webir

Clase 2

webir - Temas

- Repaso
- Normalización y stemming
- Optimización de la Lista Postings
- Consultas por Conceptos o Frases
  - Índices de Pares de Palabras
  - Índices Posicionales

- Ejemplo: se buscan obras de Shakespeare que tengan las palabras Brutus y Caesar, pero no Calpurnia
  - Brutus AND Cesar AND NOT Calpurnia
- En las obras completas de Shakespeare
- Modelo de Recuperación Booleana
  - Consultas - expresión booleana de palabras usando AND, OR y NOT
  - Los documentos son conjuntos de palabras

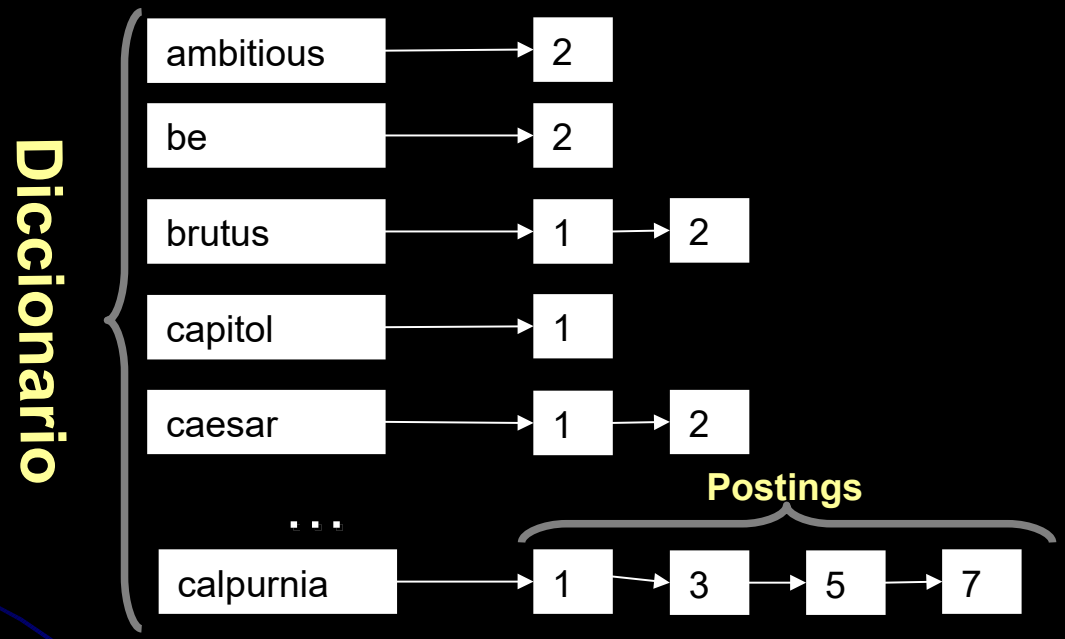
webir – Matriz de palabras-documentos

	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 – Indice Invertido

- No siempre es factible/útil construir la matriz



- Memoria vs disco – veremos estructuras de datos (cap. 5)
- Postings: Listas enlazadas o vectores de tamaño variable – frecuencia de modif.
- Procesamiento de las consultas – en general ”merging”
- Optimización, por ejemplo orden, primero los términos menos frecuentes - heurística

webir – Indice Invertido

Ejercicio

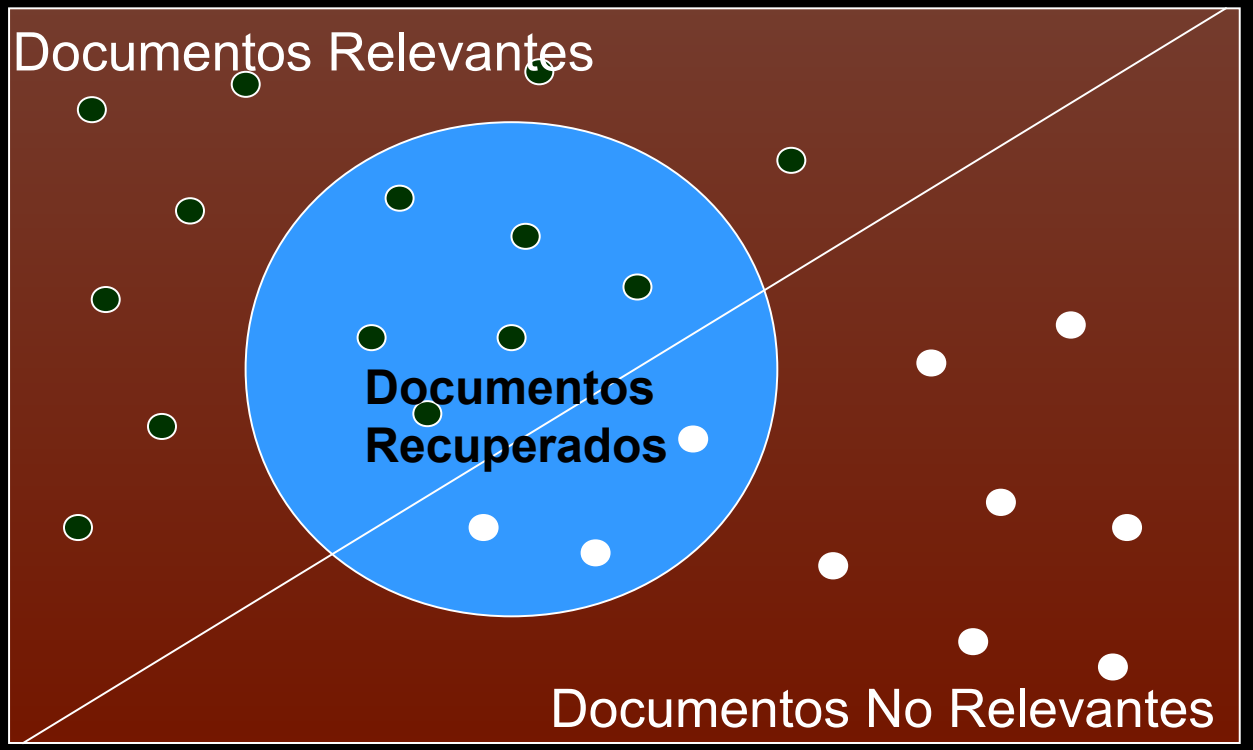
- Doc1: venta de casas en Montevideo
- Doc2: alquiler y venta de inmuebles
- Doc3: precios de casas en aumento
- Doc4: Montevideo record de precios de inmuebles

- alquiler: 2
- 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

webir – Sistemas de RI

- ***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
- Individuales
- Contexto
- Aproximación

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



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

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

webir – Extensiones Deseables – Otros Modelos

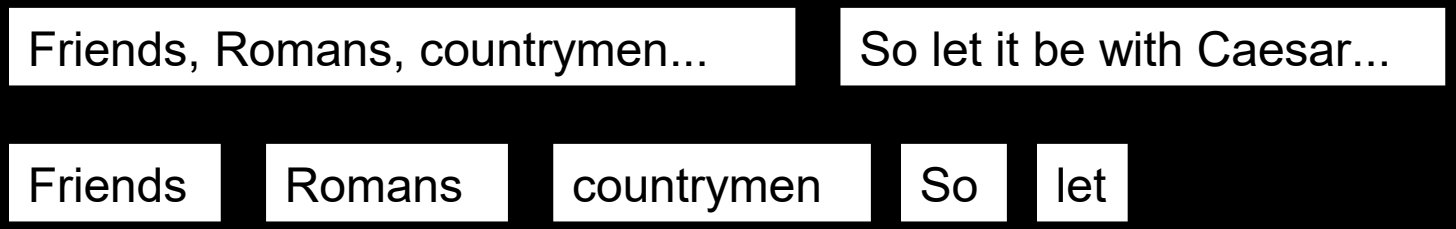
- Recuperación tolerante a errores de ortografía y otras inconsistencias
- Búsqueda de conceptos, por ej. "`sistema operativo`"
- Medidas de cercanía, por ej. `Windows` cerca de `Microsoft` o de `Gates`
- Registrar y considerar la cantidad de veces que aparecen las palabras en los documentos
  - term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

webir - Procesamiento Lingüístico para la Construcción del Indice Invertido

- Elección de la unidad “documento” a indexar
  - Páginas
  - E-mail
  - Libros o Capítulos
- Un buen sistema de RI debería ofrecer distintos niveles de granularidad
- Depende
  - Colección de documentos
  - Uso por parte de los usuarios

webir – Tokenization para la Construcción del Indice Invertido

- Separar en palabras (tokenize)



- Las palabras son unidades semánticas instanciadas que aparecen en los documentos
- Los términos son palabras (ev. normalizadas) que aparecen en el diccionario de un sistema de RI
  - Luego de diferentes procesos de normalización

webir – Tokenization para la Construcción del Indice Invertido

- Ignorar espacios en blanco y caracteres especiales
- ¿Cómo separar correctamente las palabras?

  aren't

  arent

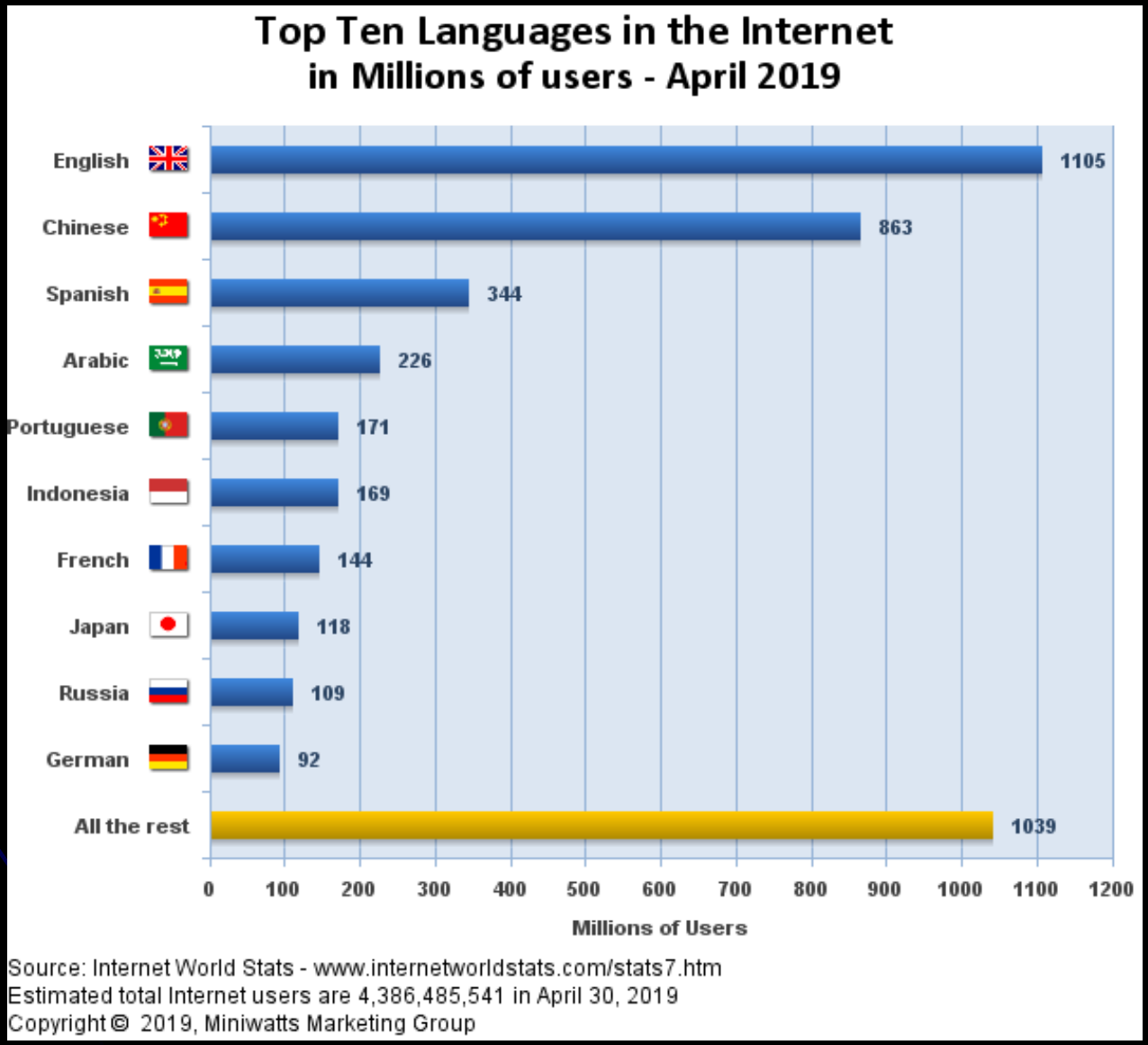
  are | n't

  aren | t

  - Sin embargo O'Higgins, O'Doole
- Depende del idioma
  - Identificación del idioma
  - Reglas, heurísticas

webir – Tokenization para la Construcción del Indice Invertido

- Identificación del idioma



webir – Tokenization para la Construcción del Indice Invertido

- Identificación del idioma
  - Problema de clasificación - Aprendizaje automático
  - Metadatos de los documentos
  - Heurísticas
    - Secuencias de letras características o palabras
- Varios idiomas en un texto

webir – Tokenization para la Construcción del Indice Invertido

- Palabras especiales – ev. dependen del idioma y/o dominio
  - C++ o C#
  - B-52
  - libertad@fing.edu.uy
  - http://www.fing.edu.uy/inco/cursos/webir/
  - Direcciones IP, por ejemplo 142.32.48.231
  - Fechas, teléfonos, etc.
  - Guiones: Hewlett-Packard, co-education
  - white space, whitespace o white-space
  - ¿Eliminar espacios en blanco?
    - Los Angeles, San Francisco, Universidad de la República

webir – Stop Words en la Construcción del Indice Invertido

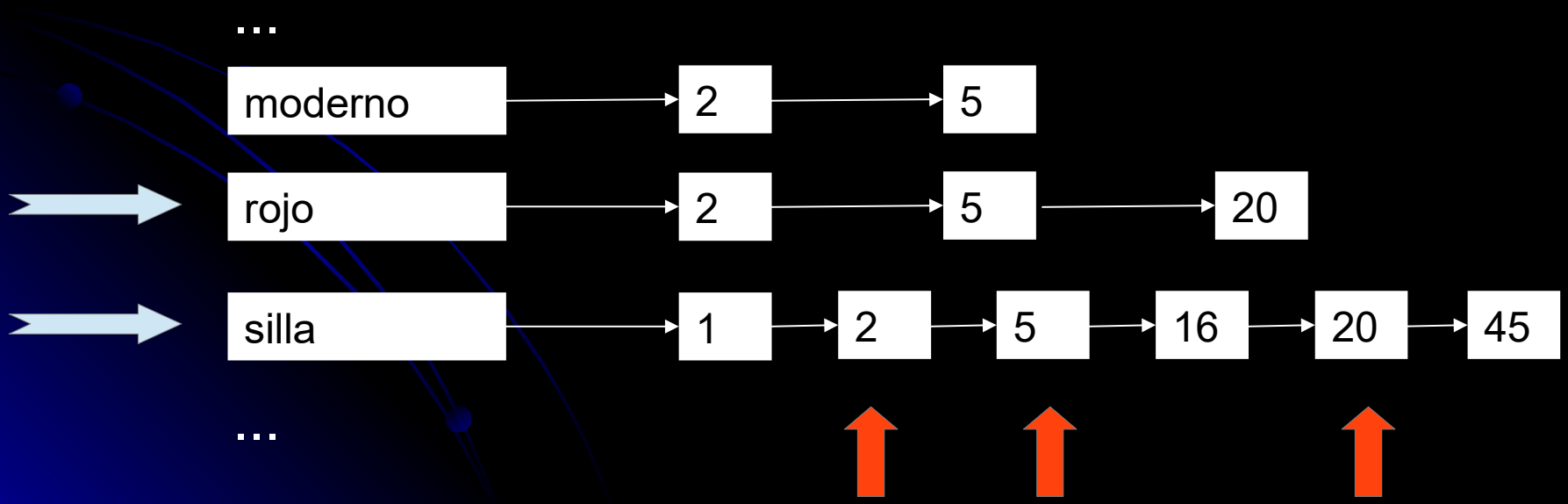
- Eliminar palabras muy comunes que no son útiles para encontrar los documentos que respondan a una necesidad de información
- Lista de palabras ordenadas por la frecuencia (número de veces) que aparecen en la colección de documentos
  - Los términos más frecuentes son ”stop words”
  - ”stop list”
  - el, la, de, hasta, por, etc.
- Reduce postings, pero hay otras formas
  - Asignar pesos
  - No se usa mucho en motores de búsqueda en Internet

webir – Normalización para la Construcción del Indice Invertido

- USA vs. U.S.A.
- Verbos - correr, corre, corrió, corren, etc.
- Llevar a una versión canónica
  - Clases de equivalencias explícitas
    - USA = {USA, U.S.A., ...}
    - correr = {correr, corre, corrió, corren, ...}
  - Reglas de transformación (implícitas)
    - Quitar caracteres como ´.´, ´-´ o ´’´
    - Verbos a infinitivo (correr = {correr, corre, corrió, corren, …})
    - Complementar con lista manual de sinónimos como {auto, automóvil, carro, …} para el momento de la consulta
  - Procesamiento en el momento de la consulta!!!

webir – Normalización para la Construcción del Indice Invertido

- Ejemplo
- Consulta: sillón colorado
- Clases de equivalencias (implícitas o éxplicitas)
  - rojo = {roja, rojo, rojizo, carmín, colorado...}
  - silla = {silla, sillón, banco}
  - Otras



webir – Normalización para la Construcción del Indice Invertido

Alternativas a las clases de equivalencia (menos eficientes):

1) Se mantienen TODAS las palabras encontradas sin normalizarlas en el diccionario
   - Lista de expansión de consultas
     - Explícita
     - Implícita
   - Cada palabra de la consulta es una disyunción de términos del diccionario
   - Menos procesamiento de la consulta, más espacio
   - Flexibilidad
     - Windows – Windows
     - windows – Windows, windows, window
     - window – window, windows
2) Expandir al construir el índice (auto, ponerlo también como automóvil) - más espacio

webir – Normalización para la Construcción del Indice Invertido

- Reglas comunes:
  - Eliminar signos no distintivos
  - Eliminar signos de puntuación/exclamación/acentuación que no son distintivos
    - cliché vs. Cliche - idioma
    - Probabilidad de que los usuarios la escriban así, España vs. Espana

webir – Normalización para la Construcción del Indice Invertido

- Reglas comunes: mayúsculas a minúsculas
  - Diferencias al comienzo de las oraciones
  - Errores en nombres
    - ferrari vs. Ferrari
  - Puede producir efectos no deseados
    - Nombres de empresas - General Motors
    - Nombres o apellidos – Bush, Black
    - C.A.T -> CAT -> cat
  - Se puede hacer de manera selectiva
    - Las del comienzo de las oraciones
    - Las de los títulos
    - ”Truecasing” - machine learning

webir – Particularidades de la Normalización

- Dependen del idioma
  - Colour vs. color
  - 3/22/2010, 22/3/2010, 22 de Marzo de 2011
  - Información en Internet en Inglés (25%), Chino (19%), Español (8%), Árabe (5%)...
  - Documentos que contienen palabras en diferentes idiomas
  - Chebyshev vs Tchebycheff, Beijing vs. Pekín – clase de equivalencia fonética - algoritmo Soundex

webir – Conjugaciones/Familias de Palabras en la Normalización

- Conjugaciones de verbos
  - am, are, is, be → be
- Familias de palabras
  - democracia, democrático, demócrata, etc.
  - car, cars, car's, cars´ → car
- "The boy´s cars are different colors …" → the boy car be different …
- Stemming
  - Proceso rudimentario, acortar, truncar
  - Algoritmo de Porter - 1980
- Lemmatization ("lemma" es la raíz de la palabra)
  - Análisis morfológico de las palabras

webir - Normalización

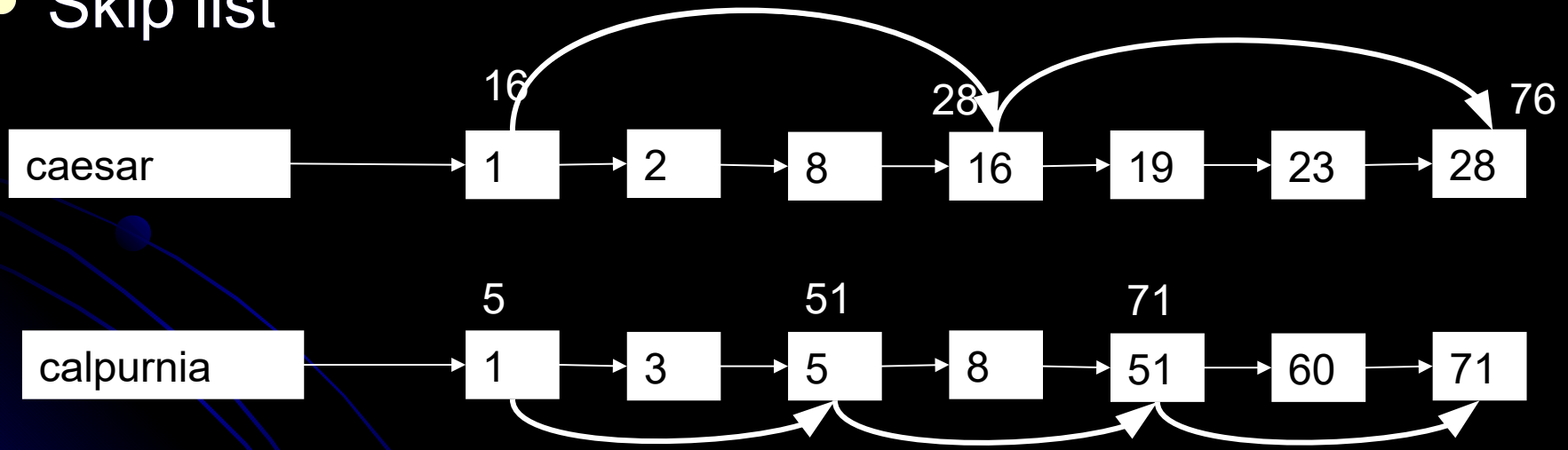
- Ejemplo Stemming
  - Oper = {operate, operating, operates, operation, operative, operational}
    - operational AND research
    - operating AND system
- Ejercicio, es cierto que:
  - En un sistema de RI Booleana Stemming NO reduce la Precisión
  - En un sistema de RI Booleana Stemming NO reduce Recall
  - Stemming aumenta el tamaño del vocabulario (diccionario)
  - Se debería usar Stemming al procesar los documentos pero no la consulta

webir - Normalización

- Ejercicio, es cierto que:
  - En un sistema de RI Booleana Stemming NO reduce la Precisión
  - En un sistema de RI Booleana Stemming NO reduce Recall
  - Stemming aumenta el tamaño del vocabulario
  - Se debería usar Stemming al procesar los documentos pero no la consulta
- El proceso de Stemming puede aumentar el Recall y puede disminuir la Precisión

webir - Optimización de la Lista Postings

- En general se recorren en forma simultánea como en “merge”
- O (m+n)
- Skip list



- ¿Dónde ubicar los punteros skip y cómo usarlos?

webir - Optimización de la Lista Postings

- Los punteros skip sólo sirven para las consultas AND
- Demasiados punteros skip
  - Muchas comparaciones
  - Mucha memoria/espacio
  - Aprox. $\sqrt{P}$ para una lista de largo P, distribuidos uniformemente - heurística

webir - Consultas por Conceptos o Frases

- Hasta 10% de las consultas
- Universidad de la República
- “Universidad de la República”
- Concepto relacionado con la cercanía de palabras
- Soluciones eficientes
  - Indices de pares de palabras - biwords
  - Indices posicionales

webir - Indices de Pares de Palabras

- Indices de pares de palabras consecutivas

Friends, Romans, countrymen...

So let it be with Caesar...

friends | romans | countrymen

so | let | it

friends romans | romans countrymen

so let | let it

- Puede dar errores, se busca “Universidad Católica Uruguay”
  - “Universidad Católica”
  - “Católica Uruguay”
  - “Iglesia Católica”

webir - Indices de Pares de Palabras Extendidos

- Se clasifican de forma sintáctica los términos
  - Sustantivos (N)
  - Verbos, Adjetivos
  - Artículos y preposiciones (X)
  - Nombres (N)
  - otros
  - → NN, NXN, NX*N, etc. son los términos del vocabulario, como “Instituto de Computación”
- Indice de frases - se extiende a ternas o más
  - Reduce los errores
  - Aumenta mucho el vocabulario
- ¡Se agregan al índice invertido, no lo sutituyen!

webir - Indices Posicionales

- En cada posting, además del docID, se guarda la lista de posiciones donde aparece el término
  - docID:<pos1, pos2, pos3, …>
  - be, 178239:
    <1, 2: <17, 25>;
    4, 5: <17, 191, 291, 430, 434>;
    5, 3: <14,19, 101>;
    … >
  - Buscar cada término
    - Usar la frecuencia - eficiencia
    - Controlar las posiciones – cercanía, frases

webir - Indices Posicionales

- Aumenta el tamaño del diccionario
  - Complejidad de las operaciones O(T) en lugar de O(N)
    - T número de términos en la colección de documentos
    - N número de documentos
- Resuelve
  - Cercanía entre palabras
  - Orden entre palabras

webir - Indices Mixtos

- Combinación de índices posicionales e índices de pares de palabras (o frases)
- Frases comunes no conviene buscarlas en índices posicionales
  - “Tabaré Vázquez”, “Tabaré Cardozo”, “Presidente Vázquez”, “Mauricio Macri”, “Zapatería Macri”
- Se puede acelerar sustancialmente aquellas consultas por frases que contienen palabras muy comunes (por separado), pero que juntas aparecen menos frecuentemente y tienen otro significado

webir - Indices Posicionales

- Ejercicio
  - ¿Qué problema puede traer la eliminación de stopwords antes de la creación del índice posicional o de pares de palabras?
  - Dar un ejemplo.
  - ¿Cómo se puede resolver?

webir – Próxima clase

- Recuperación Tolerante a Errores de Ortografía y Otras Inconsistencias
  - Búsquedas con ”Comodines”
  - Índice k-gram
  - Correcciones Ortográficas