# webir

Clase 1

# webir

- Libertad Tansini
- libertad@fing.edu.uy [webir]
- Martes y Jueves 10:00 a 11:30
- Los monitoreos por zoom

# webir – Metodología y Evaluación

- **Teórico**
- **Proyecto**
  - Investigación sobre un tema relacionado con el curso
  - Diseño de una solución a un problema
  - Posible implementación de la solución
  - El tema deberá ser validado por el docente previamente
  - Escribir un informe y hacer una presentación oral (15 min)
- **Evaluación**
  - Asistencia a clases al menos 60%
  - Proyecto
    - Monitoreos
    - Informe sobre el proyecto realizado
    - Presentación oral del proyecto

# webir – Bibliografía

Manning, Raghavan and Schütze
“Introduction to Information Retrieval”

Cambridge University Press, 2008
http://nlp.stanford.edu/IR-book/

# webir - Temas

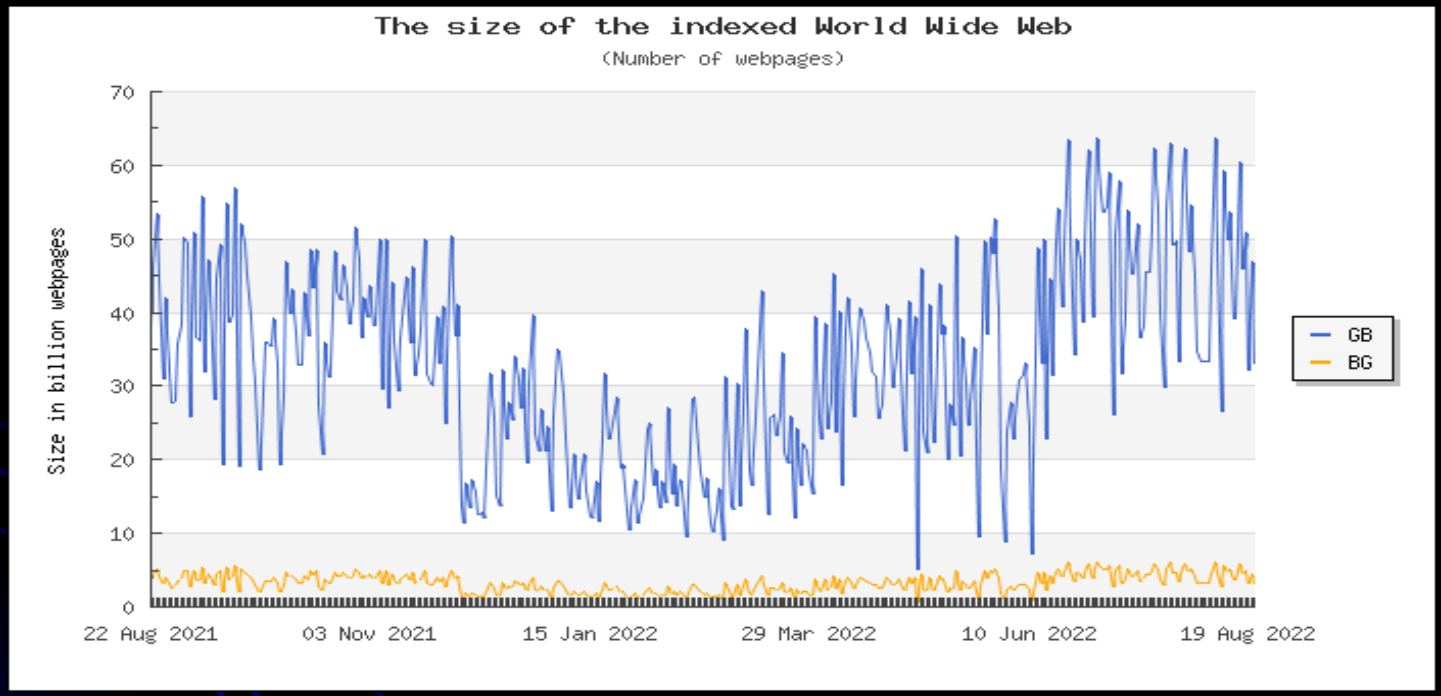
- Recuperación de información
  - Modelos
- Funcionamiento de los motores de búsqueda de (Internet)
- “Relevance feedback” y expansión de consultas
- Filtrado colaborativo
- Análisis de enlaces

# webir - Temas

- Clase
  - Recuperación de información
    - Usuario debe poder obtener de manera rápida la información que satisfaga sus necesidades
    - Gran colección de documentos - corpus
    - Documentos no estructurados
    - => Se debe poder encontrar, indexar y analizar la información
  - Motivación
  - Ejemplo de recuperación de documentos
  - Necesidad de información
  - Modelo de recuperación booleana
  - Índice invertido
  - Extensiones
  - Medidas de efectividad

# webir – Recuperación de Información

- Procesar la información que existe hoy en Internet - orden de miles de millones de palabras




- http://www.worldwidewebsize.com/
- **"The estimated minimal size of the indexed World Wide Web is based on the estimations of the numbers of pages indexed by (for example) Google and Bing"**

# webir – Recuperación de Información

- Ejemplo: corpus - obras de Shakespeare
- Se buscan obras de Shakespeare que tengan las palabras **Brutus y Caesar**, pero **no Calpurnia**
  - Brutus AND Cesar AND NOT Calpurnia
- El caso de las obras completas de Shakespeare - menos de 1 millón de palabras
- Usar comando **`grep`** de Linux
  - Busca coincidencias para una expresión regular de la entrada estándar o en una lista de archivos, e imprime las líneas que la contengan

# webir – Recuperación de Información

- A veces no es suficiente
- Colecciones o corpus **extensos** como Internet
- Operaciones de búsqueda más **complejas**
  - La palabra “Romans” cerca de “countrymen”
    - A lo más a 5 palabras de distancia o
    - Que aparezcan en la misma oración
- Devolver los documentos que contienen la información o la palabra ordenados por algún **criterio de utilidad/calidad**
  - ranking function

# webir – Matriz de palabras-documentos o de incidencia

| | 1 | 2 | 3 | 4 | 5 | 6 | |
|---|---|---|---|---|---|---|---|
| **Documentos**<br>**Términos** | Antony and Cleopatra | Julius Caesar | The Tempest | Hamlet | Othello | Macbeth | ... |
| Antony | 1 | 1 | 0 | 0 | 0 | 1 | |
| Brutus | 1 | 1 | 0 | 1 | 0 | 0 | |
| Caesar | 1 | 1 | 0 | 1 | 1 | 1 | |
| Calpurnia | 0 | 1 | 0 | 0 | 0 | 0 | |
| Cleopatra | 1 | 0 | 0 | 0 | 0 | 0 | |
| mercy | 1 | 0 | 1 | 1 | 1 | 1 | |
| worser | 1 | 0 | 1 | 1 | 1 | 0 | |
| ... | | | | | | | |

# webir – Matriz de palabras-documentos o de incidencia

| | 1 | 2 | 3 | 4 | 5 | 6 | |
|---|---|---|---|---|---|---|---|
| | Antony and Cleopatra | Julius Caesar | The Tempest | Hamlet | Othello | Macbeth | ... |
| Antony | 1 | 1 | 0 | 0 | 0 | 1 | |
| Brutus | 1 | 1 | 0 | 1 | 0 | 0 | |
| Caesar | 1 | 1 | 0 | 1 | 1 | 1 | |
| Calpurnia | 0 | 1 | 0 | 0 | 0 | 0 | |
| Cleopatra | 1 | 0 | 0 | 0 | 0 | 0 | |
| mercy | 1 | 0 | 1 | 1 | 1 | 1 | |
| worser | 1 | 0 | 1 | 1 | 1 | 0 | |
| ... | | | | | | | |

# webir – Matriz de palabras-documentos

| | 1 | 2 | 3 | 4 | 5 | 6 | |
|---|---|---|---|---|---|---|---|
| **Documentos**<br>**Términos** | Antony and Cleopatra | Julius Caesar | The Tempest | Hamlet | Othello | Macbeth | ... |
| Antony | 1 | 1 | 0 | 0 | 0 | 1 | |
| Brutus | 1 | 1 | 0 | 1 | 0 | 0 | |
| Caesar | 1 | 1 | 0 | 1 | 1 | 1 | |
| Calpurnia | 0 | 1 | 0 | 0 | 0 | 0 | |
| ... | | | | | | | |

- Operamos con los vectores para Brutus, Caesar y el complemento de Calpurnia (bit a bit)
- 110100 AND 110111 AND 101111 = 100100

# webir – Matriz de palabras-documentos

| | 1 | 2 | 3 | 4 | 5 | 6 | |
|---|---|---|---|---|---|---|---|
| **Documentos**<br><br>**Términos** | Antony and Cleopatra | Julius Caesar | The Tempest | Hamlet | Othello | Macbeth | ... |
| Antony | 1 | 1 | 0 | 0 | 0 | 1 | |
| Brutus | 1 | 1 | 0 | 1 | 0 | 0 | |
| Caesar | 1 | 1 | 0 | 1 | 1 | 1 | |
| Calpurnia | 0 | 1 | 0 | 0 | 0 | 0 | |
| ... | | | | | | | |

- Operamos con los vectores para Brutus, Caesar y el complemento de Calpurnia
- 110100 AND 110111 AND 101111 = 100100

# webir – Recuperación Booleana

- Modelo de recuperación de información
- Cualquier consulta se formula mediante una expresión booleana de palabras usando AND, OR y NOT
- Los documentos son conjuntos de palabras – “bag of words”

# webir – Indice Invertido

- ¿Es siempre factible/útil construir la matriz?

| | 1 | 2 | 3 | 4 | 5 | 6 | |
|---|---|---|---|---|---|---|---|
| | Antony and Cleopatra | Julius Caesar | The Tempest | Hamlet | Othello | Macbeth | ... |
| Antony | 1 | 1 | 0 | 0 | 0 | 1 | |
| Brutus | 1 | 1 | 0 | 1 | 0 | 0 | |
| Caesar | 1 | 1 | 0 | 1 | 1 | 1 | |
| Calpurnia | 0 | 1 | 0 | 0 | 0 | 0 | |
| Cleopatra | 1 | 0 | 0 | 0 | 0 | 0 | |
| ... | | | | | | | |

# webir – Indice Invertido

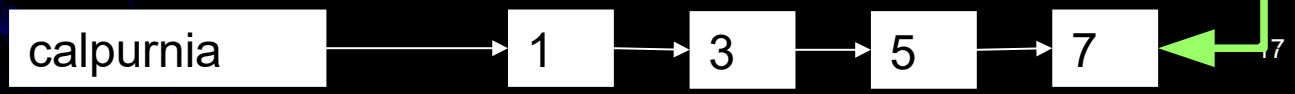
- ¿Es siempre factible/útil construir la matriz?
  - Matrices muy grandes
  - Muy pocos 1´s
  - Registrar sólo 1´s

# webir – Indice Invertido

- ¿Es siempre factible/útil construir la matriz?
  - Matrices muy grandes
  - Muy pocos 1´s
  - Registrar sólo 1´s

| Antony |
| --- |
| Brutus |
| Caesar |
| Calpurnia |
| Cleopatra |
| ... |

- Indice invertido
  - Diccionario de palabras (términos) - vocabulario
  - Lista de documentos donde aparece cada palabra (ev. también la posición) - post o posting
  - Se puede ordenar las listas por algún criterio, por ejemplo docID



# webir – Indice Invertido

- Primera aproximación a la construcción
  - Colección de documentos a indexar

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

  - Separar en palabras (tokenize)

Friends | Romans | countrymen | So | let

  - Procesamiento lingüístico para normalizar las palabras

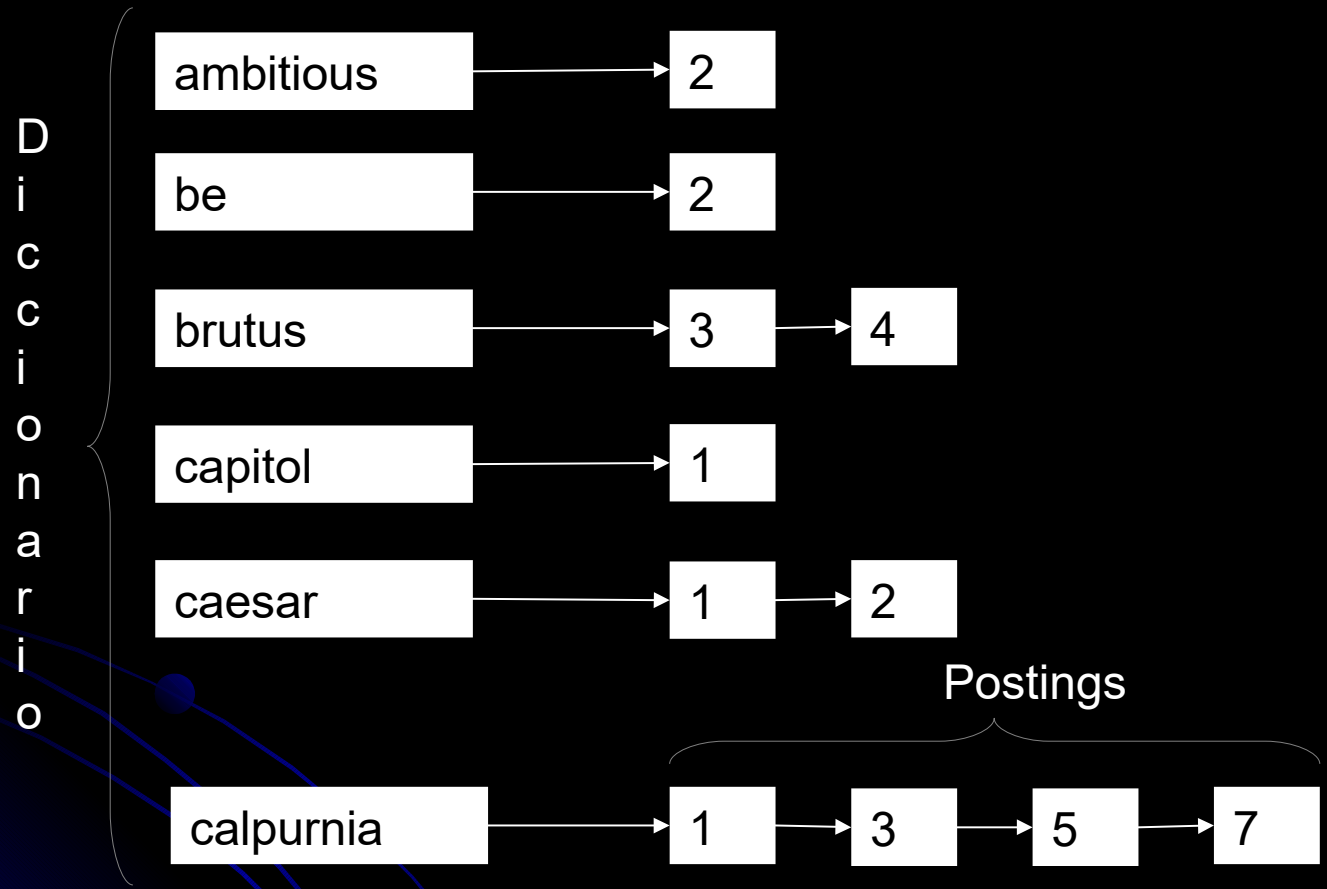
friend | roman | countryman | so | let

  - Crear el índice

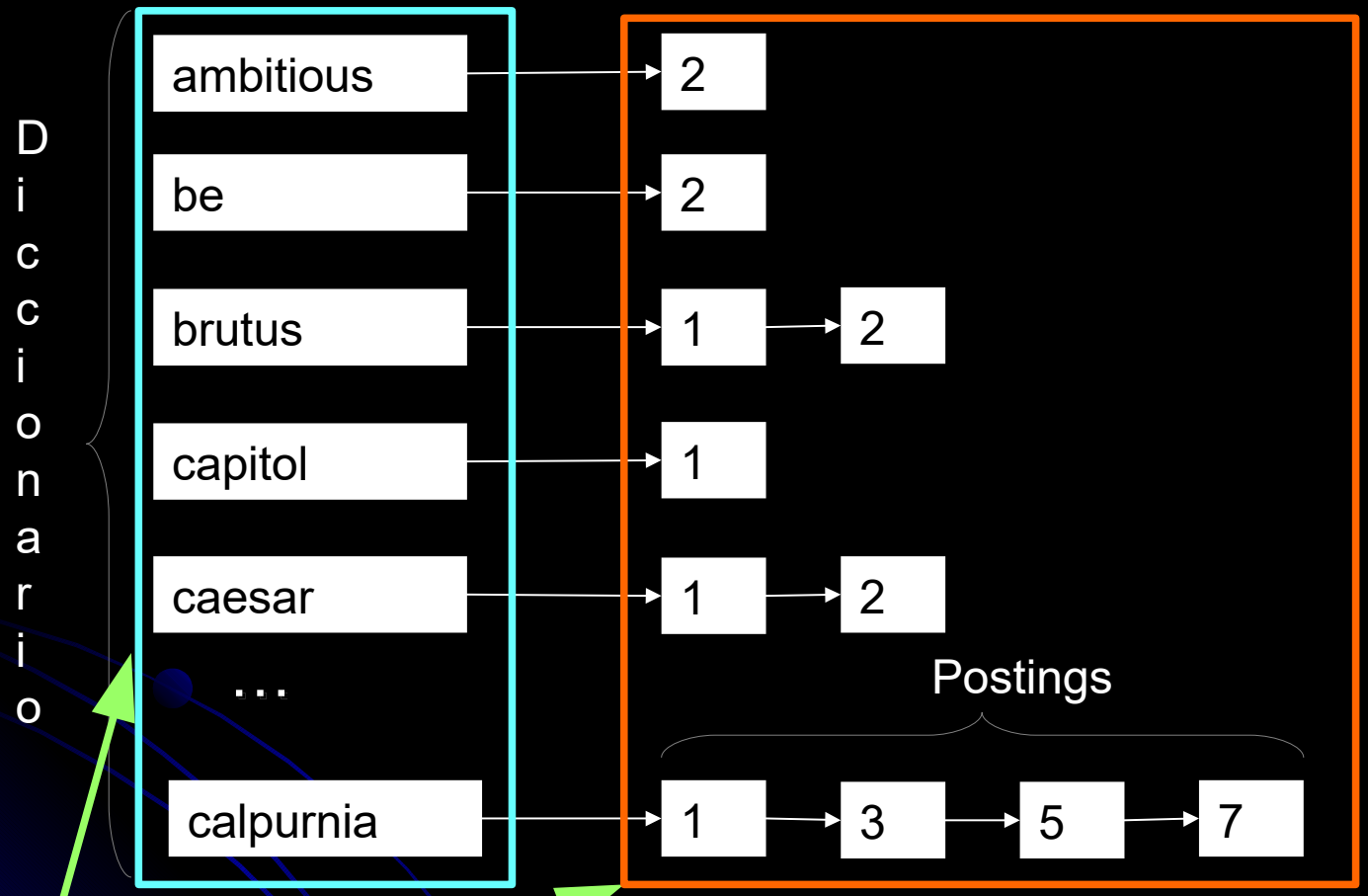
# webir – Indice Invertido

- Crear parejas de palabras y docID, (Calpurnia,1)
- Ordenar las parejas en orden alfabético por palabras
  - (Calpurnia,1),(Calpurnia,1), (Calpurnia,3), (Calpurnia,3), (Calpurnia,5), (Calpurnia,7), (Brutus,3), (Brutus,4)
- Unificar ocurrencias repetidas de palabras
  - (Calpurnia,1), (Calpurnia,3), (Calpurnia,5), (Calpurnia,7)
  - (Brutus,3), (Brutus,4)
- Agregar datos para mejorar la eficiencia
  - número total de documentos en que está cada palabra
- Ordenar postings por docID

# webir – Indice Invertido

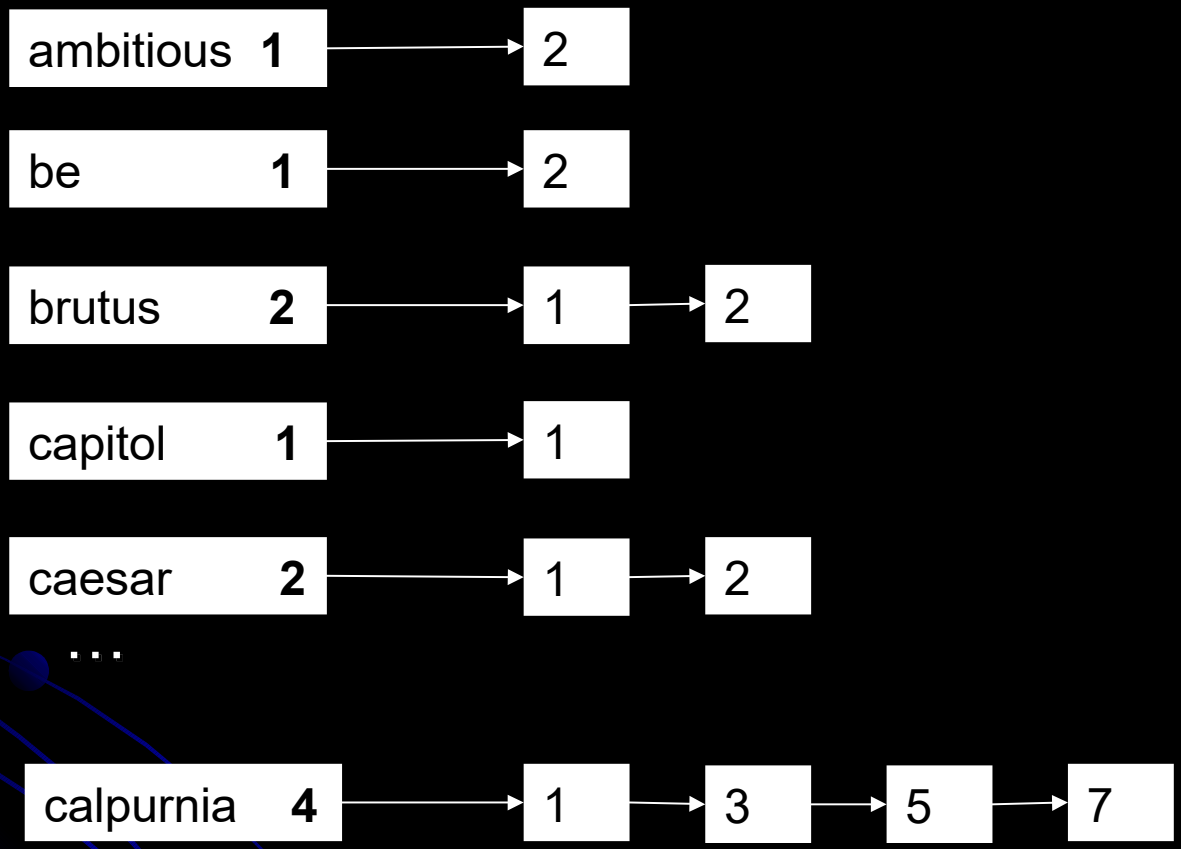


# webir – Indice Invertido




- Memoria vs disco – veremos estructuras de datos (cap. 5)
- Postings
  - Listas enlazadas o vectores de tamaño variable – frecuencia de modif.

# webir – Indice Invertido




- Procesamiento de las consultas – en general ”merging”
- Orden, primero los términos menos frecuentes - heurística
- Optimización

## Ejercicio

Dibujar el índice invertido para los siguientes documentos sin procesar ligüísticamente las palabras:

- Doc1: venta de casas en Montevideo
- Doc2: alquiler y venta de inmuebles
- Doc3: precios de casas y alquiler en aumento
- Doc4: Montevideo record de precios de inmuebles y alquiler

# webir – Indice Invertido

## Ejercicio

- Doc1: venta de casas en Montevideo
- Doc2: alquiler y venta de inmuebles
- Doc3: precios de casas y alquiler en aumento
- Doc4: Montevideo record de precios de inmuebles y alquiler

- alquiler: 2, 3, 4
- aumento: 3
- casas: 1, 3
- de: 1, 2, 3, 4
- en: 1, 3
- inmuebles: 2, 4
- Montevideo: 1, 4
- precios: 3, 4
- venta: 1, 2
- record: 4
- y: 2, 3, 4

## Ejercicio

- ¿Resultados?
- casas AND Montevideo
- inmuebles
- alquiler AND precios

- alquiler: 2, 3, 4
- aumento: 3
- casas: 1, 3
- de: 1, 2, 3, 4
- en: 1, 3
- inmuebles: 2, 4
- Montevideo: 1, 4
- precios: 3, 4
- venta: 1, 2
- record: 4
- y: 2, 3, 4

# webir – Indice Invertido

## Ejercicio

- ¿Resultados?
- casas AND Montevideo
- inmuebles
- alquiler AND precios

- ¿Orden?

- alquiler: 2, 3, 4
- aumento: 3
- casas: 1, 3
- de: 1, 2, 3, 4
- en: 1, 3
- inmuebles: 2, 4
- Montevideo: 1, 4
- precios: 3, 4
- venta: 1, 2
- record: 4
- y: 2, 3, 4

# webir – Recuperación Booleana Extendida

- ¿Medidas de cercanía?
  - a lo más a 3 palabras de distancia
  - que aparezcan en la misma oración o párrafo

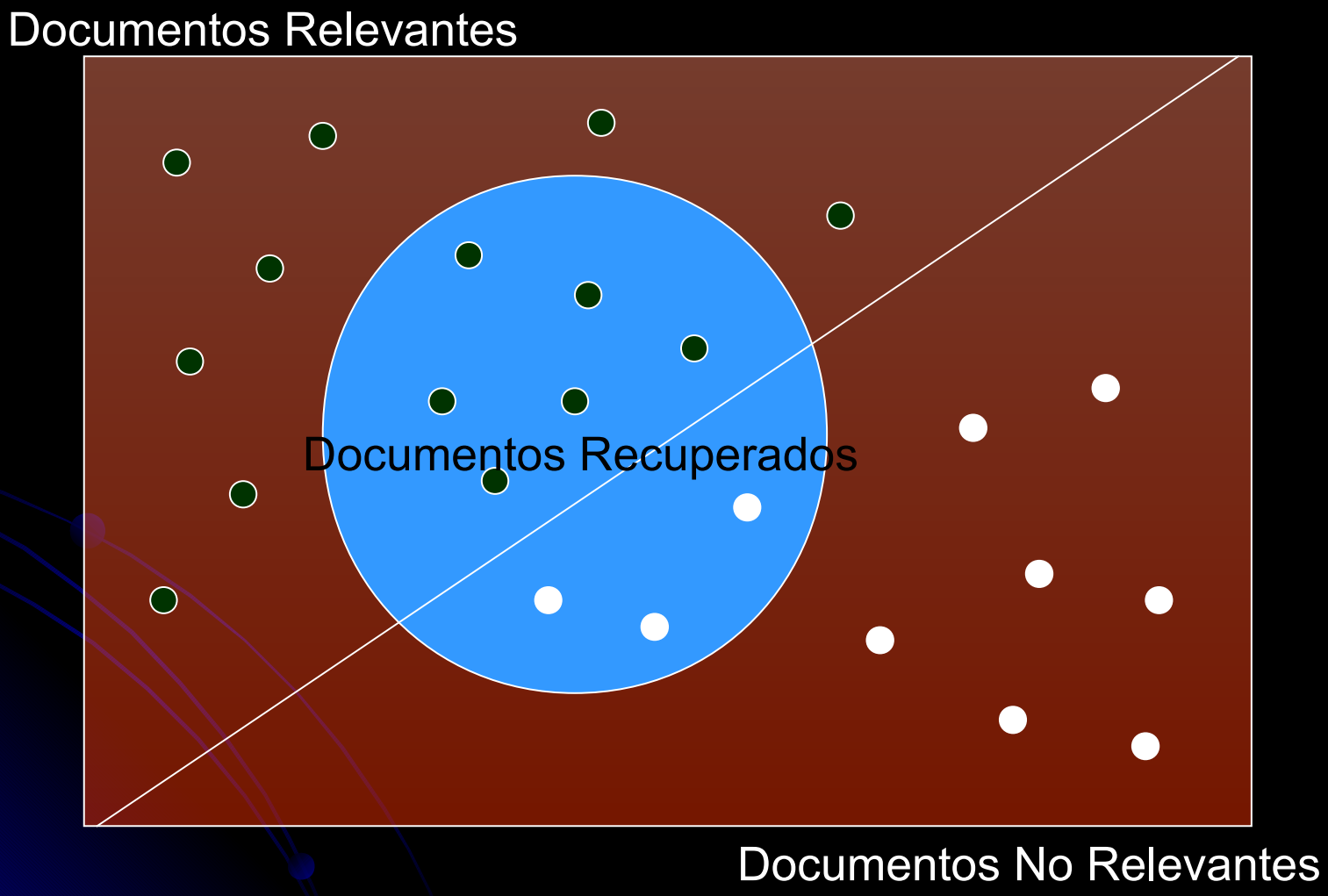
- Westlaw – http://www.westlaw.com (1975)

- *Information need:* Information on the legal theories involved in preventing the disclosure of trade secrets by employees formerly employed by a competing company.
  - *Query:* "trade secret" /s disclos! /s prevent /s employe!
- *Information need:* Requirements for disabled people to be able to access a workplace.
  - *Query:* disab! /p access! /s work-site work-place (employment /3 place)
- *Information need:* Cases about a host's responsibility for drunk guests.
  - *Query:* host! /p (responsib! liab!) /p (intoxicat! drunk!) /p guest

# webir – Extensiones Deseables – Otros Modelos

- Recuperación tolerante a errores de ortografía y otras inconsistencias
- Búsqueda de conceptos, por ej. "`sistema operativo`"
- Medidas de cercanía, por ej. `Windows` cerca de `Microsoft` o de `Gates`
- Registrar y considerar la cantidad de veces que aparecen las palabras en los documentos
  - term frequency
- Devolver los documentos ordenados por algún criterio de utilidad/calidad
  - ranking function

- ***Necesidad de información*** de un usuario no es lo mismo que la ***consulta***
- Un documento es ***relevante*** si contiene información adecuada para satisfacer su necesidad de información
- Medidas de efectividad


Documentos Relevantes
Documentos Recuperados
Documentos No Relevantes

- ***Precisión***

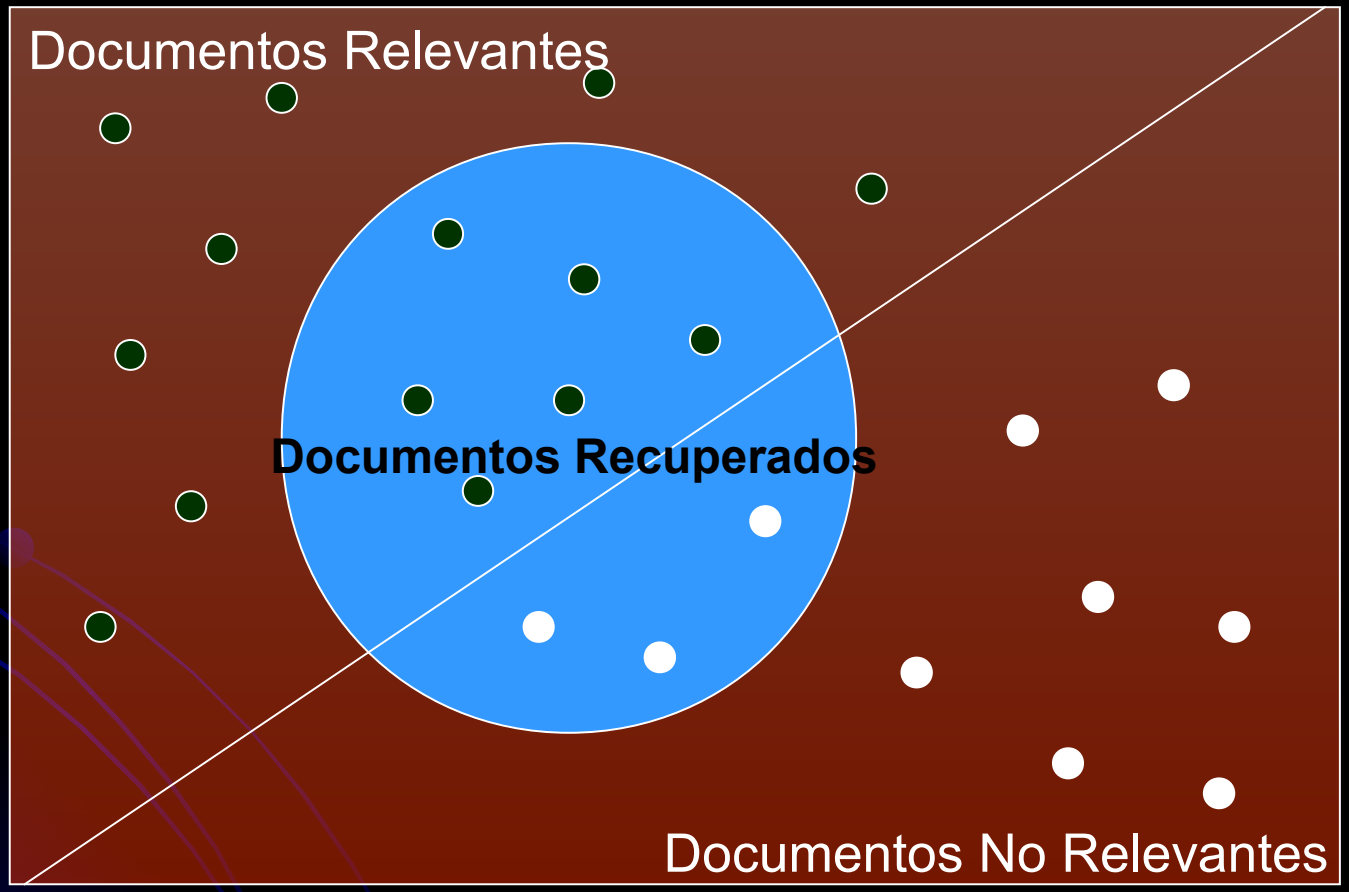
$$P = \frac{\text{Documentos relevantes recuperados}}{\text{Documentos } \mathbf{recuperados}}$$

- ***Exhaustividad en recuperación*** o ***Recall***

$$R = \frac{\text{Documentos relevantes recuperados}}{\text{Documentos } \mathbf{relevantes}}$$

- $F = 2*(P*R)/(P+R)$
- $F_\beta = (1+\beta^2)*(P*R)/(\beta^2 P+R)$

***Precisión* =** Docs relevantes recuperados/ Docs recuperados




***Recall* =** Docs relevantes recuperados/ Docs relevantes