2. Boolean Retrieval
Data Shapes
Data Shapes: Tables

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data Shapes: Trees
Data Shapes: Graphs
Data Shapes: Cubes
Data Shapes: Unstructured text


Information Retrieval

Finding unstructured data
Information Retrieval

Finding unstructured data

that satisfies an information need
Information Retrieval

Finding unstructured data

that satisfies an information need

from within large collections
Poll

Go now to:

https://eduapp-app1.ethz.ch/

or install EduApp 2.x
First, naive approach
Mr. Sherlock Holmes, who was usually very late in the mornings, save upon those not infrequent occasions when he was up all night, was seated at the breakfast table. I stood upon the hearth-rug and picked up the stick which our visitor had left behind him the night before. It was a fine, thick piece of wood, bulbous-headed, of the sort which is known as a "Penang lawyer." Just under the head was a broad silver band nearly an inch across. "To James Mortimer, M.R.C.S., from his friends of the C.C.H.," was engraved upon it, with the date "1884." It was just such a stick as the old-fashioned family practitioner used to carry--dignified, solid, and reassuring. "Well, Watson, what do you make of it?"
Holmes was sitting with his back to me, and I had given him no sign of my occupation. "How did you know what I was doing? I believe you have eyes in the back of your head."
Naive approach

Q: lawyer
Mr. Sherlock Holmes, who was usually very late in the mornings, save upon those not infrequent occasions when he was up all night, was seated at the breakfast table. I stood upon the hearth-rug and picked up the stick which our visitor had left behind him the night before. It was a fine, thick piece of wood, bulbous-headed, of the sort which is known as a "Penang lawyer." Just under the head was a broad silver band nearly an inch across. "To James Mortimer, M.R.C.S., from his friends of the C.C.H.," was engraved upon it, with the date "1884." It was just such a stick as the old-fashioned family practitioner used to carry--dignified, solid, and reassuring. "Well, Watson, what do you make of it?"

Holmes was sitting with his back to me, and I had given him no sign of my occupation. "How did you know what I was doing? I believe you have eyes in the back of your head."
CHAPTER I
Mr. Sherlock Holmes
Mr. Sherlock Holmes, who was usually very late in the mornings, save upon those not infrequent occasions when he was up all night, was seated at the breakfast table. I stood upon the hearth-rug and picked up the stick which our visitor had left behind him the night before. It was a fine, thick piece of wood, bulbous-headed, of the sort which is known as a "Penang lawyer." Just under the head was a broad silver band nearly an inch across. "To James Mortimer, M.R.C.S., from his friends of the C.C.H.," was engraved upon it, with the date "1884." It was just such a stick as the old-fashioned family practitioner used to carry--dignified, solid, and reassuring.
"Well, Watson, what do you make of it?"
Holmes was sitting with his back to me, and I had given him no sign of my occupation.
"How did you know what I was doing? I believe you have eyes in the back of your head."
CHAPTER I
Mr. Sherlock Holmes

Mr. Sherlock Holmes, who was usually very late in the mornings, save upon those not infrequent occasions when he was up all night, was seated at the breakfast table. I stood upon the hearth-rug and picked up the stick which our visitor had left behind him the night before. It was a fine, thick piece of wood, bulbous-headed, of the sort which is known as a "Penang lawyer." Just under the head was a broad silver band nearly an inch across. "To James Mortimer, M.R.C.S., from his friends of the C.C.H.," was engraved upon it, with the date "1884." It was just such a stick as the old-fashioned family practitioner used to carry--dignified, solid, and reassuring.

"Well, Watson, what do you make of it?"

Holmes was sitting with his back to me, and I had given him no sign of my occupation.

"How did you know what I was doing? I believe you have eyes in the back of your head."
Naive approach

$ grep lawyer corpus.txt
Naive approach

$ grep lawyer corpus.txt

William Whyte was. Some pragmatical seventeenth century lawyer, I He was a lawyer. That sounded ominous. What was the relation between "Her banker or her lawyer. There is that double possibility. But I am "To an English lawyer named Norton."
and send down to Fordham, the Horsham lawyer."
"I did as he ordered, and when the lawyer arrived I was asked to step "I signed the paper as directed, and the lawyer took it away with violin-player, boxer, swordsman, lawyer, and self-poisoner by cocaine which was a Penang-lawyer weighted with lead, was just such a weapon a lawyer with a good practice. They had one child, but the yellow has some claim on half Cunningham's estate, and the lawyers have been lawyer whose name was given in the paper. There we met two gentlemen, as he followed the argument of the lawyer. At least, I thought, when is known as a "Penang lawyer." Just under the head was a broad silver Holmes to the porter. "A lawyer, is he not, gray-headed, and walks Reilly the lawyer and take the defense upon myself. Take my word for "Exactly," said I. "A plausible lawyer could make it out as an act of daughter of Augusto Barelli, who was the chief lawyer and once the I showed it to Mr. Sutro, my lawyer, who lives in Harrow. He said to lawyer of yours a capable man?"
This Sutro, of course, is her lawyer. I made a mistake, I fear, in gentleman, who introduced himself as the lawyer, together with a "I only heard of it this morning," the lawyer explained.
I have been recommended to you by my lawyers, but indeed the matter to the American lawyer. ... Very good. Good-bye!
lawyer burst excitedly into the room.
"I can only suppose that this American lawyer put it in himself. What say nothing, as her lawyer had advised her to reserve her defence. We
Naive approach

Q: lawyer AND Penang
Naive approach

$ grep lawyer corpus.txt | grep Penang
Naive approach

$ grep lawyer corpus.txt | grep Penang

which was a Penang-lawyer weighted with lead, was just such a weapon is known as a "Penang lawyer." Just under the head was a broad silver
Naive approach

Q: lawyer AND Penang AND NOT silver
Naive approach

$ grep lawyer corpus.txt | grep Penang | grep -v silver
Naive approach

$ grep lawyer corpus.txt | grep Penang | grep -v silver

which was a Penang-lawyer weighted with lead, was just such a weapon
Poll

Go *now* to:

https://eduapp-app1.ethz.ch/

or install EduApp 2.x
Grepping

Regular expressions

```
foo|bar[a-z]+  
```

Wildcards

`*`
Grepping: shortcomings (1)
Grepping: shortcomings (1)

What if we want to search an extremely large collection: The Web
Grepping: shortcomings

To do it right, he told Page, you'd really have to capture a significant chunk of the World Wide Web's link structure.

Steven Levy
In The Plex: How Google Thinks, Works, and Shapes Our Lives
Grepping: shortcomings

To do it right, he told Page, you'd really have to capture a significant chunk of the World Wide Web's link structure.

Page said, sure, he'd go and download the web and get the structure.

Steven Levy
In The Plex: How Google Thinks, Works, and Shapes Our Lives
Grepping: shortcomings

To do it right, he told Page, you'd really have to capture a significant chunk of the World Wide Web's link structure.

Page said, sure, he'd go and download the web and get the structure.

He figured it would take a week or something.

Steven Levy
In The Plex: How Google Thinks, Works, and Shapes Our Lives
Grepping: shortcomings

To do it right, he told Page, you'd really have to capture a significant chunk of the World Wide Web's link structure.

Page said, sure, he'd go and download the web and get the structure.

He figured it would take a week or something.

“And of course,” he later recalled, “it took, like, years.”

Steven Levy
In The Plex: How Google Thinks, Works, and Shapes Our Lives
Grepping: shortcomings (2)
Grepping: shortcomings (2)

Proximity search?

Zurich NEAR Switzerland
Grepping: shortcomings (3)
Grepping: shortcomings (3)

How do we rank?
Usefulness of indices

Indices are everywhere

Look at the last pages of books
Data model
Document
Document

Documents
Term

Sherlock
lawyer
Switzerland
Unterwalden nid dem Wald
ETH Zürich
person
watch
run
paper
book
...

...
Relationship between documents and terms?

I han es Zündhölzli azündt
Und das het e Flamme gäh
Und i ha für d'Zigarette
Welle Füür vom Hölzli näh
Aber ds Hölzli isch dervo-
Gspickt und uf e Teppich cho
Und es hätt no fasch es Loch i Teppich gäh dervo
Inclusion

I han es Zündhölzli azündt
Und das het e Flamme gäh
Und i ha für d'Zigarette
Welle Füür vom Hölzli näh
Aber ds Hölzli isch dervo-
Gspickt und uf e Teppich cho
Und es hätt no fasch es Loch i Teppich gäh dervo

A term is included in a document

"Hölzli" ∈ D
Occurrence

I han es Zündhölzli azündt
Und das het e Flamme gäh
Und i ha für d'Zigarette
Welle Füür vom Hölzli näh
Aber ds Hölzli isch dervo-
Gspickt und uf e Teppich cho
Und es hätt no fasch es Loch i Teppich gäh dervo

A term can occur $x$ times in a document

#("Hölzli", D) = 2
Order

I han es Zündhölzli azündt
Und das het e **Flamme** gäh
Und i ha für i Zigarette
Welle **Füür** vom Hölzli näh
Aber ds Hölzli isch dervo-
Gspickt und uf e Teppich cho
Und es hätt no fasch es Loch i Teppich gäh dervo

A term can occur **x times** in a document

"Flamme" < "Füür"
I han es Zündhölzli azündt
Und das het e Flamme gäh
Und i ha für d'Zigarette
Welle Füür vom Hölzli näh
Aber ds Hölzli isch dervo-
Gspickt und uf e Teppich cho
Und es hätt no fasch es Loch i Teppich gäh dervo
Document as bag of terms
Document as set of terms
Incidence matrix
Set-of-terms model

Documents contain (or not) Terms
Incidence Matrix

Documents

1 contain (or not) 0

Terms
Incidence Matrix

\[
\text{contains}_{i,j} = \begin{cases} 
1 & \text{if document } j \text{ contains term } i \\
0 & \text{if document } j \text{ does not contain term } i
\end{cases}
\]
### Incidence Matrix

<table>
<thead>
<tr>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
</tr>
<tr>
<td>u</td>
</tr>
<tr>
<td>v</td>
</tr>
<tr>
<td>w</td>
</tr>
<tr>
<td>x</td>
</tr>
<tr>
<td>y</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 0 1 1 1 0 0 0 1 1</td>
</tr>
<tr>
<td>0 0 1 0 1 1 1 1 1 0 0</td>
</tr>
<tr>
<td>0 1 1 1 0 1 0 1 0 1</td>
</tr>
<tr>
<td>0 0 0 1 1 0 0 0 0 0</td>
</tr>
<tr>
<td>1 0 1 1 1 0 1 0 0 1</td>
</tr>
<tr>
<td>0 0 0 0 1 0 0 1 0 1</td>
</tr>
</tbody>
</table>

Ethz Zürich

Page 55
### Incidence Matrix

<table>
<thead>
<tr>
<th>Terms</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>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>u</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>v</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>w</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>x</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>y</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Contains $y, 5 = 1$
## Incidence Matrix

<table>
<thead>
<tr>
<th>Terms</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>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>u</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>w</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$\text{contains} \in \mathbb{B}^{\text{Terms} \times \text{Documents}}$
Poll

Go *now* to:

https://eduapp-app1.ethz.ch/

or install EduApp 2.x
Boolean Query
### Incidence Matrix

<table>
<thead>
<tr>
<th>Terms</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>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>u</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>v</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>w</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>x</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>y</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
## Incidence Matrix

<table>
<thead>
<tr>
<th>Terms</th>
<th>Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>1 0 1 1 1 1 0 0 0 1 1</td>
</tr>
<tr>
<td>u</td>
<td>0 0 1 0 1 1 1 1 1 0 0</td>
</tr>
<tr>
<td>v</td>
<td>0 1 1 1 1 0 1 0 1 0 1</td>
</tr>
<tr>
<td>w</td>
<td>0 0 0 1 1 0 0 0 0 0 0</td>
</tr>
<tr>
<td>x</td>
<td>1 0 1 1 1 0 1 0 0 1 1</td>
</tr>
<tr>
<td>y</td>
<td>0 0 0 0 1 0 0 1 0 1 1</td>
</tr>
</tbody>
</table>

The matrix represents the incidence of terms in documents. Each row corresponds to a term, and each column corresponds to a document.
### Incidence Matrix

![Incidence Matrix Diagram](image)
Boolean Query

NOT u
## Incidence Matrix

<table>
<thead>
<tr>
<th></th>
<th>t</th>
<th>u</th>
<th>v</th>
<th>w</th>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

**Documents**
## Incidence Matrix

<table>
<thead>
<tr>
<th></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>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>u</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

\[
(0 \ 0 \ 1 \ 0 \ 1 \ 1 \ 1 \ 1 \ 1 \ 0 \ 0)
\]
Incidence Matrix

\[
\begin{pmatrix}
0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1
\end{pmatrix}
\]
## Incidence Matrix

<table>
<thead>
<tr>
<th>Documents</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>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 NOT u</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1 1</td>
</tr>
</tbody>
</table>
Incidence Matrix

\[
\begin{pmatrix}
1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\
\end{pmatrix}
\]

NOT u

Documents

1 2 3 4 5 6 7 8 9 10
Boolean Query

u AND x
## Incidence Matrix

<table>
<thead>
<tr>
<th>Terms</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>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>u</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>v</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>w</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>x</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>y</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
## Incidence Matrix

<table>
<thead>
<tr>
<th>Terms</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>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>u</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>x</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Incidence Matrix

Terms

\[ u \]

\[
\begin{pmatrix}
0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0
\end{pmatrix}
\]

\[ x \]

\[
\begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1
\end{pmatrix}
\]

\[ u \text{ AND } x \]

\[
\begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0
\end{pmatrix}
\]
Incidence Matrix

\[
\begin{pmatrix}
0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0
\end{pmatrix}
\]
Boolean Query

u OR x
Incidence Matrix

\[
\begin{pmatrix}
\text{Terms} \\
\begin{array}{cccccccccccc}
\text{t} & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 \\
\text{u} & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\
\text{v} & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\
\text{w} & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{x} & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\
\text{y} & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \\
\end{array}
\end{pmatrix}
\]
Incidence Matrix

<table>
<thead>
<tr>
<th>Terms</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>10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\begin{pmatrix}
0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\
\end{pmatrix}
\]

\[
\begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1
\end{pmatrix}
\]
## Incidence Matrix

<table>
<thead>
<tr>
<th></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>10</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>u</strong></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><strong>x</strong></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

\[ (0 \ 0 \ 1 \ 0 \ 1 \ 1 \ 1 \ 1 \ 1 \ 0 \ 0) \]

\[ (0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 1 \ 0 \ 1) \]

\[ (0 \ 0 \ 1 \ 0 \ 1 \ 1 \ 1 \ 1 \ 1 \ 0 \ 1) \]

**OR**
Incidence Matrix

\[ u \text{ AND } x \]

\[
\begin{pmatrix}
0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 1
\end{pmatrix}
\]
Poll

Go now to:

https://eduapp-app1.ethz.ch/

or install EduApp 2.x
Effectiveness of an IR system
Information need

Impact of global warming on the development of butterfly ecosystems in Southern part of the Canton of Uri between 2011 and 2013
Information need

Impact of global warming on the development of butterfly ecosystems in Southern part of the Canton of Uri between 2011 and 2013

$ butterfly AND climate AND uri AND south

$ lepidoptera AND ur AND NOT winter

$ (central NEAR (switzerland OR sweden)) AND NOT (lu OR nw OR ow OR stockholm) AND warmth
Results
Results
Results

Returned results

Relevant

Not relevant
Results

Precision = \frac{\text{Relevant}}{\text{Not relevant}}
Results

\[
\text{Precision} = \frac{\text{true positives}}{\text{returned}}
\]
Results

Returned results

Relevant

Not relevant

Precision =
Results

Precision = 50%
Results

Precision =
Results

Precision = 100%
Results

Returned results

Relevant

Not relevant

Precision =
Results

**Precision = 0%**
Results

Precision =
Results

[Diagram showing returned results with 4 relevant and 2 not relevant] 

Precision = 80%
Results

Positives

Relevant

Not relevant
Results

### Relevant

- Positives
- Negatives

### Not relevant

- Positives
- Negatives
Results

Relevant

Positives

Negatives

Recall =
Results

Recall = \frac{\text{true positives}}{\text{relevant}}
Results

Recall = 50%
Results

Relevant

Positives

Negatives

Recall = 100%
Results

Relevant

Positives

Negatives

Recall = 0%
Results

Recall = 67%
Type I vs. Type II error

- **Positives**
  - Relevant
  - Type I error: (false positives)
  - Not relevant

- **Negatives**
  - Type II error: (false negatives)
  - Relevant
  - Not relevant

Type I error: (false positives)
Type II error: (false negatives)
High precision

Relevant

Positives

Negatives

Not relevant
High recall

Positives

Negatives

Relevant

Not relevant
Compromise

High precision

High recall
Incidence Matrix

\[
\begin{pmatrix}
1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\
0 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1
\end{pmatrix}
\]
### Incidence Matrix

<table>
<thead>
<tr>
<th>Terms</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>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>u</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>w</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Typically 500,000</td>
</tr>
<tr>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Typically 500,000
Incidence Matrix

Typically 500,000

Typically one million
Incidence Matrix

Documents

Typically one million

Terms

Typically 500,000

500 billion booleans!
Incidence Matrix

Documents

1 2 3 4 5 6 7 8 9 10

t u v w

Typically 500,000

Typically one million

Terms

t u v w x y

500 billion booleans!

How many zeros?
Incidence Matrix

Documents

1 2 3 4 5 6 7 8 9 10

Typically 1,000 terms per document

Typically one million

Terms

t u v w x y

Typically 500,000

How many zeros?
Incidence Matrix

Typically 500,000

Typically one million

How many zeros?
Incidence Matrix

Documents

1 2 3 4 5 6 7 8 9 10

t u v w x y

Typically 500,000

Typically one million

Terms

How many zeros?

one billion 1s
499 billion 0s
Incidence Matrix

- Typically 500,000
- Typically one million
- 99.8% empty!
### Incidence Matrix

#### Terms
- t
- u
- v
- w
- x
- y

#### Documents
- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10

Typically 500,000

Typically one million

**99.8% empty!**

**This is space-inefficient.**
Can we store this in a better way?

<table>
<thead>
<tr>
<th>Terms</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>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1 1</td>
</tr>
<tr>
<td>u</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0 0</td>
</tr>
<tr>
<td>v</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1 1</td>
</tr>
<tr>
<td>w</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0 0</td>
</tr>
<tr>
<td>x</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0 0</td>
</tr>
<tr>
<td>y</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1 1</td>
</tr>
</tbody>
</table>
Poll

Go now to:

https://eduapp-app1.ethz.ch/

or install EduApp 2.x
Can we store this in a better way?

<table>
<thead>
<tr>
<th>Terms</th>
<th>t</th>
<th>u</th>
<th>v</th>
<th>w</th>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Documents

1 2 3 4 5 6 7 8 9 10
Can we store this in a better way?

<table>
<thead>
<tr>
<th>Terms</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>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>u</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>v</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>w</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>x</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>

Documents:

- t: 1, 4, 9, 10
- u: 5, 6, 7
- v: 2, 4, 6, 8, 10
- w: 5
- x: 1, 3, 4, 7
- y: 5, 8, 10
### Inverted index (inverted file)

<table>
<thead>
<tr>
<th>Terms</th>
<th>1, 4, 9, 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td></td>
</tr>
<tr>
<td>u</td>
<td>5, 6, 7</td>
</tr>
<tr>
<td>v</td>
<td>2, 4, 6, 8, 10</td>
</tr>
<tr>
<td>w</td>
<td>5</td>
</tr>
<tr>
<td>x</td>
<td>1, 3, 4, 7</td>
</tr>
<tr>
<td>y</td>
<td>5, 8, 10</td>
</tr>
</tbody>
</table>
Inverted index (inverted file)

<table>
<thead>
<tr>
<th>Terms</th>
<th>Vocabulary</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>1, 4, 9, 10</td>
</tr>
<tr>
<td>u</td>
<td>5, 6, 7</td>
</tr>
<tr>
<td>v</td>
<td>2, 4, 6, 8, 10</td>
</tr>
<tr>
<td>w</td>
<td>5</td>
</tr>
<tr>
<td>x</td>
<td>1, 3, 4, 7</td>
</tr>
<tr>
<td>y</td>
<td>5, 8, 10</td>
</tr>
</tbody>
</table>
## Inverted index (inverted file)

<table>
<thead>
<tr>
<th>Terms</th>
<th>Posting</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t$</td>
<td>1, 4, 9, 10</td>
</tr>
<tr>
<td>$u$</td>
<td>5, 6, 7</td>
</tr>
<tr>
<td>$v$</td>
<td>2, 4, 6, 8, 10</td>
</tr>
<tr>
<td>$w$</td>
<td>5</td>
</tr>
<tr>
<td>$x$</td>
<td>1, 3, 4, 7</td>
</tr>
<tr>
<td>$y$</td>
<td>5, 8, 10</td>
</tr>
</tbody>
</table>
Inverted index (inverted file)

<table>
<thead>
<tr>
<th>Terms</th>
<th>Posting</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>1, 4, 9, 10</td>
</tr>
<tr>
<td>u</td>
<td>5, 6, 7</td>
</tr>
<tr>
<td>v</td>
<td>2, 4, 6, 8, 10</td>
</tr>
<tr>
<td>w</td>
<td>5</td>
</tr>
<tr>
<td>x</td>
<td>1, 3, 4, 7</td>
</tr>
<tr>
<td>y</td>
<td>5, 8, 10</td>
</tr>
</tbody>
</table>

Assumption:
Each document has a Document ID (docID)
Inverted index (inverted file)

<table>
<thead>
<tr>
<th>Terms</th>
<th>Postings list</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>1, 4, 9, 10</td>
</tr>
<tr>
<td>u</td>
<td>5, 6, 7</td>
</tr>
<tr>
<td>v</td>
<td>2, 4, 6, 8, 10</td>
</tr>
<tr>
<td>w</td>
<td>5</td>
</tr>
<tr>
<td>x</td>
<td>1, 3, 4, 7</td>
</tr>
<tr>
<td>y</td>
<td>5, 8, 10</td>
</tr>
</tbody>
</table>
### Inverted index (inverted file)

<table>
<thead>
<tr>
<th>Terms</th>
<th>Postings</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>1, 4, 9, 10</td>
</tr>
<tr>
<td>u</td>
<td>5, 6, 7</td>
</tr>
<tr>
<td>v</td>
<td>2, 4, 6, 8, 10</td>
</tr>
<tr>
<td>w</td>
<td>5</td>
</tr>
<tr>
<td>x</td>
<td>1, 3, 4, 7</td>
</tr>
<tr>
<td>y</td>
<td>5, 8, 10</td>
</tr>
</tbody>
</table>
Inverted index (inverted file)

<table>
<thead>
<tr>
<th>Terms</th>
<th>Dictionary</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>1, 4, 9, 10</td>
</tr>
<tr>
<td>u</td>
<td>5, 6, 7</td>
</tr>
<tr>
<td>v</td>
<td>2, 4, 6, 8, 10</td>
</tr>
<tr>
<td>w</td>
<td>5</td>
</tr>
<tr>
<td>x</td>
<td>1, 3, 4, 7</td>
</tr>
<tr>
<td>y</td>
<td>5, 8, 10</td>
</tr>
</tbody>
</table>
Document frequency

<table>
<thead>
<tr>
<th>Terms</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>4</td>
</tr>
<tr>
<td>u</td>
<td>3</td>
</tr>
<tr>
<td>v</td>
<td>5</td>
</tr>
<tr>
<td>w</td>
<td>1</td>
</tr>
<tr>
<td>x</td>
<td>4</td>
</tr>
<tr>
<td>y</td>
<td>3</td>
</tr>
</tbody>
</table>

- t: 1, 4, 9, 10
- u: 5, 6, 7
- v: 2, 4, 6, 8, 10
- w: 5
- x: 1, 3, 4, 7
- y: 5, 8, 10
Document frequency

<table>
<thead>
<tr>
<th>Terms</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>1, 4, 9, 10</td>
</tr>
<tr>
<td>u</td>
<td>5, 6, 7</td>
</tr>
<tr>
<td>v</td>
<td>2, 4, 6, 8, 10</td>
</tr>
<tr>
<td>w</td>
<td>5</td>
</tr>
<tr>
<td>x</td>
<td>1, 3, 4, 7</td>
</tr>
<tr>
<td>y</td>
<td>5, 8, 10</td>
</tr>
</tbody>
</table>

4

3

2

Most frequent term
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!
<table>
<thead>
<tr>
<th>You</th>
<th>come</th>
<th>most</th>
<th>carefully</th>
<th>upon</th>
<th>your</th>
<th>hour</th>
</tr>
</thead>
<tbody>
<tr>
<td>Take</td>
<td>fair</td>
<td>thy</td>
<td>hour</td>
<td>Laertes</td>
<td>time</td>
<td>be</td>
</tr>
<tr>
<td>My</td>
<td>hour</td>
<td>is</td>
<td>almost</td>
<td>come</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Possess</td>
<td>it</td>
<td>merely</td>
<td>That</td>
<td>it</td>
<td>should</td>
<td>come</td>
</tr>
</tbody>
</table>
Linguistic pre-processing

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
## Assign document IDs

<table>
<thead>
<tr>
<th>Word</th>
<th>ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>you</td>
<td>1</td>
</tr>
<tr>
<td>come</td>
<td>1</td>
</tr>
<tr>
<td>most</td>
<td>1</td>
</tr>
<tr>
<td>carefully</td>
<td>1</td>
</tr>
<tr>
<td>upon</td>
<td>1</td>
</tr>
<tr>
<td>your</td>
<td>1</td>
</tr>
<tr>
<td>hour</td>
<td>1</td>
</tr>
<tr>
<td>take</td>
<td>2</td>
</tr>
<tr>
<td>fair</td>
<td>2</td>
</tr>
<tr>
<td>thy</td>
<td>2</td>
</tr>
<tr>
<td>hour</td>
<td>2</td>
</tr>
<tr>
<td>Laertes</td>
<td>2</td>
</tr>
<tr>
<td>time</td>
<td>2</td>
</tr>
<tr>
<td>be</td>
<td>2</td>
</tr>
<tr>
<td>thine</td>
<td>2</td>
</tr>
<tr>
<td>my</td>
<td>3</td>
</tr>
<tr>
<td>hour</td>
<td>3</td>
</tr>
<tr>
<td>is</td>
<td>3</td>
</tr>
<tr>
<td>almost</td>
<td>3</td>
</tr>
<tr>
<td>come</td>
<td>3</td>
</tr>
<tr>
<td>possess</td>
<td>4</td>
</tr>
<tr>
<td>it</td>
<td>4</td>
</tr>
<tr>
<td>merely</td>
<td>4</td>
</tr>
<tr>
<td>that</td>
<td>4</td>
</tr>
<tr>
<td>it</td>
<td>4</td>
</tr>
<tr>
<td>should</td>
<td>4</td>
</tr>
<tr>
<td>come</td>
<td>4</td>
</tr>
<tr>
<td>to</td>
<td>4</td>
</tr>
<tr>
<td>this</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>Sort</td>
<td></td>
</tr>
<tr>
<td></td>
<td>almost</td>
</tr>
<tr>
<td></td>
<td>come</td>
</tr>
<tr>
<td></td>
<td>fair</td>
</tr>
<tr>
<td></td>
<td>hour</td>
</tr>
<tr>
<td></td>
<td>is</td>
</tr>
<tr>
<td></td>
<td>Laertes</td>
</tr>
<tr>
<td></td>
<td>my</td>
</tr>
<tr>
<td></td>
<td>take</td>
</tr>
<tr>
<td></td>
<td>this</td>
</tr>
<tr>
<td></td>
<td>to</td>
</tr>
<tr>
<td></td>
<td>your</td>
</tr>
</tbody>
</table>
Merge

almost: 3
be: 2
carefully: 1
come: 1 3 4

is: 3

Laertes: 2
merely: 4
most: 1
my: 3
possess: 4
should: 4

take: 2
that: 4
thine: 2
this: 4
thy: 2
time: 2
to: 4
upon: 1
you: 1
your: 1
merely
most
my
possess
should
take
that

Laertes

thine
this
thy
time
to
upon
you
your

almost
carefully
come
fair
hour
is
it
Add document frequency

<table>
<thead>
<tr>
<th>almost</th>
<th>be</th>
<th>carefully</th>
<th>come</th>
<th>fair</th>
<th>hour</th>
<th>is</th>
<th>it</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
<td>2</td>
<td></td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Laertes</th>
<th>merely</th>
<th>most</th>
<th>my</th>
<th>possess</th>
<th>should</th>
<th>take</th>
<th>that</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>thine</th>
<th>this</th>
<th>thy</th>
<th>time</th>
<th>to</th>
<th>upon</th>
<th>you</th>
<th>your</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
Single word query

hour
Single word query

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
Single word query

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  

Single word query

Results

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

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

**Document 3**
My **hour** is almost come,
Boolean (conjunctive) query

hour AND come
Single word query

- almost
- be
- carefully
- come
- fair
- hour
- is
- it
- Laertes
- merely
- most
- my
- possess
- should
- take
- that
- thine
- this
- thy
- time
- to
- upon
- you
- your
Single word query

almost  1  3
be  1  2
carefully  1  1
come  3  1  3  4
fair  1  2
hour  3  1  2  3
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
Single word query

come

hour
Single word query
Single word query

Results

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

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

---

**Intersect**

- come
  - 3
  - 1
  - 3
  - 4

- hour
  - 3
  - 1
  - 2
  - 3

- 1
  - 3
Boolean (conjunctive) query

hour OR this
Single word query
Single word query

almost 1 3
be 1 2
carefully 1 1
come 3 1 3 4
fair 1 2
hour 3 1 2 3
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
Single word query

hour

this
Single word query
Single word query
You come most carefully upon your hour.
Take thy fair hour, Laertes; time be thine,
My hour is almost come,
Possess it merely. That it should come to this!
Intersection algorithm

List A: 1 ➔ 2 ➔ 4 ➔ 5 ➔ 8 ➔ 9 ➔ 10 ➔ 12

List B: 1 ➔ 3 ➔ 4 ➔ 6 ➔ 7 ➔ 8 ➔ 11 ➔ 12
Intersection algorithm

Initialization:
two pointers
Intersection algorithm

List A

Initialization: two pointers

List B

Intersection of A and B
Intersection algorithm

List A

List B

Intersection of A and B
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
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
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
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
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
Intersection algorithm

List A

List B

Intersection of A and B
Intersection algorithm

List A

List B

Intersection of A and B

1 2 4 5 8 9 10 12

1 3 4 6 7 8 11 12

1 4
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
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
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
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
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
Intersection algorithm

List A

List B

Intersection of A and B
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
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
Intersection algorithm

List A

List B

Intersection of A and B
Union algorithm

List A

1 → 2 → 4 → 5 → 8 → 9 → 10 → 12

List B

1 → 3 → 4 → 6 → 7 → 8 → 11 → 12
Union algorithm

List A

Initialization: two pointers

List B
Union algorithm

List A

Initialization:
two pointers

List B

Union of A and B
Union algorithm

List A

1 → 2 → 4 → 5 → 8 → 9 → 10 → 12

List B

1 → 3 → 4 → 6 → 7 → 8 → 11 → 12

Union of A and B
Union algorithm

List A

1 → 2 → 4 → 5 → 8 → 9 → 10 → 12

List B

1 → 3 → 4 → 6 → 7 → 8 → 11 → 12

Union of A and B

1
Union algorithm

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

List B: 1 → 3 → 4 → 6 → 7 → 8 → 11 → 12

Union of A and B: 1
Union algorithm

List A

1 → 2 → 4 → 5 → 8 → 9 → 10 → 12

List B

1 → 3 → 4 → 6 → 7 → 8 → 11 → 12

Union of A and B

1 → 2
Union algorithm

List A

1 → 2 → 4 → 5 → 8 → 9 → 10 → 12

List B

1 → 3 → 4 → 6 → 7 → 8 → 11 → 12

Union of A and B

1 → 2
Union algorithm

List A
1 ➔ 2 ➔ 4 ➔ 5 ➔ 8 ➔ 9 ➔ 10 ➔ 12

List B
1 ➔ 3 ➔ 4 ➔ 6 ➔ 7 ➔ 8 ➔ 11 ➔ 12

Union of A and B
1 ➔ 2 ➔ 3
Union algorithm

List A

List B

Union of A and B
Union algorithm

List A

1 → 2 → 4 → 5 → 8 → 9 → 10 → 12

List B

1 → 3 → 4 → 6 → 7 → 8 → 11 → 12

Union of A and B

1 → 2 → 3 → 4
Union algorithm

List A
1 → 2 → 4 → 5 → 8 → 9 → 10 → 12

List B
1 → 3 → 4 → 6 → 7 → 8 → 11 → 12

Union of A and B
1 → 2 → 3 → 4 → 5
Union algorithm

List A

1 → 2 → 4 → 5 → 8 → 9 → 10 → 12

List B

1 → 3 → 4 → 6 → 7 → 8 → 11 → 12

Union of A and B

1 → 2 → 3 → 4 → 5 → 6
Union algorithm

List A
1 → 2 → 4 → 5 → 8 → 9 → 10 → 12

List B
1 → 3 → 4 → 6 → 7 → 8 → 11 → 12

Union of A and B
1 → 2 → 3 → 4 → 5 → 6 → 7
Union algorithm

List A

1 → 2 → 4 → 5 → 8 → 9 → 10 → 12

List B

1 → 3 → 4 → 6 → 7 → 8 → 11 → 12

Union of A and B

1 → 2 → 3 → 4 → 5 → 6 → 7 → 8
Union algorithm

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

List B: 1 → 3 → 4 → 6 → 7 → 8 → 11 → 12

Union of A and B: 1 → 2 → 3 → 4 → 5 → 6 → 7 → 9 → 8
Union algorithm

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

List B:
1 → 3 → 4 → 6 → 7 → 8 → 11 → 12

Union of A and B:
1 → 2 → 3 → 4 → 5 → 6 → 7 → 10 → 9 → 8
Union algorithm

List A
1 → 2 → 4 → 5 → 8 → 9 → 10 → 12

List B
1 → 3 → 4 → 6 → 7 → 8 → 11 → 12

Union of A and B
1 → 2 → 3 → 4 → 5 → 6 → 7
11 ← 10 ← 9 ← 8
Union algorithm

List A

List B

Union of A and B
Union algorithm

List A
1 → 2 → 4 → 5 → 8 → 9 → 10 → 12

List B
1 → 3 → 4 → 6 → 7 → 8 → 11 → 12

Union of A and B
1 → 2 → 3 → 4 → 5 → 6 → 7
12 ← 11 ← 10 ← 9 ← 8
Optimizing

eth AND zurich AND bibliopole

The inverted index has the document frequencies!
Optimizing

eth AND zurich AND bibliopole

10,000 docs 10 docs

100,000 docs
Optimizing

bibliopole AND eth AND zurich

10 docs

10,000 docs

100,000 docs

Sort query terms by increasing document frequency
Optimizing

Maintain intermediate result in memory

```
bibliopole AND eth AND zurich
```

10 docs

10,000 docs

100,000 docs

100,000 docs
Optimizing

bibliopole AND eth AND zurich
10 docs
10,000 docs
100,000 docs

Maintain intermediate result in memory
Optimizing

Maintain intermediate result in memory
More complex queries

(epf OR eth) AND zurich AND bibliopole

We can approximate the (estimated) size of the postings list by summing the two original lists
More complex queries

(epf OR eth) AND NOT zurich AND bibliopole

We can adapt the intersection algorithm to "virtually" walk down the negated list
This week's reading

Chapter 1

Boolean Retrieval