Stemming – Lematización

UCR – ECCI

CI-2414 Recuperación de Información

Prof. M.Sc. Kryscia Daviana Ramírez Benavides

Introducción

- Frecuentemente se usa una palabra para representar una consulta, pero sólo una variante de esa palabra está presente en un documento relevante (singular, plural, gerundio, etc.).
- El problema puede ser solucionado con la sustitución de una palabra por todas sus formas.
- Una de las técnicas para mejorar la realización de los sistemas RI es ofrecer a los usuarios formas de encontrar variantes morfológicas de los términos de búsqueda:
  - Por ejemplo: si un usuario usa el término “lematización” en su búsqueda es probable que esté interesado en variantes tales como “lematizado” y “lexema”.

Introducción (cont.)

- Se usa el término *junción* para significar el hecho de fundir o combinar como término general para el proceso de juntar variantes morfológicas.
- La *junción* puede ser:
  - Manual: usando alguna clase de expresiones regulares.
  - Automática: a través de programas denominados lexematizadores.
- La lematización también se usa en los sistemas RI para reducir el tamaño de los índices
  - Dado que un lexema corresponde normalmente a varios términos se pueden alcanzar factores de compresión del 50% almacenando lexemas en vez de términos.

Definición

- El *stemming* o lematización es hallar el lema (*stem*) de las palabras y no tiene que tener significado.
- Un algoritmo de *stemming* es un proceso de normalización lingüística en el cual las diferentes formas que puede adoptar una palabra son reducidos a una única forma común, a la cual se denomina *stem* o lema.
- *Stem* o lema es la porción de la palabra después de eliminar sus afijos. Por ejemplo: **perr** para las palabras **perros**, **perros**, **perrito**, **perrote**, etc.

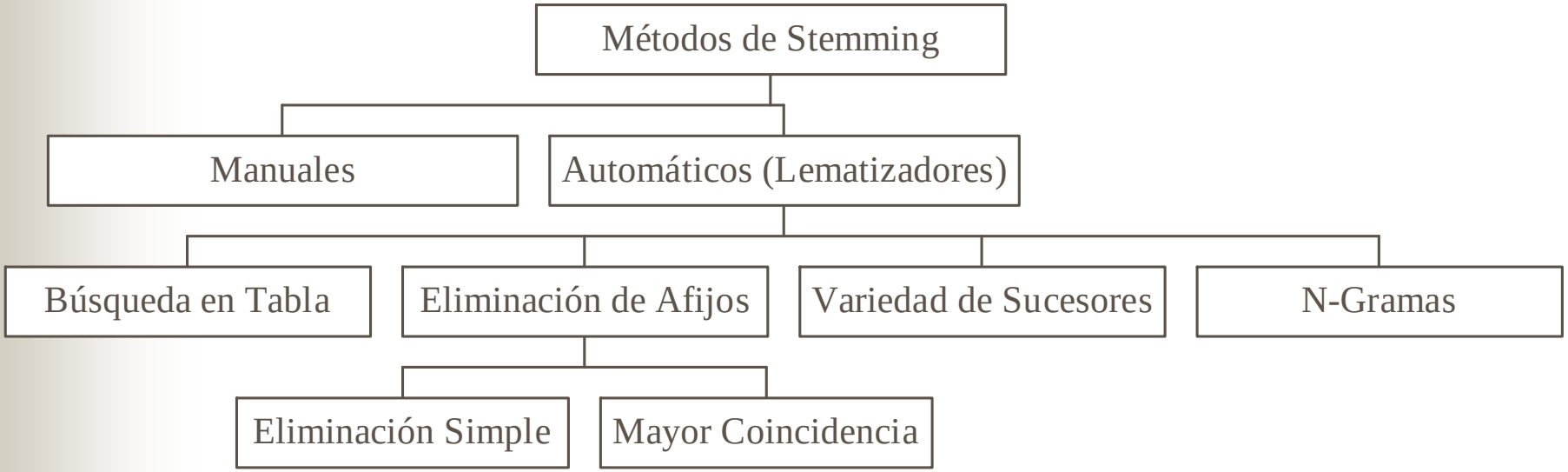
¿Cuándo lematizar?

- Los términos pueden ser lematizados en el momento de:
  - La indexación.
    - Eficiente
    - Compresión de índices
    - Pérdida de término completos
  - La búsqueda.
- La ventaja de la lematización durante la búsqueda radica en:
  - No se pierde información sobre los términos completos.
- La desventaja de lematizar durante la búsqueda es por la eficiencia, dado que los términos al no estar lematizados se requieren recursos en el momento de la búsqueda.

¿Cuándo lematizar? (cont.)

- Las ventajas de la lematización durante la indexación radican en:
  - Su eficiencia, dado que los términos ya están lematizados no se requieren recursos en el momento de la búsqueda.
  - La compresión de los índices.
  - Mejora la formulación de consultas (incrementa el *recall*).
- La desventaja de lematizar durante la indexación es que se pierde la información sobre los términos completos. A menos que se use un espacio adicional para almacenar las formas lematizadas y sin lematizar.

Taxonomía de los Algoritmos de *Stemming*



Taxonomía de los Algoritmos de *Stemming* Automáticos

- Los términos y sus correspondientes lexemas pueden estar almacenados en una tabla realizando la lematización a través de búsquedas en esa tabla.
- Los algoritmos de eliminación de afijos eliminan sufijos y/o prefijos del término y deja un lexema.
- Los lematizadores de variedad de sucesores usan, como base para la lematización, las frecuencias de las secuencias de letras en un cuerpo del texto.
- Los métodos de *n*-gramas juntan los términos en base al número de bigramas o *n*-gramas que comparten.

Valoración de Lematizadores

- Exactitud:
  - Existen dos formas en que un lematizador puede ser inexacto:
    - *Over-Stemming* (**Hiperlematización**): Al eliminar demasiados caracteres del término durante el proceso se pueden juntar términos que no están relacionados (son de dominios diferentes), causa la recuperación de documentos no relevantes.
    - *Under-Stemming* (**Hipolematización**): Al eliminar menos caracteres de lo debido durante el proceso impidiendo que se junten términos relacionados, causando la no recuperación de documentos relevantes.
- Efectividad de la recuperación (convocatoria y precisión).
- Nivel de compresión de los índices.
- Los lematizadores no suelen ser juzgados por su exactitud lingüística, aunque los lexemas que producen suelen ser muy similares a las raíces.

Búsqueda en Tabla

- Consiste en almacenar en una tabla todos los términos índice y sus lexemas.
- Los términos de las consultas y de los índices podrán ser lematizados a través de búsquedas en la tabla. Usando un árbol-B o una dispersión estas búsquedas serán muy rápidas.
- Es sencillo.

Búsqueda en Tabla (cont.)

- Ejemplo: “Presentar”

Palabra	Combinaciones de Sufijos
Presentarla	arla
Presentarlas	arlas
Presentarle	arle
Presentarles	arles
Presentarlo	arlo
Presentarlos	arlos
Presentarse	arse
Presentase	ase
Presentásemos	ásemos
Presente	e
Presentémonos	émonos

Palabra	Combinaciones de Sufijos
Presentable	able
Presentables	ables
Presentación	ación
Presentaciones	aciones
Presentado	ado
Presentador	ador
Presentadores	adores
Presentándonos	ándonos
Presentar	ar
Presentáramos	áramos
Presentaríamos	aríamos

Búsqueda en Tabla (cont.)

- Ventajas:
  - Usando un árbol-B o una dispersión estas búsquedas serán muy rápidas.
  - Es sencillo.
- Desventajas:
  - Construcción de la tabla.
  - Difícil para palabras específicas a un dominio.
  - Puede provocar una sobrecarga de almacenamiento.
  - No existen tablas estándar (ni para Castellano ni para Inglés).
    - Si existieran, muchos de los términos encontrados en los documentos no estarían representados porque son dependientes de la temática y no pertenecen al vocabulario estándar.

Eliminación de Afijos

- Este tipo de algoritmos elimina los sufijos y/o prefijos de los términos dejando un lexema, transformando en ocasiones el lexema resultante.
- Un ejemplo simple es uno que elimina los plurales de los términos en Inglés, se expresa como un conjunto de reglas de las cuáles sólo se usa la primera aplicable:
  - si una palabra acaba en ‘ies’ pero no en ‘eies’ ni en ‘aies’ entonces ‘ies’ pasa a ‘y’.
  - si una palabra acaba en ‘es’ pero no en ‘aes’ ni en ‘ees’ ni en ‘oes’ entonces ‘es’ pasa a ‘e’.
  - si una palabra acaba en ‘s’ pero no en ‘us’ ni en ‘ss’ entonces ‘s’ pasa a null.

Eliminación de Afijos (cont.)

- No son reglas heurísticas.
- Son reglas que aplicadas a las palabras nos dan su forma común.
- Se basan en reglas gramaticales aplicadas al revés.
- El más conocido y usado es el algoritmo de PORTER, el cuál tiene 30-40 reglas, es para el idioma inglés y sólo elimina sufijos.
- Sólo suelen eliminarse sufijos:
  - Es más sencillo.
  - Hay más sufijos que prefijos.

Eliminación de Afijos (cont.)

- Ventajas:
  - Con un número de reglas pequeño se obtiene gran eficiencia.
  - Ante una nueva palabra se puede sacar su raíz fácilmente.
- Desventajas:
  - Dependen del idioma.
  - Hay que construir la tabla de reglas.
  - El conjunto de reglas empleado es crítico en la calidad del lematizador.

Variedad de Sucesores

- Están basados en trabajos de lingüística estructural que intentan determinar los límites de las palabras y los morfemas, basándose en la distribución de fonemas en un gran cuerpo de pronunciaciones.
- El método también puede usar letras en lugar de fonemas y un cuerpo de texto en lugar de pronunciaciones transcritas fonéticamente.
- Consiste en agrupar palabras con la misma raíz, elimina los sufijos de las palabras.

Variedad de Sucesores (cont.)

- Sea a una palabra de longitud n:
  - a_i es un prefijo de longitud i de a.
- Sea D el corpus de palabras:
  - $D(a_i)$ es el subconjunto de D que contiene aquellos términos cuyas primeras i letras coinciden exactamente con a_i.
- La variedad de sucesores de a_i. denotada por $S(a_i)$ se define como el número de letras distintas que ocupan la posición i+1 en las palabras de $D(a_i)$.
- Una palabra de longitud n tiene n variedades de sucesores: $S(a_1)$, $S(a_2)$, ..., $S(a_n)$.

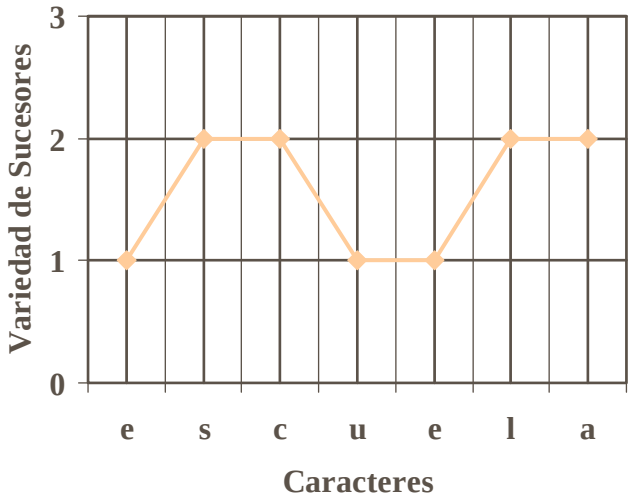
Variedad de Sucesores (cont.)

- La variedad de sucesores de una cadena es el número de caracteres diferentes que siguen a esa cadena en las palabras de un cuerpo de texto
  - Considérese un cuerpo de texto que consta de las siguientes palabras: “escolar”, “escuela”, “escuelas”, “estado”, “escuelilla”, “preescolar”.
  - Para determinar la variedad de sucesores para “escuela” se debe usar el siguiente procedimiento:
    - La primera letra de ‘escuela’ es ‘e’, que está seguida en el cuerpo de texto por un carácter: ‘s’. De esta forma, la variedad de sucesores para ‘e’ es uno.
    - La siguiente variedad de sucesores para ‘escuela’ es dos, porque para ‘es’ están los caracteres ‘c’ y ‘t’ en el cuerpo de texto.
    - Así sucesivamente.

Variedad de Sucesores (cont.)

- La tarea es determinar el *stem* de “escuela” en el siguiente corpus: “escolar”, “escuela”, “escuelas”, “estado”, “escuelilla”, “preescolar”.

Stem	Variedad de Sucesores	Caracteres
e	1	s
es	2	c - t
esc	2	o - u
escu	1	e
escue	1	l
escuel	2	a - i
escuela	2	s - blanco



Variedad de Sucesores (cont.)

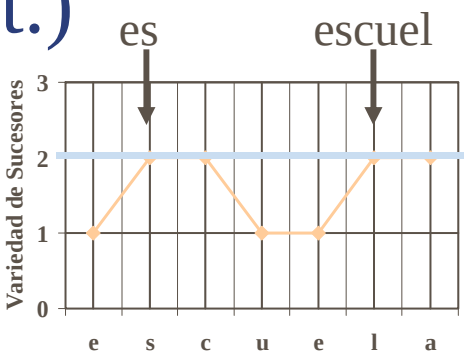
- Una vez que se han calculado las variedades de sucesores para una palabra dada se puede usar esta información para segmentar la palabra.
- Cuatro formas de hacerlo:
  - *Método del valor de corte:* Se selecciona un valor de corte para las variedades de sucesores y se identifica un límite cada vez que se alcanza ese valor de corte. El problema con este método es la selección del valor de corte:
    - Si es muy pequeño se harán cortes incorrectos
    - Si es demasiado grande se perderán cortes correctos
  - *Método de los picos y valles:* Se hace el corte de segmento después de los caracteres cuya variedad de sucesores excede a la del carácter que lo precede y a la del que lo sigue. Este método elimina el problema de la selección del valor de corte

Variedad de Sucesores (cont.)

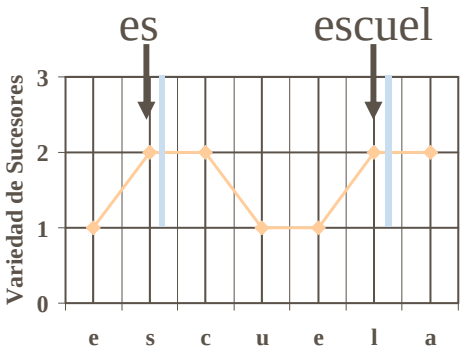
- Cuatro formas de hacerlo (cont.):
  - *Método de palabra completa:* Se hace el corte después de un segmento si éste es una palabra completa en el corpus.
  - *Método de la entropía:* Este método aprovecha la distribución de las variedades de sucesores y funciona como sigue:
    - Sea $|D_{ai}|$ el número de palabras en un cuerpo de texto que comienzan con la secuencia de letras a_i de longitud i.
    - Sea $|D_{aij}|$ el número de palabras en D_{ai} con sucesor j.
    - La probabilidad de que un miembro de D_{ai} tenga sucesor j viene dada por la entropía de $|D_{ai}|$:
    $$H_{ai} = \sum_{j=1}^{26} -\frac{|D_{aij}|}{|D_{ai}|} * \log_2 \frac{|D_{aij}|}{|D_{ai}|}$$
    - Usando esta ecuación se calculan las entropías de una palabra, se selecciona un valor de corte y se identifican los límites de segmento cuando se supera este valor de corte. Se podrían definir de forma similar unas medidas de entropía para predecesores.

Variedad de Sucesores (cont.)

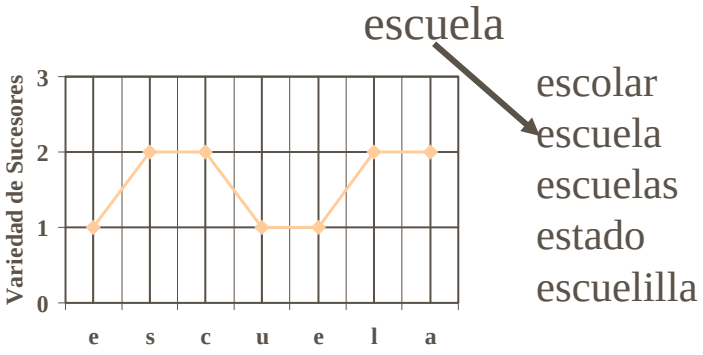
- *Método del valor de corte:*



- *Método de los picos y valles:*



- *Método de palabra completa:*



Variedad de Sucesores (cont.)

- Selección de lexemas:
  - Después de segmentar una palabra hay que seleccionar el lexema. Para ello, algunos investigadores han propuesto la siguiente regla:
    - Si el primer segmento aparece en menos o igual de 12 palabras del corpus entonces el primer segmento es el lexema.
    - Si no, el segundo segmento es el lexema.
  - La condición está basada en la observación de que si un segmento aparece en más de 12 palabras será con mucha probabilidad un prefijo.
  - No se considera lexema a ningún segmento más allá del segundo debido a la escasez de prefijos múltiples.

Variedad de Sucesores (cont.)

- Selección de lexemas (cont.):
  - Tomando una subcadena, a medida que se añaden caracteres, disminuye la variedad de sucesores.
  - Llegado a un punto determinado (cantidad de caracteres de la subcadena), la variedad de sucesores comienza a aumentar.
  - Este es el punto que marca cual es la raíz de la palabra.

Variedad de Sucesores (cont.)

- Ejemplo:
  - La tarea es determinar el *stem* de “escolar” en el siguiente corpus: “escolar”, “escuela”, “escuelas”, “estado”, “escuelilla”, “preescolar”.
  - Se usará el método de picos y valles.

Stem	Variedad de Sucesores	Caracteres
e	1	s
es	2	c - t
esc	2	o - u
esco	1	l
escol	1	a
escola	1	r
escolar	1	blanco

  - El *stem* de escolar es **es**.

Variedad de Sucesores (cont.)

- Ejemplo (cont.):
  - La tarea es determinar el *stem* de “escuela” en el siguiente corpus: “escolar”, “escuela”, “escuelas”, “estado”, “escuelilla”, “preescolar”.
  - Se usará el método de picos y valles.

Stem	Variedad de Sucesores	Caracteres
e	1	s
es	2	c - t
esc	2	o - u
escu	1	e
escue	1	l
escuel	2	a - i
escuela	2	s - blanco

  - El *stem* de escuela es **escuel**.

Variedad de Sucesores (cont.)

- Ejemplo (cont.):
  - La tarea es determinar el *stem* de “escuelas” en el siguiente corpus: “escolar”, “escuela”, “escuelas”, “estado”, “escuelilla”, “preescolar”.
  - Se usará el método de picos y valles.

Stem	Variedad de Sucesores	Caracteres
e	1	s
es	2	c - t
esc	2	o - u
escu	1	e
escue	1	l
escuel	2	a - i
escuela	2	s - blanco
escuelas	1	blanco

  - El *stem* de escuelas es **escuel**.

Variedad de Sucesores (cont.)

- Ejemplo (cont.):
  - La tarea es determinar el *stem* de “estado” en el siguiente corpus: “escolar”, “escuela”, “escuelas”, “estado”, “escuelilla”, “preescolar”.
  - Se usará el método de picos y valles.

Stem	Variedad de Sucesores	Caracteres
e	1	s
es	2	c - t
est	1	a
esta	1	d
estad	1	o
estado	1	blanco

  - El *stem* de estado es **es**.

Variedad de Sucesores (cont.)

- Ejemplo (cont.):
  - La tarea es determinar el *stem* de “escuelilla” en el siguiente corpus: “escolar”, “escuela”, “escuelas”, “estado”, “escuelilla”, “preescolar”.
  - Se usará el método de picos y valles.

Stem	Variedad de Sucesores	Caracteres
e	1	s
es	2	c - t
esc	2	o - u
escu	1	e
escue	1	l
escuel	2	a - i
escueli	1	l
escuelil	1	l
escuelill	1	a
escuelilla	1	blanco

  - El *stem* de escuelilla es **escuel**.

Variedad de Sucesores (cont.)

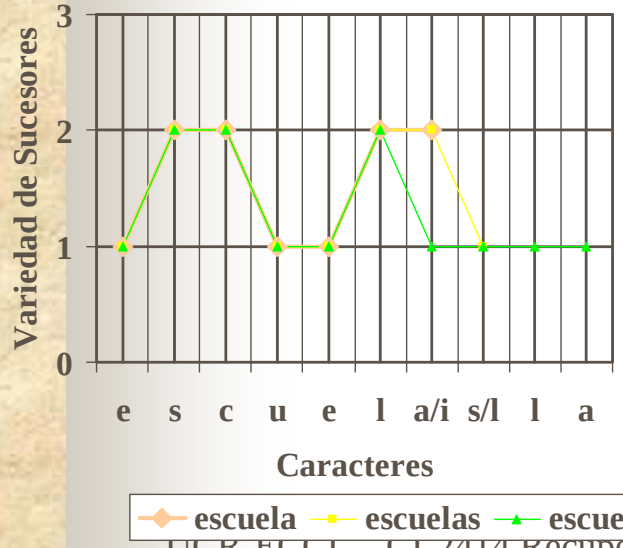
- Ejemplo (cont.):
  - La tarea es determinar el *stem* de “preescolar” en el siguiente corpus: “escolar”, “escuela”, “escuelas”, “estado”, “escuelilla”, “preescolar”.
  - Se usará el método de picos y valles.

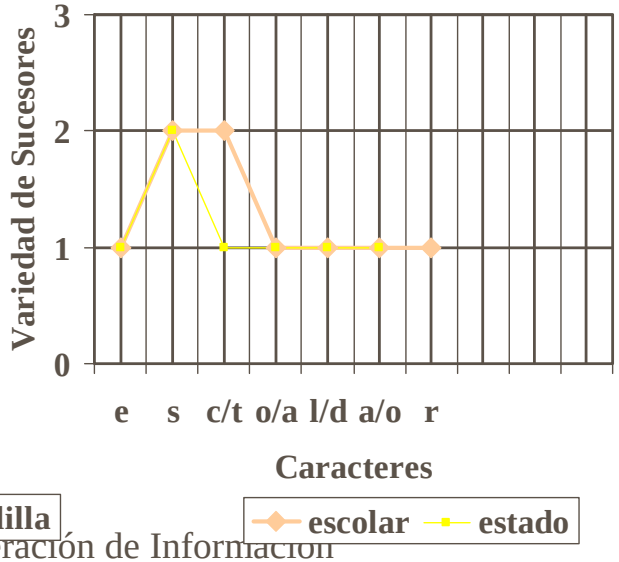
Stem	Variedad de Sucesores	Caracteres
p	1	r
pr	1	e
pre	1	e
pree	1	s
prees	1	c
preesc	1	o
preesco	1	l
preescol	1	a
preescola	1	r
preescolar	1	blanco

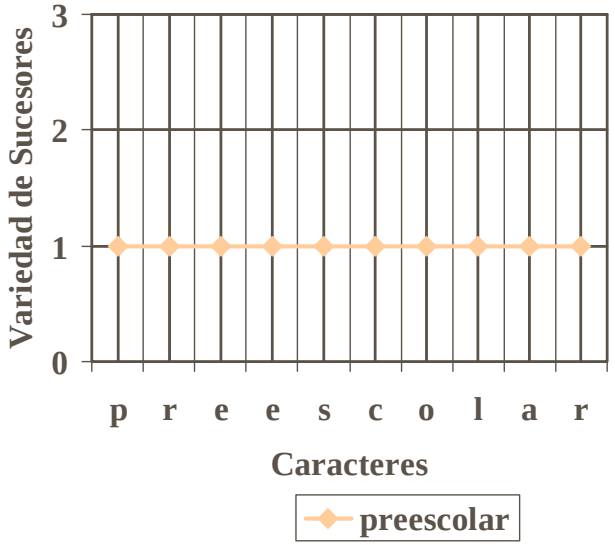
  - El *stem* de preescolar es **preescolar**.

Variedad de Sucesores (cont.)

- Ejemplo (cont.):
  - El corpus: “escolar”, “escuela”, “escuelas”, “estado”, “escuelilla”, “preescolar” se obtiene:
    - Se agrupan escolar y estado, con el *stem* **es**.
    - Se agrupan escuela, escuelas y escuelilla, con el *stem* **escuel**.
    - Se agrupan preescolar, con el *stem* **preescolar**.







Variedad de Sucesores (cont.)

- Ventajas:
  - Es sencillo.
  - Es fácil.
  - Es automático, obtiene la lematización del corpus de texto.
- Desventajas:
  - El problema con la selección del valor de corte para obtener la “raíz”:
    - Si es muy pequeño se harán cortes incorrectos
    - Si es demasiado grande se perderán cortes correctos
  - Sólo elimina los sufijos de las palabras.
  - Es un poco más el rendimiento, computacionalmente hablando.

N-Gramas – Bigramas

- Está basado en el método de los bigramas compartidos:
- Es heurístico.
- Aunque está incluido como método de lematización resulta un poco confuso porque no se produce tal lematización.
- En esta aproximación se calculan medidas de asociación entre pares de términos basándose en los bigramas únicos compartidos.
- Una vez que se han contado los bigramas únicos para la pareja de términos se calcula una medida de similitud basada en el coeficiente de Dice => **S = 2*C/(A+B)**

N-Gramas – Bigramas (cont.)

- Tal medida de similitud se determina para todos los pares de términos de la base de datos, formándose la matriz de similitud.
- Una vez que se dispone de la matriz de similitud se asocian los términos usando alguno de los métodos de agrupamiento conocidos:
  - Puesto que son pocos los términos relacionados, la mayoría de los valores de similitud son ceros. Por tanto, serán útiles las técnicas clásicas de manipulación y almacenamiento de matrices escasas.
  - Si se usa un **valor de umbral** de similitud de **0.6**, la mayoría de los agrupamientos que se forman son correctos y en casi ningún caso se producen asociaciones falsas.

N-Gramas – Bigramas (cont.)

- Ejemplo:
  - La tarea es determinar el *stem* de cada palabra del siguiente corpus:
    - escolar => es sc co ol la ar = 6
    - escuela => es sc cu ue el la = 6
    - escuelas => es sc cu ue el la as = 7
    - estado => es st ta ad do = 5
    - escuelilla => es sc cu ue el li il ll la = 9
    - preescolar => pr re ee es sc co ol la ar = 9

N-Gramas – Bigramas (cont.)

- Ejemplo:
  - $S_{12} = 2*3/(6+6) = 1/2$
  - $S_{13} = 2*3/(6+7) = 6/13$
  - $S_{14} = 2*1/(6+5) = 2/11$
  - $S_{15} = 2*3/(6+9) = 6/15 = 2/5$
  - $S_{16} = 2*6/(6+9) = 12/15 = 4/5$
  - $S_{23} = 2*6/(6+7) = 12/13$
  - $S_{24} = 2*1/(6+5) = 2/11$
  - $S_{25} = 2*6/(6+9) = 12/15 = 4/5$
  - $S_{26} = 2*3/(6+9) = 6/15 = 2/5$

N-Gramas – Bigramas (cont.)

- Ejemplo:
  - $S_{34} = 2*1/(7+5) = 2/12 = 1/6$
  - $S_{35} = 2*6/(7+9) = 12/16 = 3/4$
  - $S_{36} = 2*3/(7+9) = 6/16 = 3/8$
  - $S_{45} = 2*1/(5+9) = 2/14 = 1/7$
  - $S_{46} = 2*1/(5+9) = 2/14 = 1/7$
  - $S_{56} = 2*3/(9+9) = 6/18 = 1/3$

N-Gramas – Bigramas (cont.)

	escolar	**escuela**	**escuelas**	**estado**	**escuelilla**	**preescolar**
escolar	1	**1/2**	**6/13**	**2/11**	**2/5**	**4/5**
escuela	1/2	1	**12/13**	**2/11**	**4/5**	**2/5**
escuelas	6/13	12/13	1	**1/6**	**3/4**	**3/8**
estado	2/11	2/11	1/6	1	**1/7**	**1/7**
escuelilla	2/5	4/5	3/4	1/7	1	**1/3**
preescolar	4/5	2/5	3/8	1/7	1/3	1

- Las palabras escolar y preescolar forman un grupo, y su *stem* es escolar, sus valores dan más de 0,6.
- Las palabras escuela, escuelas y escuelilla forman un grupo, y su *stem* es escuel, sus valores dan más de 0,6.
- La palabra estado forman un grupo, y su *stem* es estado, sus valores dan más de 0,6.

N-Gramas – Bigramas (cont.)

- Ventajas:
  - Agrupa palabras similares, aún con prefijos y sufijos.
  - Es sencillo y fácil.
  - Si se usa un **valor de umbral** de similitud de **0.6**, la mayoría de los agrupamientos que se forman son correctos y en casi ningún caso se producen asociaciones falsas.
- Desventajas:
  - Aunque está incluido como método de lematización resulta un poco confuso porque no se produce tal lematización.

Errores Comunes

- *Over-Stemming*: Términos con diferentes significados son transformados a una misma raíz. Por ejemplo: universidad – universo.
- *Under-Stemming*: Términos con similar significado no son reducidos a una misma raíz. Por ejemplo: máquina – maquinaría.
- El ***over-stemming*** reduce la **precisión** y el ***under-stemming*** reduce el ***recall***.

Conclusiones

- La indexación se hace de forma más rápida.
- Se reduce el tamaño del índice hasta un 50%.
- Aumenta la eficiencia.
- En el Web no se usa, ya que existen muchos idiomas.

Ejercicio

- Dado el corpus de texto: colocar, colocación, coloso, vocación, evocación, gesto.
  - Utilizar variedad de sucesores con el método de picos y valles para obtener los diferentes grupos de palabras que se unen y el lema que las representa.
  - Utilizar *n*-gramas (usando bigramas) para obtener los diferentes grupos de palabras que se unen y el lema que las representa.

Referencias Bibliográficas

- La información fue tomada de:
  - Libro de texto del curso.
  - http://www.gedlc.ulpgc.es/docencia/seminarios/rit/.