&gt;</td>
</tr>
<tr>
<td><strong>4.</strong></td>
<td>@</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
<td>I</td>
<td>J</td>
<td>K</td>
<td>L</td>
<td>M</td>
<td>N</td>
</tr>
<tr>
<td><strong>5.</strong></td>
<td>P</td>
<td>Q</td>
<td>R</td>
<td>S</td>
<td>T</td>
<td>U</td>
<td>V</td>
<td>W</td>
<td>X</td>
<td>Y</td>
<td>Z</td>
<td>[</td>
<td>\</td>
<td>]</td>
<td>^</td>
</tr>
<tr>
<td><strong>6.</strong></td>
<td>`</td>
<td>a</td>
<td>b</td>
<td>c</td>
<td>d</td>
<td>e</td>
<td>f</td>
<td>g</td>
<td>h</td>
<td>i</td>
<td>j</td>
<td>k</td>
<td>l</td>
<td>m</td>
<td>n</td>
</tr>
<tr>
<td><strong>7.</strong></td>
<td>p</td>
<td>q</td>
<td>r</td>
<td>s</td>
<td>t</td>
<td>u</td>
<td>v</td>
<td>w</td>
<td>x</td>
<td>y</td>
<td>z</td>
<td>{</td>
<td></td>
<td></td>
<td>}</td>
</tr>
</tbody>
</table>
### ASCII Table

<table>
<thead>
<tr>
<th>Control characters</th>
<th>Control characters</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.</td>
<td>SP</td>
</tr>
<tr>
<td>1.</td>
<td>! &quot; # $ % &amp; ' ( ) * + , - . /</td>
</tr>
<tr>
<td>2.</td>
<td>0 1 2 3 4 5 6 7 8 9 : : &lt; = &gt; ?</td>
</tr>
<tr>
<td>3.</td>
<td>@ A B C D E F G H I J K L M N O</td>
</tr>
<tr>
<td>4.</td>
<td>` a b c d e f g h i j k l m n o</td>
</tr>
<tr>
<td>5.</td>
<td>P Q R S T U V W X Y Z [ \ ] ^ _</td>
</tr>
<tr>
<td>6.</td>
<td>p q r s t u v w x y z {</td>
</tr>
</tbody>
</table>

50 (hex)
## ASCII Table

<table>
<thead>
<tr>
<th>.0</th>
<th>.1</th>
<th>.2</th>
<th>.3</th>
<th>.4</th>
<th>.5</th>
<th>.6</th>
<th>.7</th>
<th>.8</th>
<th>.9</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>SP</td>
<td>!</td>
<td>&quot;</td>
<td>#</td>
<td>$</td>
<td>%</td>
<td>&amp;</td>
<td>'</td>
<td>(</td>
<td>)</td>
<td>*</td>
<td>+</td>
<td>,</td>
<td>-</td>
<td>.</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>:</td>
<td>:</td>
<td>&lt;</td>
<td>=</td>
<td>&gt;</td>
</tr>
<tr>
<td>4</td>
<td>@</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
<td>I</td>
<td>J</td>
<td>K</td>
<td>L</td>
<td>M</td>
<td>N</td>
</tr>
<tr>
<td>5</td>
<td>P</td>
<td>Q</td>
<td>R</td>
<td>S</td>
<td>T</td>
<td>U</td>
<td>V</td>
<td>W</td>
<td>X</td>
<td>Y</td>
<td>Z</td>
<td>[</td>
<td>\</td>
<td>]</td>
<td>^</td>
</tr>
<tr>
<td>6</td>
<td>\</td>
<td>a</td>
<td>b</td>
<td>c</td>
<td>d</td>
<td>e</td>
<td>f</td>
<td>g</td>
<td>h</td>
<td>i</td>
<td>j</td>
<td>k</td>
<td>l</td>
<td>m</td>
<td>n</td>
</tr>
<tr>
<td>7</td>
<td>p</td>
<td>q</td>
<td>r</td>
<td>s</td>
<td>t</td>
<td>u</td>
<td>v</td>
<td>w</td>
<td>x</td>
<td>y</td>
<td>z</td>
<td>{</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

50 (hex) 0101000000
## UTF-8

<table>
<thead>
<tr>
<th>Character</th>
<th>Codepoint</th>
<th>Codepoint in binary</th>
<th>UTF-8 (variable length)</th>
</tr>
</thead>
</table>
**UTF-8**

<table>
<thead>
<tr>
<th>Character</th>
<th>Codepoint</th>
<th>Codepoint in binary</th>
<th>UTF-8 (variable length)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>U+0050</td>
<td>1010 000</td>
<td></td>
</tr>
</tbody>
</table>
UTF-8

<table>
<thead>
<tr>
<th>Character</th>
<th>Codepoint</th>
<th>Codepoint in binary</th>
<th>UTF-8 (variable length)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>U+0050</td>
<td>1010 000</td>
<td></td>
</tr>
</tbody>
</table>

Less than 7 bits
## UTF-8

<table>
<thead>
<tr>
<th>Character</th>
<th>Codepoint</th>
<th>Codepoint in binary</th>
<th>UTF-8 (variable length)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>U+0050</td>
<td>1010 000</td>
<td>01010000</td>
</tr>
</tbody>
</table>

Less than 7 bits
# UTF-8

<table>
<thead>
<tr>
<th>Character</th>
<th>Codepoint</th>
<th>Codepoint in binary</th>
<th>UTF-8 (variable length)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>U+0050</td>
<td>1010 000</td>
<td>01010000</td>
</tr>
<tr>
<td>π</td>
<td>U+03C0</td>
<td>11 1100 0000</td>
<td></td>
</tr>
</tbody>
</table>
UTF-8

<table>
<thead>
<tr>
<th>Character</th>
<th>Codepoint</th>
<th>Codepoint in binary</th>
<th>UTF-8 (variable length)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>U+0050</td>
<td>1010 000</td>
<td>01010000</td>
</tr>
<tr>
<td>π</td>
<td>U+03C0</td>
<td>11 1100 0000</td>
<td></td>
</tr>
</tbody>
</table>

Less than 7 bits

Less than 11 bits
## UTF-8

<table>
<thead>
<tr>
<th>Character</th>
<th>Codepoint</th>
<th>Codepoint in binary</th>
<th>UTF-8 (variable length)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>U+0050</td>
<td>1010 000</td>
<td>01010000</td>
</tr>
<tr>
<td>π</td>
<td>U+03C0</td>
<td>11 1100 0000</td>
<td>11001111 10000000</td>
</tr>
</tbody>
</table>

Less than 7 bits

Less than 11 bits
# UTF-8

<table>
<thead>
<tr>
<th>Character</th>
<th>Codepoint</th>
<th>Codepoint in binary</th>
<th>UTF-8 (variable length)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>U+0050</td>
<td>1010 000</td>
<td>01010000</td>
</tr>
<tr>
<td>π</td>
<td>U+03C0</td>
<td>11 1100 0000</td>
<td>11001111 10000000</td>
</tr>
<tr>
<td>€</td>
<td>U+20AC</td>
<td>10 0000 1010 1100</td>
<td></td>
</tr>
</tbody>
</table>
UTF-8

<table>
<thead>
<tr>
<th>Character</th>
<th>Codepoint</th>
<th>Codepoint in binary</th>
<th>UTF-8 (variable length)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>U+0050</td>
<td>1010 000</td>
<td>01010000</td>
</tr>
<tr>
<td>π</td>
<td>U+03C0</td>
<td>11 1100 0000</td>
<td>11001111 10000000</td>
</tr>
<tr>
<td>€</td>
<td>U+20AC</td>
<td>10 0000 1010 1100</td>
<td></td>
</tr>
</tbody>
</table>

Less than 7 bits
Less than 11 bits
Less than 16 bits
# UTF-8

<table>
<thead>
<tr>
<th>Character</th>
<th>Codepoint</th>
<th>Codepoint in binary</th>
<th>UTF-8 (variable length)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>U+0050</td>
<td>1010 000</td>
<td>01010000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Less than 7 bits</td>
<td></td>
</tr>
<tr>
<td>π</td>
<td>U+03C0</td>
<td>11 1100 0000</td>
<td>11001111 10000000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Less than 11 bits</td>
<td></td>
</tr>
<tr>
<td>€</td>
<td>U+20AC</td>
<td>10 0000 1010 1100</td>
<td>11100010 10000010 10101100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Less than 16 bits</td>
<td></td>
</tr>
</tbody>
</table>
Collecting documents: first challenge

ASCII?
UTF-8?
ISO-Latin-1?
...

Collecting documents: first challenge

User-defined
Annotated in document
Machine learning
...

...
Collecting documents: first challenge

Software-specific encoding
Collecting documents: first challenge

Data-specific encoding

`&amp;`

`&lt;`

`...`
Collecting documents: first challenge

Binary encoding
Collecting documents: first challenge

Classification (e.g., ML)

Language
UTF-8
Encoding
Type

PDF
Collecting documents: second challenge
Collecting documents: second challenge

Fine-grained documents
(Example: E-Mail archive)
Collecting documents: second challenge
Collecting documents: second challenge
Collecting documents: second challenge

Grouping to a single document (Example: LaTeX source)
Term Vocabulary
Building an inverted index

**Document 1**
You come most carefully upon your hour.

**Document 2**
Take thy fair hour, Laertes; time be thine,

**Document 3**
My hour is almost come,

**Document 4**
Possess it merely. That it should come to this!
Building an inverted index

**Document 1**
You come most carefully upon your hour

**Document 2**
Take thy fair hour Laertes time be thine

**Document 3**
My hour is almost come

**Document 4**
Possess it merely That it should come to this
Building an inverted index

This is not trivial!

name@example.com

isn't

Jake O'Neill

the learn-it-all-by-heart methodology

website vs. web site

Kanton Basel Stadt
Corner cases

Hewlett-Packard
State-of-the-art
co-education
the hold-him-back-and-drag-him-away
maneuver data base
San Francisco
Los Angeles-based company
cheap San Francisco-Los Angeles fares
York University vs. New York University

Credits: H. Schütze, LMU
Corner cases: numbers

3/20/91
20/3/91
Mar 20, 1991
B-52
100.2.86.144
(800) 234-2333
800.234.2333
English

I would like a coffee
You would like a coffee
He would like a coffee
We would like a coffee
You would like a coffee
They would like a coffee
I want a coffee
You want a coffee
He wants a coffee
We want a coffee
You want a coffee
They want a coffee
Swedish

Jag **skulle vilja ha** en kaffe
Du **skulle vilja ha** en kaffe
Han **skulle vilja ha** en kaffe
Vi **skulle vilja ha** en kaffe
Ni **skulle vilja ha** en kaffe
De **skulle vilja ha** en kaffe
German

Ich möchte einen Kaffee
Du möchtest einen Kaffee
Er möchte einen Kaffee
Wir möchten einen Kaffee
Ihr möchtet einen Kaffee
Sie möchten einen Kaffee
German

Donaudampfschiffahrts-Elektrizitätenhauptbetriebswerkbaununterbeamtengesellschaft
Swiss German

I ha tänkt, du heggish en kaffee welle trinke
中华人民共和国，简称中国，是位于东亚的社会主义国家，首都位于北京[12]。中国领土面积约为960万平方公里，是世界上纯陆地面积第二大、陆地面积第三大、总面积第三大或第四大的国家[註 11][13]，当中划分为23个省份[註 12]、5个自治区、4个直辖市和2个特别行政区。
मैं इस टर्म अच्छा पढ़ रहा हूँ।
मैं इस टर्म अच्छा पढ़ रही हूँ।
मुझे इंफर्मैशन रिट्रीवल बहुत पसंद है।
 मैं इस टर्म अच्छा पढ़ रहा हूँ।
 मैं इस टर्म अच्छा पढ़ रही हैं।
 मुझे इंफ़र्मेशन रिट्रीवल बहुत पसंद है।

Male speaker
1st person

Oblique case

Female speaker

3rd person

To me
Polysynthetic Languages
Siberian Yupik

neghyaghtughyugumayaghpetaallu

Source: Polysynthetic Language: Central Siberian Yupik
W J De Reuse
Polysynthetic Languages
Siberian Yupik

neghyaghtughyugumayaghpetaallu

eat
Polysynthetic Languages
Siberian Yupik

eat  go to
Polysynthetic Languages
Siberian Yupik

neghyaghtughyugumayaghpetaallu

eat  go to  want
Polysynthetic Languages
Siberian Yupik

Past

neghyaghtughyugumayaghpetaallu

eat  go to  want
Polysynthetic Languages
Siberian Yupik

<Frustration>
Past
neghyaghtughyugumayaghpetaallu

eat   go to   want
Polysynthetic Languages
Siberian Yupik

<Frustration>
Past
neghyaghtughyugumayaghpetaallu

eat  go to  want

inferential
evidential
(turns out)
Polysynthetic Languages
Siberian Yupik

<Frustration>
Past 3rd/3rd
neghyaghtughyugumayaghpetaaullu

eat   go to   want

inferential
evidential
(turns out)
Polysynthetic Languages
Siberian Yupik

<Frustration>
Past 3rd/3rd
neghyaghtughyugumayaghpetaallu

eat  go to  want also

inferential evidential (turns out)
Polysynthetic Languages
Siberian Yupik

neghyaghtughyugumayaghanpettaallu

Also, it turns out she/he wanted to go eat it, but...
Tokenize

You come most carefully upon your hour

Take fair thy hour Laertes time be thine

My hour is almost come

Possess it merely That it should come to this
Token

You come most carefully upon your hour

Take fair thy hour Laertes time be thine

My hour is almost come

Possess that merely That it should come to this

Token = grouped sequence of characters
Type

You come most carefully upon your hour

Take fair thy hour Laertes time be thine

My hour is almost come

Possess it merely That it should come to this

Type = equivalence class (same sequences)
Type

Type = equivalence class (same sequences)
Stop words

You come most carefully upon your hour
Take fair thy hour Laertes time be thine
My hour is almost come
Possess it merely That it should come to this
Reuters's list

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
Trend

Large list (200-300)

No list at all
Equivalence classes of terms (types)

- U.S.A.
  - USA
- today
  - to-day
- window
  - windows
  - Windows
- Zurich
  - Zürich
  - Zuerich
  - Züri
Normalization

U.S.A.  USA

to-day  today

Rules that remove characters are the easy part.
Expansion (rather than equivalence classes)

Introduces asymmetry!

Windows → Windows
windows → windows, Windows, window
window → window, windows
When to expand?

Upon indexing

Lift |
When to expand?

Upon indexing

Lift | Elevator

Lift | 1 | 4 | 5 | 6
Elevator | 1 | 4 | 5 | 6

Expansion

Lift | 1 | 5
Elevator | 1 | 4 | 6
When to expand?

Upon indexing:

Upon querying:

- Lift
- Elevator

Expansion

Lift
- 1
- 4
- 5
- 6

Elevator
- 1
- 4
- 6
When to expand?

Upon indexing

Lift | Expansion

Lift → 1 → 4 → 5 → 6
Elevator → 1 → 4 → 5 → 6

Upon querying

Lift OR Elevator

Lift → 1 → 5
Elevator → 1 → 4 → 6
Normalization: accents and diacritics

cliché → cliche

Zürich → Zuerich
Normalization: lowercasing

C.A.T → cat
Schweiz → schweiz
Normalization: truecasing

The company Apple has launched a new iProduct.

Apples fall from trees.
Stemming

"Chop" letters of the word
Porter Stemmer

(m>0) ENCI -> ENCE  valenci -> valence
(m>0) ANCI -> ANCE  hesitanci -> hesitance
(m>0) IZER -> IZE    digitizer -> digitize
(m>0) ABLI -> ABLE   conformabli -> conformable
(m>0) ALLI -> AL     radicalli -> radical
(m>0) ENTLI -> ENT   differentli -> different
(m>0) ELI -> E       vileli - > vile
(m>0) OUSLI -> OUS   analogousli -> analogous
(m>0) IZATION -> IZE vietnamization -> vietnamize
(m>0) ATION -> ATE   predication -> predicate
(m>0) ATOR -> ATE    operator -> operate
(m>0) ALISM -> AL    feudalism -> feudal
(m>0) IVENESS -> IVE decisiveness -> decisive
(m>0) FULNESS -> FUL hopefulness -> hopeful
(m>0) OUSNESS -> OUS callousness -> callous

https://tartarus.org/martin/PorterStemmer/
Porter Stemmer

Such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation.

Porter
such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation.

Lovins
such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation.

Paice
such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation.

Source: the textbook (Introduction to Information Retrieval)
Lemmatization

Building equivalence classes or expanding with

Natural Language Processing
The Vauquois Triangle

Interlingua

Semantic Structure

Semantic Structure

Syntactic Structure

Syntactic Structure

Words

Words

Semantic

Syntactic

Direct

Transfer
Lemmatization

Full morphological analysis

computer
compute
computes
computed
computation
computing
...

Lemmatization or Stemming?

Lemmatization does not help and can even degrade performance for English documents.
Lemmatization or Stemming?

Lemmatization can help with language that have more morphology.
Skip lists
Reminder: Postings (Standard Inverted Index)

almost: 1 3
be: 1 2
carefully: 1 1
come: 3 1
fair: 1 2
hour: 3 1
is: 1 3
it: 1 4
Laertes: 1 2
merely: 1 4
most: 1 1
my: 1 3
possess: 1 4
should: 1 4
take: 1 2
that: 1 4
thine: 1 2
this: 1 4
thy: 1 2
time: 1 2
to: 1 4
upon: 1 1
you: 1 1
your: 1 1
Reminder: Intersection algorithm

List A

1 ➔ 2 ➔ 4 ➔ 5 ➔ 8 ➔ 9 ➔ 10 ➔ 12

List B

1 ➔ 3 ➔ 4 ➔ 6 ➔ 7 ➔ 8 ➔ 11 ➔ 12

Intersection of A and B

1 ➔ 4 ➔ 8 ➔ 12

End
Another example

List A: 1 → 2 → 3 → 4 → 5 → 6 → 7 → 8 → 9 → 10 → 12

List B: 1 → 3 → 10 → 11 → 12 → 14

Intersection of A and B
Another example

List A

List B

Intersection of A and B
Another example

List A

1 → 2 → 3 → 4 → 5 → 6 → 7 → 8 → 9 → 10 → 12

List B

1 → 3 → 10 → 11 → 12 → 14

Intersection of A and B

1
Another example

List A

1 → 2 → 3 → 4 → 5 → 6 → 7 → 8 → 9 → 10 → 12

List B

1 → 3 → 10 → 11 → 12 → 14

Intersection of A and B

1 → 3
Another example

List A
1 → 2 → 3 → 4 → 5 → 6 → 7 → 8 → 9 → 10 → 12

List B
1 → 3 → 10 → 11 → 12 → 14

Intersection of A and B
1 → 3
Another example

List A

List B

Intersection of A and B
Another example

List A

1 → 2 → 3 → 4 → 5 → 6 → 7 → 8 → 9 → 10 → 12

List B

1 → 3 → 10 → 11 → 12 → 14

Intersection of A and B

1 → 3
Another example

List A: 1 → 2 → 3 → 4 → 5 → 6 → 7 → 8 → 9 → 10 → 12

List B: 1 → 3 → 10 → 11 → 12 → 14

Intersection of A and B: 1 → 3
Another example

List A

List B

Intersection of A and B
Another example

List A

List B

Intersection of A and B
Another example

List A

1 → 2 → 3 → 4 → 5 → 6 → 7 → 8 → 9 → 10 → 12

List B

1 → 3 → 10 → 11 → 12 → 14

Intersection of A and B

1 → 3 → 10
Another example

List A

List B

Intersection of A and B
Another example

List A  1 → 2 → 3 → 4 → 5 → 6 → 7 → 8 → 9 → 10 → 12

List B  1 → 3 → 10 → 11 → 12 → 14

Intersection of A and B  1 → 3 → 10 → 12
Another example

List A: 1 → 2 → 3 → 4 → 5 → 6 → 7 → 8 → 9 → 10 → 12

List B: 1 → 3 → 10 → 11 → 12 → 14

Intersection of A and B: 1 → 3 → 10 → 12
Another example

List A: 1 → 2 → 3 → 4 → 5 → 6 → 7 → 8 → 9 → 10 → 12

List B: 1 → 3 → 10 → 11 → 12 → 14

Intersection of A and B: 1 → 3 → 10 → 12
Another example

How could we make this faster?
Another example

List A: 1 → 2 → 3 → 4 → 5 → 6 → 7 → 8 → 9 → 10 → 12

List B: 1 → 3 → 10 → 11 → 12 → 14

Intersection of A and B: 1 → 3
Another example

List A

1 → 2 → 3 → 4 → 5 → 6 → 7 → 8 → 9 → 10 → 12

List B

1 → 3 → 10 → 11 → 12 → 14

Intersection of A and B

1 → 3

Magical pointer
Another example

List A: 1 → 2 → 3 → 4 → 5 → 6 → 7 → 8 → 9 → 10 → 12

List B: 1 → 3 → 10 → 11 → 12 → 14

Intersection of A and B: 1 → 3 → 10
In practice
In practice

Two short skips:
In practice

Two short skips:
many comparisons
waste of space
In practice

Two long skips:
- few comparisons
- not many real opportunities to skip
In practice

Two long skips:
In practice

\[ \sqrt{\text{Number of postings}} \]
Phrase search
Phrase queries

![Google search results for "eth zurich"](image)
Phrase queries
With a posting list

"ETH Zürich"

Quotes = phrase search
With a posting list

ETH

1 → 2 → 4 → 5 → 8 → 9 → 10

Zurich

1 → 3 → 4 → 6 → 7 → 8 → 11
With a posting list

ETH

1 → 2 → 4 → 5 → 8 → 9 → 10

Zurich

1 → 3 → 4 → 6 → 7 → 8 → 11

We have no information on the proximity of the terms!
Phrase search: approaches
Phrase search: approaches

Biword indices
Phrase search: approaches

Biword indices

Positional indices
Bi-word indices

Help ETH Zurich to flexibly react to new challenges and to set new accents in the future.
Bi-word indices

Help ETH Zurich to flexibly react to new challenges and to set new accents in the future.
Bi-word indices

Help ETH Zurich to flexibly react to new challenges and to set new accents in the future.
Bi-word indices

Help ETH Zurich to flexibly react to new challenges and to set new accents in the future.
Help ETH Zurich to flexibly react to new challenges and to set new accents in the future.
Bi-word indices

Help ETH Zurich to flexibly react to new challenges and to set new accents in the future.
Help ETH Zurich to flexibly react to new challenges and to set new accents in the future.
Help ETH Zurich to flexibly react to new challenges and to set new accents in the future.
Help ETH Zurich to flexibly react to new challenges and to set new accents in the future.

Bi-word indices

Index

Help ETH

ETH Zurich

Zurich to

to flexibly

flexibly react

react to
Bi-word indices

"ETH Zurich"
Bi-word indices

"ETH Zurich"
Bi-word indices

"Help ETH Zurich to flexibly react"
Bi-word indices

"Help ETH Zurich to flexibly react"

"Help ETH" AND "ETH Zurich" AND "ETH to" AND "to flexibly" AND "flexibly react"
Bi-word indices

"Help ETH Zurich to flexibly react"

"Help ETH" AND "ETH Zurich" AND "ETH to" AND "to flexibly" AND "flexibly react"
Bi-word indices

"Help ETH Zurich to flexibly react"

"Help ETH" AND "ETH Zurich" AND "ETH to" AND "to flexibly" AND "flexibly react"
Inconvenient?
Inconvenient?

"Help ETH Zurich to flexibly react"

"Help ETH" AND "ETH Zurich" AND "ETH to" AND "to flexibly" AND "flexibly react"
Inconvenient?

"Help ETH Zurich to flexibly react"

"Help ETH" AND "ETH Zurich" AND "ETH to" AND "to flexibly" AND "flexibly react"

Help ETH Zurich to introduce techniques to flexibly react...
Inconvenient?

"Help ETH Zurich to flexibly react"|

"Help ETH" AND "ETH Zurich" AND "ETH to" AND "to flexibly" AND "flexibly react"|

Help ETH Zurich to introduce techniques to flexibly react...
Inconvenient?

"Help ETH Zurich to flexibly react"

"Help ETH" AND "ETH Zurich" AND "ETH to" AND "to flexibly" AND "flexibly react"

Help ETH Zurich to introduce techniques to flexibly react...
Inconvenient?

"Help ETH Zurich to flexibly react"

"Help ETH" AND "ETH Zurich" AND "ETH to" AND "to flexibly" AND "flexibly react"

Help ETH Zurich to introduce techniques to flexibly react...

✅  ✅  ✅
Inconvenient?

"Help ETH Zurich to flexibly react"

"Help ETH" AND "ETH Zurich" AND "ETH to" AND "to flexibly" AND "flexibly react"

Help ETH Zurich to introduce techniques to flexibly react...
Inconvenient?

"Help ETH Zurich to flexibly react"

"Help ETH" AND "ETH Zurich" AND "ETH to" AND "to flexibly" AND "flexibly react"

Help ETH Zurich to introduce techniques to flexibly react...
Inconvenient?

"Help ETH Zurich to flexibly react"

"Help ETH" AND "ETH Zurich" AND "ETH to" AND "to flexibly" AND "flexibly react"

Help ETH Zurich to introduce techniques to flexibly react...

False positive!
Part-of-speech tagging

Help ETH Zurich to introduce techniques so as to flexibly react...
Part-of-speech tagging

Help ETH Zurich to introduce techniques so as to flexibly react...
Part-of-speech tagging

Help ETH Zurich to introduce techniques so as to flexibly react...
Part-of-speech tagging

NX*N

Help ETH Zurich to introduce techniques so as to flexibly react...
Part-of-speech tagging

Help ETH Zurich to introduce techniques so as to flexibly react...

\[ Nx^*N \]
Part-of-speech tagging

NX*N

Help ETH Zurich to introduce techniques so as to flexibly react...
Bi-word indices

"Help ETH Zurich to flexibly react" | "Help ETH" AND "ETH Zurich" AND "ETH to" AND "to flexibly" AND "flexibly react" | Intersect

<table>
<thead>
<tr>
<th>Help ETH</th>
<th>ETH Zurich</th>
<th>Zurich to</th>
<th>to flexibly</th>
<th>flexibly react</th>
<th>react to</th>
</tr>
</thead>
</table>

172
Phrase indices (3)

"Help ETH Zurich to flexibly react"

"Help ETH Zurich" AND "ETH Zurich to" AND "ETH to flexibly" AND "to flexibly react"

<table>
<thead>
<tr>
<th>Intersection</th>
<th>Help ETH Zurich</th>
<th>ETH Zurich to</th>
<th>Zurich to flexibly</th>
<th>to flexibly react</th>
<th>flexibly react to</th>
</tr>
</thead>
<tbody>
<tr>
<td>Help ETH Zurich</td>
<td>🟢</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ETH Zurich to</td>
<td></td>
<td>🟢</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zurich to flexibly</td>
<td></td>
<td></td>
<td>🟢</td>
<td></td>
<td></td>
</tr>
<tr>
<td>to flexibly react</td>
<td></td>
<td></td>
<td></td>
<td>🟢</td>
<td></td>
</tr>
<tr>
<td>flexibly react to</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>🟢</td>
</tr>
</tbody>
</table>
Phrase indices (4)

"Help ETH Zurich to flexibly react"

"Help ETH Zurich to" AND "ETH Zurich to flexibly" AND "Zurich to flexibly react"

Intersect
Improves on false positive issue

"Help ETH Zurich to flexibly react"

"Help ETH Zurich" AND "ETH Zurich to" AND "Zurich to flexibly" AND "to flexibly react"

Help ETH Zurich to introduce techniques to flexibly react...

True negative
Inconvenient
Inconvenient

Size of the vocabulary increases
Inconvenient

Size of the vocabulary increases exponentially

\((\# \text{ Terms})^n\)
Help ETH Zurich to flexibly react to new challenges and to set new accents in the future.
Help ETH Zurich to flexibly react to new challenges and to set new accents in the future.
Positional index

Help ETH Zurich to flexibly react to new challenges and to set new accents in the future.
Positional index

Help ETH Zurich to flexibly react to new challenges and to set new accents in the future.

term frequency
Help ETH Zurich to flexibly react to new challenges and to set new accents in the future.
Help ETH Zurich to flexibly react to new challenges and to set new accents in the future.

Positional indexing
Help ETH Zurich to flexibly react to new challenges and to set new accents in the future.
Help ETH Zurich to flexibly react to new challenges and to set new accents in the future.
Help ETH Zurich to flexibly react to new challenges and to set new accents in the future.
Help ETH Zurich to flexibly react to new challenges and to set new accents in the future.
Help ETH Zurich to flexibly react to new challenges and to set new accents in the future.
Positional index

"ETH Zurich"]
Positional index

"ETH Zurich"
Positional index

"ETH Zurich"
This week's reading

Chapter 2

The Term Vocabulary and Postings Lists