webir

Clase 3

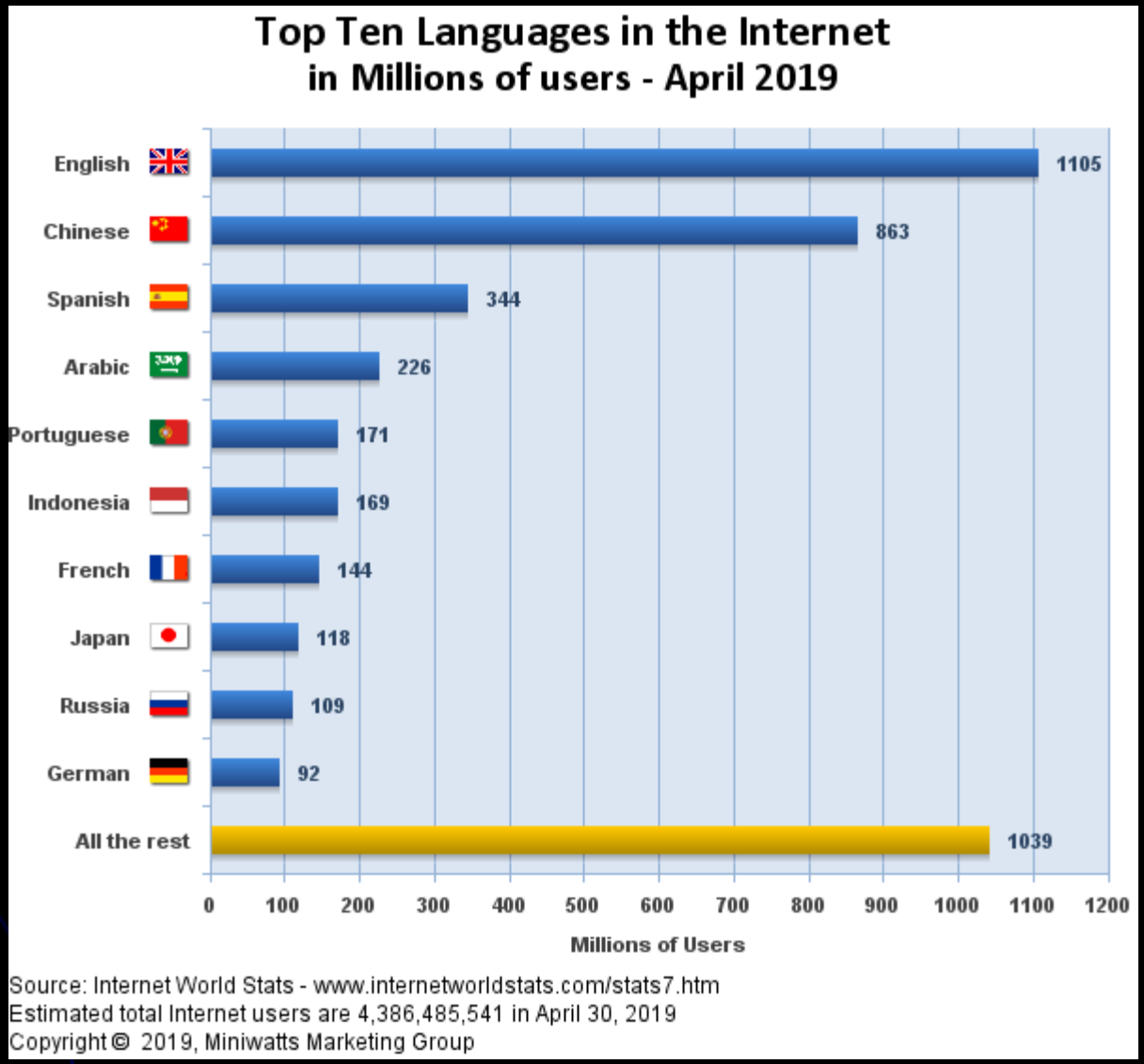
webir - Temas

- Repaso
  - Consultas por Conceptos o Frases
  - Índices de Pares de Palabras
  - Índices Posicionales
- Recuperación Tolerante a Errores de Ortografía y Otras Inconsistencias
  - Estructuras auxiliares
  - Búsquedas con ”Comodines”
  - Índice k-gram
  - Correcciones Ortográficas

webir – Construcción del Indice Invertido

- Granularidad
- Tokenize
- Identificación del idioma
- Normalización
- Crear el índice

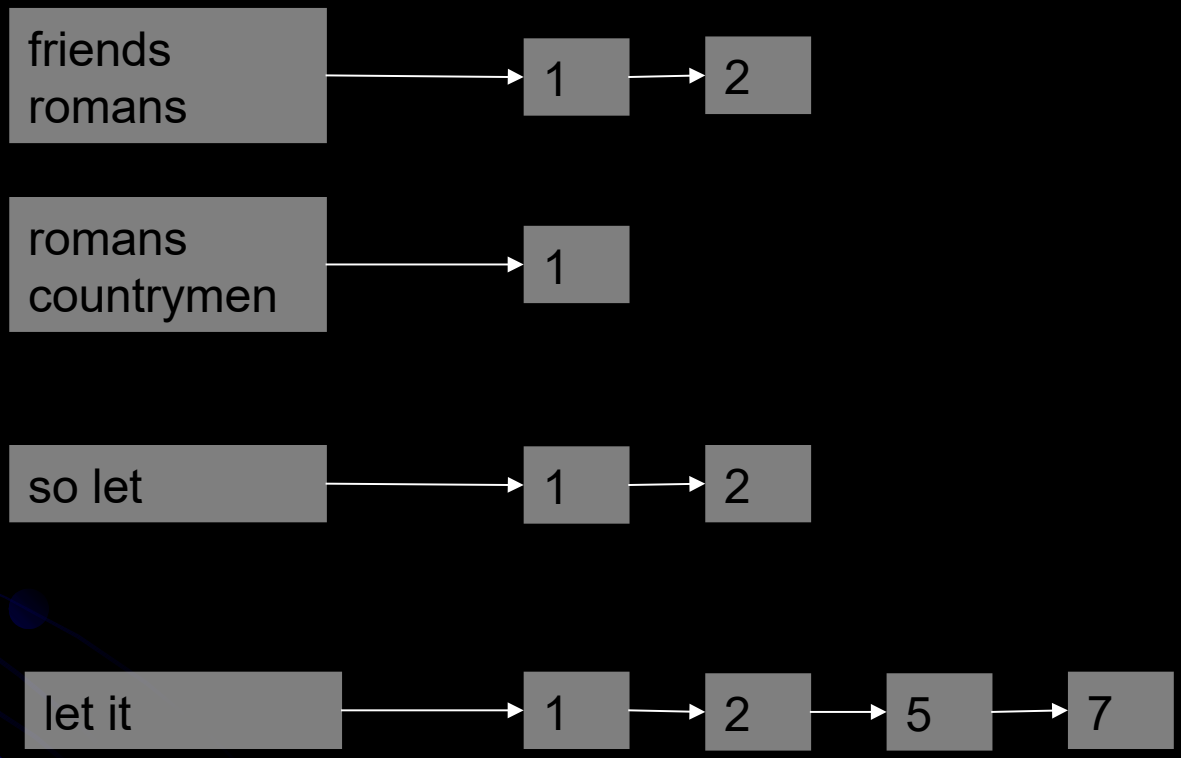
- Optimización de la Lista Postings



webir - Consultas por Conceptos o Frases

- “Universidad de la República”
- Concepto relacionado con la cercanía de palabras
- Soluciones eficientes
  - Índices de pares de palabras – biwords
    - Menor cantidad de errores que sólo el índice invertido común
    - Índice de frases - ternas o más - reduce los errores - aumenta el vocabulario
  - Índices posicionales

webir - Indices de Pares de Palabras



- Se clasifican de forma sintáctica los términos
- ¡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>;
    … >
- 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

webir - Indices Mixtos

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

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 resuelve?

webir - Recuperación Tolerante a Errores de Ortografía y Otras Inconsistencias

- Estructuras de datos auxiliares para las búsquedas en el vocabulario del diccionario
- Búsquedas con ”comodines”
  - o*u*i*e*a – orquídea
  - Automat* - automático, autómata, automatizar
- Errores de ortografía
- Búsquedas de términos fonéticamente similares

webir - Estructuras de Datos Auxiliares para el Vocabulario del Diccionario

- Búsqueda de términos
  - Hashing
  - Árboles de búsqueda
- Depende de
  - Cuántos términos
  - Cantidad estática
  - Sólo se agregan términos o pueden desaparecer
  - Frecuencia de acceso de los términos

webir - Estructuras de Datos Auxiliares para el Vocabulario del Diccionario

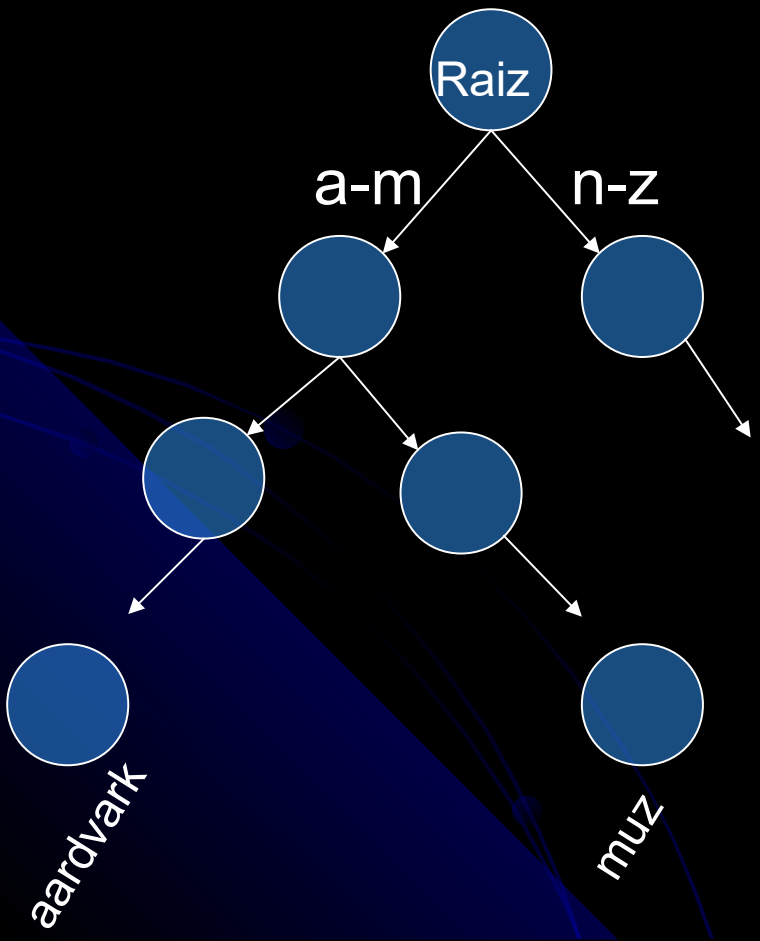
- O(1) caso promedio?
- Hashing
  - Transformar cada término en un entero
  - Resolver colisiones de forma simple
  - Espacio grande para que hayan pocas o ninguna colisión
    - Demasiado espacio para Internet
    - Obsoleto en poco tiempo para Internet
    - No hay forma de encontrar términos parecidos

webir - Estructuras de Datos Auxiliares para el Vocabulario del Diccionario

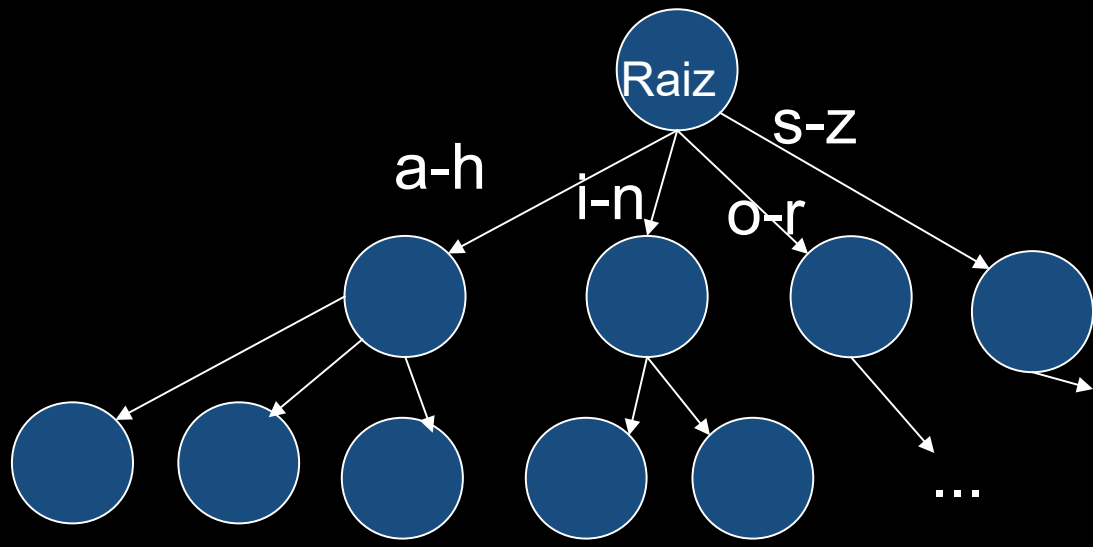
- Árboles
  - Binarios
    - c/nodo tiene hasta 2 hijos
    - O($\log_2$ N) depende de mantenerlo balanceo
  - Árboles-B – usados para diccionarios
    - c/nodo tiene entre [a,b] hijos
    - Balanceados
    - Práctico para levantar mayor porción del árbol de disco
- Requieren un orden único asociado a los caracteres
  - No siempre ocurre por ej. Chino

webir - Estructuras de Datos Auxiliares para el Vocabulario del Diccionario

Árbol binario

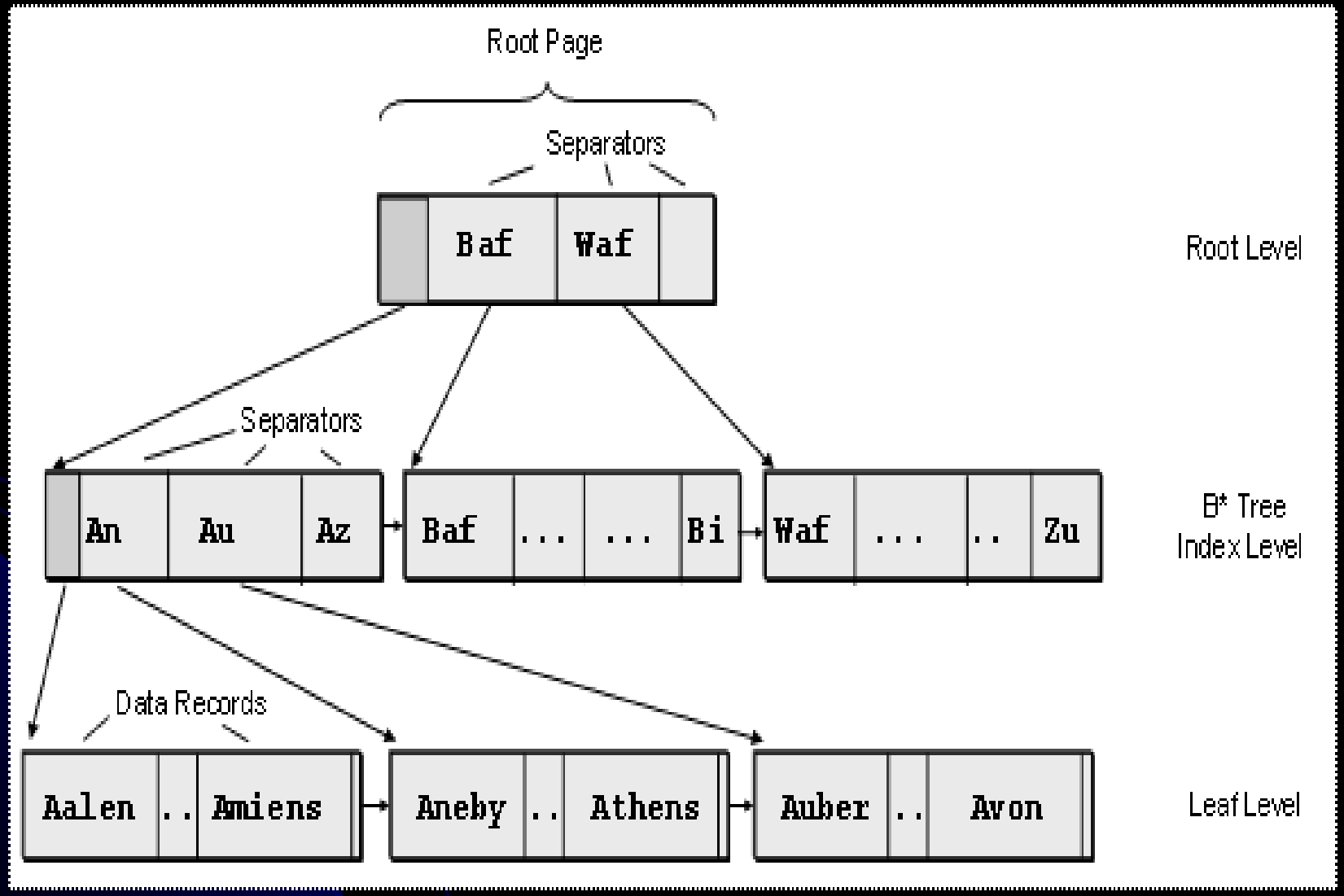


Árbol-B



Árbol-B con cant. nodos en el intervalo [2-4]

webir - Estructuras de Datos Auxiliares para el Vocabulario del Diccionario



webir - Búsquedas con ”Comodines”

- No se conoce completamente la palabra - S*dney
- Se busca conscientemente distintas versiones de la misma palabra - color y colour
- Se buscan variantes de la palabra – operation, operational, operative, etc.
- Se busca palabras en otros idiomas – Universit* Stuttgart

webir - Búsquedas con ”Comodines”

- mon* - se puede encontrar en un árbol de búsqueda para las palabras del vocabulario
  - El conjunto W de palabras con prefijo mon
  - |W| búsquedas en el diccionario
- *mon - se puede encontrar en un árbol de búsqueda para las palabras invertidas del vocabulario (nom*)
  - Árbol invertido
  - Árbol de búsqueda para las palabras invertidas del vocabulario (lemon - nomel)
- ¿se*mon?

webir - Búsquedas con ”Comodines”

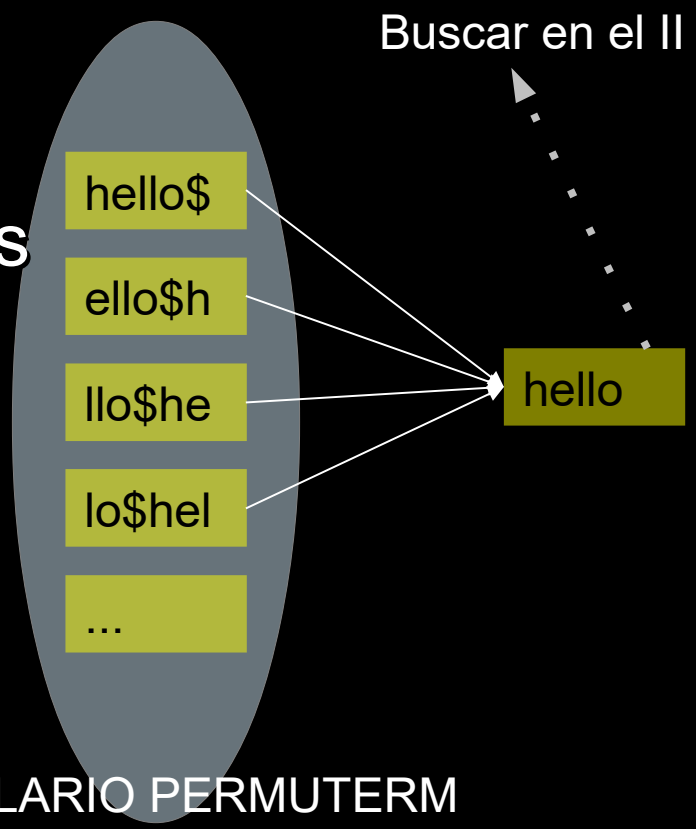
- mon* - se puede encontrar en un árbol de búsqueda para las palabras del vocabulario
  - El conjunto W de palabras con prefijo mon
  - |W| búsquedas en el diccionario
- *mon - se puede encontrar en un árbol de búsqueda para las palabras invertidas del vocabulario (nom*)
  - Árbol invertido
- se*mon
  - mediante ambos árboles tomar intersección de resultados de se* y *mon
  - después buscar en vocabulario, recorrer las listas de postings

webir - Búsquedas con ”Comodines”

- Caso general de búsquedas con ”comodines”
  - o*u*i*e*a – orquídea
- Búsqueda de la palabra q_w ev. con más de un comodín = se busca conjunto W de palabras que resuelven la consulta
  - Primero encontrar un conjunto Q tal que $W \subset Q$
  - Controlar las palabras de Q que cumplen las condiciones para hallar W
- Índice Permuterm
- Índice k-gram

webir - Índice Permuterm

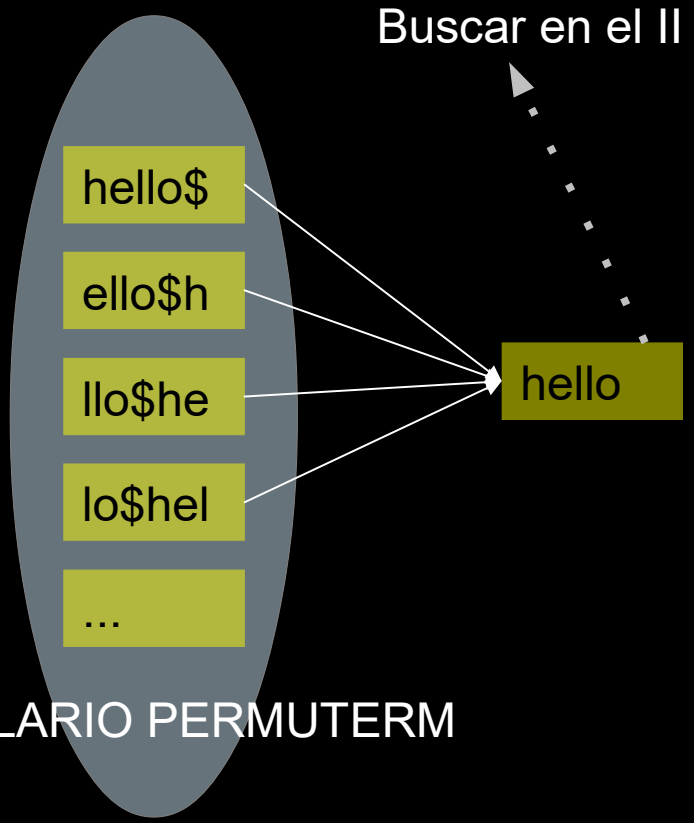
- Se agrega $ al final
- Permutaciones de las palabras
  - hello$
  - ello$h
  - llo$he
  - lo$hel
  - ...



- En el índice permuterm cada permutación apunta al término original
- Se agrega un árbol de búsqueda

webir - Índice Permuterm

- Consulta m*n
  - Se busca n$m* (ev. en un árbol de búsqueda)
  - Luego en el índice invertido común
- Se incrementa el tamaño del diccionario

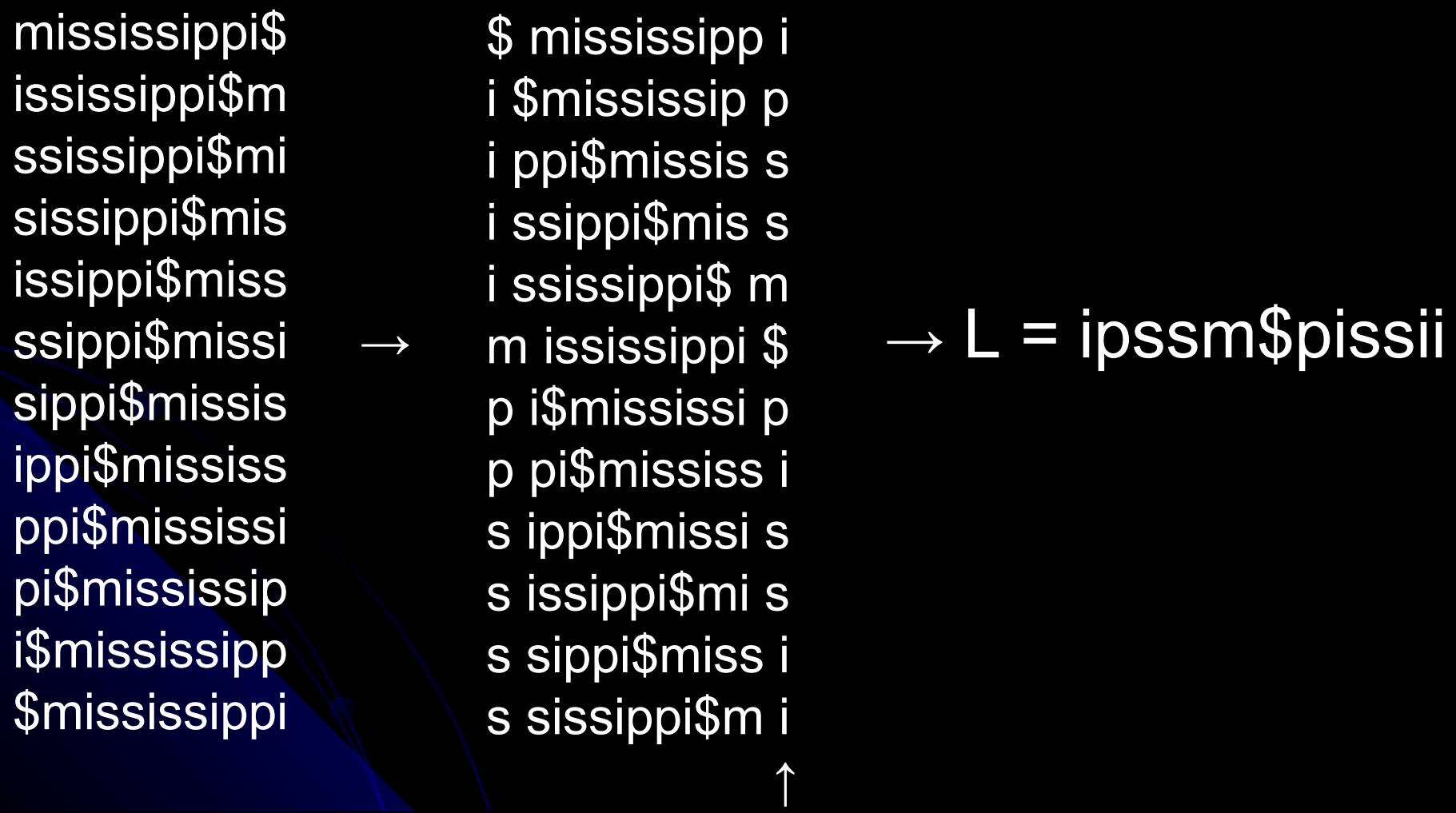


webir - Índice Permuterm

- Se filtran los términos que cumplan
- Finalmente se buscan los términos que cumplen las condiciones en el diccionario

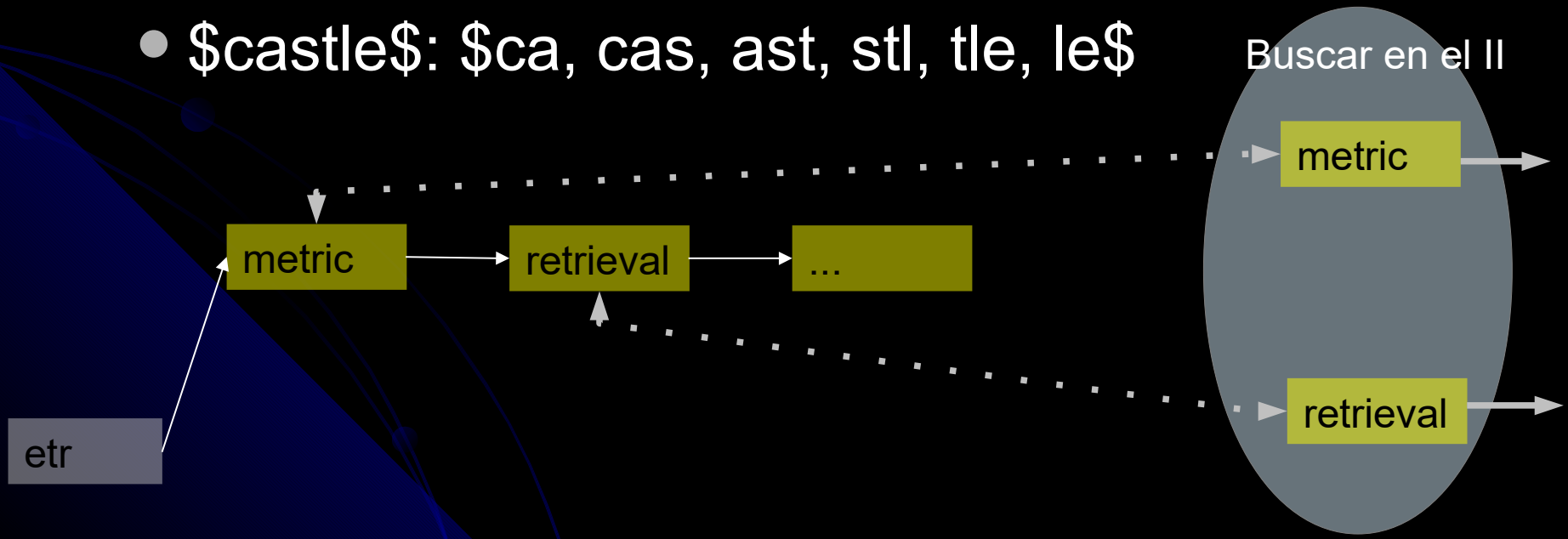
- Resolver per*fe*o
  - Buscar las palabras o$per* (= per*o)
    - {perfecto, perchero, per}
  - Filtrar las que tengan ”fe” en el medio
  - perfecto si, pero no perchero

“Compressed Permuterm Index” Ferragina y Venturini - 2010



webir - Índice k-gram

- Secuencias de k caracteres
  - 3-gram de la palabra castle: cas, ast y stl
- $ para comenzar y finalizar las subsecuencias de una palabra
  - $castle$: $ca, cas, ast, stl, tle, le$



webir - Índice k-gram

- Resolver re*ve
  - Buscar $re AND ve$
  - relive, remove, retrieve
- Buscar los términos que cumplen las condiciones en el diccionario
- Resolver red*
  - Buscar $re AND red
  - Error retired
- Filtrado posterior para eliminar errores
  - String matching

webir - Correcciones Ortográficas

- britney spears = britian spears, britney's spears, brandy spears, prittany spears
- Tipos de correcciones
  - Encontrar el/los términos más cercanos
  - Elegir entre dos (o más) términos cercanos,
    - grnt = grunt, grant
    - Mayor frecuencia en la colección de documentos
    - Mayor frecuencia en las consultas de los usuarios
- Dos técnicas usada en conjunto:
  - Distancia de edición
  - k-gram overlap

webir - Correcciones Ortográficas – Acciones

- Devolver documentos con el término original y otros términos que son correcciones del original
- Devolver documentos con términos que son correcciones del original sólo si el original NO aparece en el diccionario
- Devolver documentos con términos que son correcciones del original sólo si el original devuelve menos de *m* resultados
- Si el original devuelve menos de *m* resultados se presentan al usuario términos alternativos que son correcciones del original

webir - Correcciones Ortográficas

- Dos alternativas: corregir los términos de la consulta por separado o de forma conjunta
  - Términos aislados
    - Corregir las palabras de la consulta en forma individual, aún en caso de conceptos o frases
    - Algunas palabras puede ser que NO se corrigan ya que no se detectan como errores
    - ”flew form Heathrow”
  - Corrección sensible al contexto

webir - Correcciones Ortográficas – Distancia de Edición

- Distancia de edición entre s_1 y s_2 = mínimo número de operaciones de edición necesarias para transformar s_1 en s_2 (Levenshtein)
  - Insertar un caracter
  - Borrar un caracter
  - Reemplazar un caracter por otro
- Se puede asignar pesos a las operaciones
  - Reemplazar ”a” por ”p” es menos probable que reemplazar ”a” por ”s” – teclado
- Algoritmo de programación dinámica

webir - Correcciones Ortográficas – Distancia de Edición

- ¿Con que términos del vocabulario se debe comparar la palabra de la consulta?
  - ¿Todos?
- Heurísticas
  - Términos que empiecen con la misma letra
  - Versión del índice Permuterm sin $
  - Omitir algunas letras del final o del comienzo y buscar en el índice Permuterm

webir - Correcciones Ortográficas – k-gram

Buscar en el índice k-gram un conjunto de palabras con muchas subsecuencias en común con la original para buscar las más cercanas – limitar aún más y usar dist. edición

- ¿k? ¿Cuántas palabras?
- ¿Cuántas subsecuencias en común?
  - Coeficiente Jaccard (medida de coincidencias)

$$|A \cap B| / |A \cup B|$$

  - Conjuntos de subsecuencias que superen un umbral – luego se calcula distancia de edición

webir - Correcciones Ortográficas - Corrección Sensible al Contexto

- Corregir las palabras de la consulta en forma individual
- Buscar como conceptos o frases con y sin las correcciones
- Usar frecuencias (por ejemplo en el índice de pares de palabras) para acotar las búsquedas
  - En el corpus
  - En las consultas
  - ”flew form Heathrow” → ”flew from Heathrow”

webir - Correcciones Ortográficas

- Ejercicio
  - Explicar porqué la distancia en edición entre s1 y s2 nunca es mayor que max{|s1|, |s2|}

webir – Cronograma Proyectos

- Semana del 15/4
  - Definición de grupos
    - se deben registrar los grupos
    - se recomienda leer proyectos de años anteriores
    - empezar a definir el tema/problema sobre el que quieren hacer el proyecto
- Semanas del 22/4 y 29/4
  - Monitoreos de los grupos para validar tema del proyecto
- No hay clases
- **1era entrega hasta el 13/5**
    - definición concreta del problema a abordar, una descripción alto nivel de las componentes del sistema y las herramientas que piensan que van a utilizar en cada una de ellas (si ya pudieron investigarlo)

Proyectos

- El tema específico deberá ser validado
- Los proyectos constan de tres actividades
  - Investigación sobre un tema relacionado con el curso
  - Diseño de una solución a un problema (usando lo aprendido)
  - Implementación de la solución
- Los proyectos deben concluirse con
  - Informe sobre el trabajo realizado
  - Una presentación oral del trabajo (max. 15 minutos)

Proyectos - Optimización en SRI

- **Análisis de Consultas**
  - "Improving search engines by query clustering“, Ricardo Baeza-Yates, Carlos Hurtado, Marcelo Mendoza, 2007
  - "Agglomerative clustering of a search engine query log“, D Beeferman, A Berger, Proceedings of the sixth ACM SIGKDD, 2000
  - "CETR: content extraction via tag ratios", Tim Weninger, William H. Hsu, Jiawei Han, Proceedings of the 19th international conference on World wide web, 2010
  - Intención de usuario - Aprendizaje supervisado, semi supervisado y no supervisado
    - Weka homepage: http://www.cs.waikato.ac.nz/~ml/weka/
    - http://people.cs.uchicago.edu/~vikass/svmlin.html.
    - "Semi-Supervised Learning Literature Survey", Xiaojin Zhu, University of Wisconsin – Madison
- **Manejo y búsqueda de información no tradicional (multimedia, imágenes, etc.).**

Proyectos - Optimización en SRI

- **Seguridad y RI**
  - “Misuse detection for information retrieval systems”, Cathey, Ma, Goharian, Grossman, Proceedings of the twelfth international conference on Information and knowledge management, 2003
  - “Using relevance feedback to detect misuse for information retrieval systems”, Ma, Goharian, Proceedings of the thirteenth ACM international conference on Information and knowledge management, 2004
  - “Query length impact on misuse detection in information retrieval systems”, Ma, Goharian, Proceedings of the 2005 ACM symposium on Applied computing, 2005
- **Perfil de Usuario**
  - "Deriving Concept-based User Profiles from Search Engine Logs“, Kenneth Wai-Ting Leung, Dik Lun Lee, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2007
  - "Using A Graph-based Ontological User Profile For Personalizing Search", Mariam Daoud, Lynda Tamine-Lechani, Mohand Boughanem, CIKM’08, 2008.

Proyectos - Optimización en SRI

- **Filtrado Colaborativo**
  - "Item-based collaborative filtering recommendation algorithms", Badrul Sarwar, George Karypis, Joseph Konstan, John Reidl, Proceedings of the 10th international conference on World Wide Web, 2001
  - "Recommendation as Classification: Using Social and Content-Based Information in Recommendation", Chumki Basu, Haym Hirsh, William Cohen, AAAI-98 Proceedings, 1998
- **Trust**
  - "Trust-aware recommender systems", Paolo Massa,Paolo Avesani, Proceedings of the 2007 ACM conference on Recommender systems, 2007.
  - "Trust-aware collaborative filtering for recommender systems", P. Massa and P. Avesani, Proc. of Federated Int. Conference On The Move to Meaningful Internet: CoopIS, DOA, ODBASE, 2004
  - "Incorporating similarity and trust for collaborative filtering", Su Chen , Tiejian Luo , Wei Liu , Yanxiang Xu, Proceedings of the 6th international conference on Fuzzy systems and knowledge discovery, 2009
  - "Trust in recommender systems", John O'Donovan, Barry Smyth, International Conference on Intelligent User Interfaces, 2005

Proyectos - Redes Sociales y RI

- **Extracción del contexto social, noticias, preferencias**
  - Facebook, twitter, otras
- **Autoridad en redes sociales**
  - “Authoritative sources in a hyperlinked environment”, J. Kleinberg, Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998
  - "The Anatomy of a Large-Scale Hypertextual Web Search Engine", Sergey Brin, Larry Page, Proceedings of the 7th international conference on World Wide Web (WWW), 1998
  - “Predicting Positive and Negative Links in Online Social Networks”, J. Leskovec, D. Huttenlocher, J. Kleinberg., Proc. 19th International World Wide Web Conference, 2010
  - "Extracting reputation in multi agent systems by means of social network topology“, Josep M. Pujol, Ramon Sangüesa, Jordi Delgado, Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 1 table of contents, 2002
- **Recomendaciones**
  - "FilmTrust: Movie Recommendations using Trust in Web-based Social Networks“, Jennifer Golbeck, James Hendler, 2006
  - "Recommending collaboration with social networks: a comparative evaluation“, David W. McDonald, Proceedings of the SIGCHI conference on Human factors in computing systems, 2003
  - "Tag-Based Contextual Collaborative Filtering", Reyn Nakamoto, Shinsuke Nakajima, Jun Miyazaki, Shunsuke Uemura, IAENG International Journal of Computer Science, 2008
  - “Personalization via friendsourcing”, BERNSTEIN,TAN, SMITH, CZERWINSKI, HORVITZ, ACM Transactions on Computer-Human Interaction (TOCHI) , 2010
- **Small-World Phenomena**
  - "The Small-World Phenomenon and Decentralized Search”, J. Kleinberg. A short essay as part of Math Awareness Month, appearing in SIAM News 37(3), 2004

Proyectos - Web Crawling y Estructura de la Web

1. **http://www.cwr.cl/projects/WIRE/**
   - Por temas
   - Por zonas geográficas

2. **Estructura de la Web**
   - “Authoritative sources in a hyperlinked environment”, J. Kleinberg, Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998
   - "The Anatomy of a Large-Scale Hypertextual Web Search Engine", Sergey Brin, Larry Page, Proceedings of the 7th international conference on World Wide Web (WWW), 1998
   - "Inferring Web communities from link topology“, D. Gibson, J. Kleinberg, P. Raghavan, Proc. 9th ACM Conference on Hypertext and Hypermedia, 1998