

Prácticas de Algorítmica

Segunda parte del proyecto
coordinado SAR-ALT:
Búsqueda aproximada de cadenas

Curso 2021-2022



Objetivos



Objetivos

El objetivo de esta segunda parte del proyecto conjunto SAR-ALT es:

- 1 Crear un sistema eficiente para realizar la búsqueda aproximada de una cadena respecto de todas las cadenas del diccionario.

Como subtareas:

- A) Extender la distancia entre cadenas a distancia cadena-trie.
 - B) Medir experimentalmente qué algoritmo es mejor para unos datos concretos.
- 2 Utilizar esa búsqueda aproximada para ampliar el motor de recuperación de información desarrollado en SAR de forma que acepte consultas con búsqueda aproximada.



Búsqueda aproximada de cadenas

Búsqueda aproximada de cadenas

Una de las mejores referencias recomendadas para estudiar con detalle la búsqueda aproximada de cadenas es el artículo:

“Indexing Methods for Approximate Dictionary Searching: Comparative Analysis” (por Leonid Boytsov)

Disponible en <http://boytsov.info/pubs/jea2011.pdf>

Otra referencia, mucho más rudimentaria, es la página de wikipedia:

https://en.wikipedia.org/wiki/Approximate_string_matching

Vamos a introducir rápidamente el problema situando en contexto cuál es el objetivo concreto en este proyecto:

- La búsqueda puede ser **online** o bien **offline**: se llama **online** cuando el diccionario no puede ser preprocesado para generar algún tipo de índice que permita mejorar la eficiencia. Cuando sí podemos realizar algún tipo de preproceso se llama **offline** o **indexing**.

Vamos a suponer que sí se puede realizar un preproceso, estamos en el caso **offline**.

Búsqueda aproximada de cadenas

- Es posible orientar la búsqueda **a nivel de secuencia o de palabras**:
 - A nivel **de secuencia** alguien podría escribir “wiki pedia” en lugar de “wikipedia” y recuperarse del error (se consideran errores llamados *split* y *merge*).
 - A nivel **de palabras** se separa el texto por sus separadores y luego se limita a buscar las palabras más cercanas. En el ejemplo anterior buscaría las palabras cercanas a “wiki” y a “pedia”. En este caso se trabaja con un diccionario de palabras.

Nosotros vamos a limitarnos a trabajar a nivel de palabra.

- Otro aspecto a tener en cuenta cuando utilizamos un diccionario es saber si éste va a ser actualizado de manera frecuente o no. Si casi nunca se actualiza no será importante la eficiencia o el coste de la parte de preprocesado. Vamos a suponer que el diccionario **no se actualiza** (en caso de querer actualizarlo, podemos regenerar las estructuras de datos desde cero).
- También podemos distinguir entre búsquedas que tienen en cuenta el contexto (otras palabras que forman parte de dicha búsqueda) o no. Vamos a tratar con el caso más simple donde **no se tiene en cuenta el contexto**.

Búsqueda aproximada de cadenas

- Otro punto a tener en cuenta es si queremos que el sistema nos sugiera palabras mientras estamos escribiendo el término a buscar (búsqueda **predictiva**). El sistema debería completar una palabra a partir de su prefijo *incluso si el usuario se equivoca al escribir ese prefijo*. Únicamente buscaremos sugerencias de manera **no predictiva**.

Resumiendo: vamos a limitarnos a palabras ya escritas, que pueden contener errores ortográficos. Buscaremos las palabras candidatas en un diccionario que podremos haber preprocesado y que normalmente no va a ser modificado. Esta suele ser la tarea de un corrector ortográfico:

https://en.wikipedia.org/wiki/Spelling_suggestion

Existen bibliotecas python con esta funcionalidad, por ejemplo:
<http://pyenchant.github.io/pyenchant/api/enchant.html>

```
>>> import enchant
>>> d = enchant.Dict("en_US")    # create dictionary for US English
>>> d.suggest("enchnt")
['enchant', 'enchants', 'enchanter', 'penchant', 'incant', 'enchain', 'enchanted']
```



Búsqueda aproximada de cadenas

El objetivo es diseñar una clase Python que recibe un diccionario de términos/palabras, lo preprocesa y después ofrece un método `suggest` que recibe:

- La palabra a buscar.
- Un umbral o nivel de tolerancia (para limitar la búsqueda).

Y devuelve:

- Una lista de las palabras del diccionario que estén próximas a la buscada (dentro del umbral indicado).

Para ello necesitamos evaluar la distancia entre dos palabras. Vamos a limitarnos a estudiar **distancias de edición**:

La distancia de edición entre dos cadenas $x, y \in \Sigma^$ es el número mínimo de operaciones de edición para convertir la primera en la segunda (o viceversa).*

Las distancias de edición más utilizadas son:

- Distancia de Levenshtein.
- Distancia de Damerau-Levenstein.

Distancias de edición

Distancias de edición

Las operaciones de edición consideradas para calcular la distancia de Levenshtein entre x e y son:

- Insertar un carácter (denotado $\lambda \rightarrow y_j$)
- Borrar un carácter (denotado $x_i \rightarrow \lambda$)
- Sustituir un carácter por otro (denotado $x_i \rightarrow y_j$)

Aunque hemos definido la distancia de edición como el número de operaciones de edición, es posible considerar el caso **ponderado** donde las operaciones de edición pueden tener costes diferentes. Por ejemplo, tiene sentido que sustituir una consonante por otra diferente tenga un coste mayor a sustituir una vocal por la misma vocal acentuada. En el caso ponderado el coste de cada operación vendría dado por:

- $\gamma(\lambda \rightarrow y_j)$
- $\gamma(x_i \rightarrow \lambda)$
- $\gamma(x_i \rightarrow y_j)$

Distancias de edición

En este proyecto vamos a limitarnos al caso **no ponderado**. Es decir, aunque se pueda contemplar que algunas operaciones de distancia de edición tienen un coste mayor que otras, el coste de todas las inserciones o borrados es el mismo sin importar el símbolo del alfabeto. Por otra parte, todas las operaciones de sustitución dependen únicamente de si el símbolo que se sustituye es el mismo (un acierto, de coste cero) o es diferente (coste mayor que cero).

- $\gamma(\lambda \rightarrow y_j) = 1$

- $\gamma(x_i \rightarrow \lambda) = 1$

- $\gamma(x_i \rightarrow y_j) = \begin{cases} 0 & \text{si } x_i = y_j \\ 1 & \text{si } x_i \neq y_j \end{cases}$



Distancia de Levenshtein

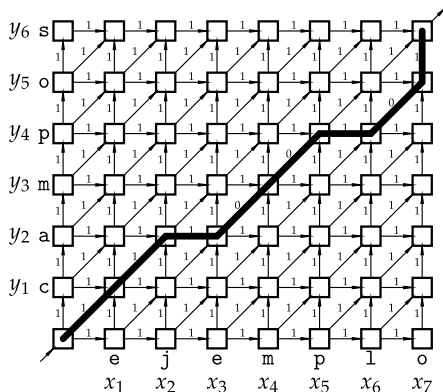
En la siguiente ecuación recursiva $D(i, j)$ denota el coste de convertir el prefijo de longitud i de x en el prefijo de longitud j de y :

$$D(i, j) = \min \begin{cases} 0 & \text{si } i=0 \text{ y } j=0 \\ D(i-1, j) + 1 & \text{si } i>0 \\ D(i, j-1) + 1 & \text{si } j>0 \\ D(i-1, j-1) & \text{si } i > 0, j > 0, x_i = y_j \\ D(i-1, j-1) + 1 & \text{si } i > 0, j > 0, x_i \neq y_j \end{cases}$$

De manera que $d(x, y) = D(|x|, |y|)$.

Distancia de Levenshtein

El siguiente gráfico ilustra la matriz de estados cuando se calcula la distancia de Levenshtein entre las cadenas *campos* y *ejemplo*:



Observa que todas las transiciones tienen coste 1 exceptuando las diagonales donde $x_i = y_j$.

Distancia de Levenshtein

Es más, con este sencillo ejemplo en Python (utilizando la biblioteca numpy) podemos mostrar los pesos de las diagonales para algunos pares de palabras:

```
def matriz(term, ref):  
    vref = np.array(list(ref))  
    return np.vstack([vref != letter for letter in term])+0  
  
for term,ref in [("campos","ejemplo"),  
                 ("reconocibles","irreconocible")]:  
    print(term,ref,levenshtein_dist(term,ref))  
    print(matriz(term,ref))
```

campos ejemplo 5.0

```
[[1 1 1 1 1 1 1]  
 [1 1 1 1 1 1 1]  
 [1 1 1 0 1 1 1]  
 [1 1 1 1 0 1 1]  
 [1 1 1 1 1 1 0]  
 [1 1 1 1 1 1 1]]
```

Distancia de Levenshtein

reconocibles irreconocible 3.0

```
[[1 0 0 1 1 1 1 1 1 1 1 1]
 [1 1 1 0 1 1 1 1 1 1 1 0]
 [1 1 1 1 0 1 1 1 0 1 1 1]
 [1 1 1 1 1 0 1 0 1 1 1 1]
 [1 1 1 1 1 1 0 1 1 1 1 1]
 [1 1 1 1 1 0 1 0 1 1 1 1]
 [1 1 1 1 0 1 1 1 0 1 1 1]
 [0 1 1 1 1 1 1 1 1 0 1 1]
 [1 1 1 1 1 1 1 1 1 1 0 1]
 [1 1 1 1 1 1 1 1 1 1 0 1]
 [1 1 1 0 1 1 1 1 1 1 1 0]
 [1 1 1 1 1 1 1 1 1 1 1 1]]
```

Se puede observar en general que, si no nos interesasen distancias muy altas, bastaría con calcular los valores cercanos a la diagonal principal.

Atención

En la figura anterior (sacada de los apuntes, se compara *campos* y *ejemplo*) la diagonal principal va de la esquina inferior izquierda a la superior derecha mientras que ahora va desde la superior izquierda.

Distancia de Damerau-Levenshtein

Se trata de una extensión de la distancia de Levenshtein en la que se añade un nuevo tipo de operación de edición:

- Trasponer o intercambiar dos símbolos adyacentes $ab \leftrightarrow ba$. Con la notación de editar x para obtener y sería $x_{i-1} = y_j$ y $x_i = y_{j-1}$ para $i > 0, j > 0$.

Una observación **obvia** es que, como incluye las operaciones de edición de la distancia de Levenshtein, se cumplirá siempre que:

$$\text{Damerau-Levenshtein}(x, y) \leq \text{Levenshtein}(x, y) \quad \forall x, y \in \Sigma^*$$

Por ejemplo, la distancia de Levenshtein entre “algoritmo” y “algoritmo” es 2 (dos sustituciones “i” \rightarrow “t”, “t” \rightarrow “i”) mientras que la distancia de Damerau-Levenshtein es 1 (intercambio “it” \leftrightarrow “ti”).

Distancia de Damerau-Levenshtein

Existen 2 variantes de esta distancia:

- La versión **no restringida** es la distancia como se entiende de la definición: buscar la forma de aplicar el menor número de operaciones para pasar de una cadena a la otra.
- En la versión **restringida** vamos a suponer que una vez intercambiados dos símbolos, éstos **no** se pueden utilizar en otras operaciones de edición.

Resulta que la versión restringida:

- 1 A veces no da la distancia mínima.
Por ejemplo, $d("ba", "acb")$ es 3 en la versión restringida pero sólo 2 en la versión no restringida.
- 2 No cumple la desigualdad triangular, por lo que no es una métrica.
Ejemplo: $d("ca", "ac") + d("ac", "abc") < d("ca", "abc")$
Esto hace que no pueda utilizarse en algunas circunstancias donde esa propiedad es necesaria.
- 3 Pero es más sencilla y rápida de calcular que la no restringida.

Distancia de Damerau-Levenstein restringida

El cálculo de la versión restringida de Damerau-Levenstein mediante programación dinámica es relativamente sencillo, basta con utilizar la siguiente ecuación recursiva:

$$D(i, j) = \min \begin{cases} 0 & \text{si } i=0 \text{ y } j=0 \\ D(i-1, j) + 1 & \text{si } i>0 \\ D(i, j-1) + 1 & \text{si } j>0 \\ D(i-1, j-1) & \text{si } i > 0, j > 0, x_i = y_j \\ D(i-1, j-1) + 1 & \text{si } i > 0, j > 0, x_i \neq y_j \\ D(i-2, j-2) + 1 & \text{si } i > 1, j > 1, x_{i-1} = y_j, x_i = y_{j-1} \end{cases}$$

Observa que simplemente añade un caso más (última línea) a la ecuación recursiva de la distancia de Levenshtein.

Distancia de Damerau-Levenstein no restringida

Si en la versión restringida hemos supuesto que $ab \rightarrow ba$ tiene coste 1, para la versión no restringida bastaría con suponer un conjunto de operaciones:

$aub \rightarrow bva$ tiene coste $1 + |u| + |v|$ para cualquier $u, v \in \Sigma^*$.

Explicación: para obtener bva a partir de aub **condicionado a que exista una transposición** $ab \rightarrow ba$ deberíamos:

- 1 Borrar u en aub , lo cual tiene coste $|u|$ borrados,
- 2 La transposición $ab \rightarrow ba$ con coste 1 y, finalmente,
- 3 Insertar v en ba , con coste $|v|$ inserciones.

Observa que una alternativa a pasar de aub a bva sin $ab \rightarrow ba$ sería:

- 1 Sustituir $a \rightarrow b$ (i.e. $aub \rightarrow bub$),
- 2 Obtener v a partir de u (i.e. $bub \rightarrow bvb$),
- 3 Sustituir $b \rightarrow a$ (i.e. $bvb \rightarrow bva$).

Esta alternativa tiene coste $2 + d(u, v)$, por lo que sólo conviene la transposición si $1 + |u| + |v| < 2 + d(u, v)$ (i.e. si $d(u, v) > |u| + |v| - 1$). A medida que crecen $|u|$ y $|v|$ será más improbable que valga la pena considerar las transposiciones.

Distancia de Damerau-Levenstein “intermedia”

Es posible implementar la distancia de Damerau-Levenstein no restringida con contadores de talla $|\Sigma|$, pero es un poco complejo.

Vamos a basarnos en la *intuición* de que a medida que $|u|$ y $|v|$ crecen es más y más improbable que tenga sentido aplicar una transposición $a \leftrightarrow b$ en $aub \rightarrow bva$ para proponer una solución que llamaremos *intermedia* y que consiste en considerar únicamente los casos donde $|u| + |v| \leq \text{cte}$ para una constante cte prefijada. En esta práctica consideraremos $\text{cte} = 1$ de modo que únicamente consideraremos las siguientes operaciones de edición donde $a, b, c, d \in \Sigma$:

- $ab \leftrightarrow ba$ coste 1
- $acb \leftrightarrow ba$ coste 2
- $ab \leftrightarrow bca$ coste 2

Observa...

Observa que no vale la pena contemplar la siguiente operación de edición:

- $acb \leftrightarrow bda$ coste 3

Puesto que ese mismo coste se consigue sin necesidad de una trasposición.

Tareas a realizar en el proyecto

Tareas a realizar en el proyecto

- 1 Implementar una serie de distancias de edición entre cadenas en su versión básica.
- 2 Implementar estas distancias de alguna forma “mejorada” que reciba un umbral `threshold` de modo que se pueda dejar de calcular cualquier distancia que supere dicho umbral. Si la distancia calculada es mayor a `threshold` el método devolverá `threshold+1` (incluso al final).
- 3 Evitar la ejecución de la distancia de edición entre aquellas cadenas en las que algún tipo de cota optimista nos permita saber que dicha distancia será mayor al umbral establecido.
- 4 Extender esta distancia, originalmente definida entre pares de cadenas, para calcular la distancia entre una cadena y un **trie**.
- 5 Realizar un estudio experimental de medidas de tiempo para determinar qué versiones de las anteriores son las más eficientes para unos datos concretos.
- 6 Modificar el Indexador y el Recuperador de SAR para permitir la búsqueda aproximada de cadenas utiliznado las versiones más eficientes según el punto anterior.



Tareas a realizar en el proyecto (tarea 1)

1 Implementar una serie de distancias de edición entre cadenas.

Debes implementar, utilizando programación dinámica:

- Distancia de Levenshtein (inserción, borrado y sustitución tienen coste 1).
- Distancia de Damerau-Levenstein restringida.
- Distancia de Damerau-Levenstein “intermedia” (es decir, considerar las operaciones de edición $aub \rightarrow bva$ cuando $|u| + |v| \leq cte$) para $cte = 1$.

Muchos de estos algoritmos ya están disponibles en los apuntes o en wikipedia. El objetivo es tener una referencia para asegurarse de que el resto de las implementaciones dan el mismo resultado y para, posteriormente, comparar la eficiencia con respecto a estas versiones.

Opcional

De manera **opcional** puedes implementar la versión general de Damerau-Levenstein.



Tareas a realizar en el proyecto (tarea 1)

Para comprobar la correcta realización de esta tarea se os proporciona un fichero `test_tarea1_plantilla.py` que lanza unos tests.

El resultado esperado de los mismos es:

algoritmo	algortimo	levenshtein	2	restricted	1	intermediate	1
algoritmo	algortximo	levenshtein	3	restricted	3	intermediate	2
algoritmo	lagortimo	levenshtein	4	restricted	2	intermediate	2
algoritmo	agaloritom	levenshtein	5	restricted	4	intermediate	3
algoritmo	algormio	levenshtein	3	restricted	3	intermediate	2
acb	ba	levenshtein	3	restricted	3	intermediate	2

Tareas a realizar en el proyecto (tarea 2)

- 2 Implementar estas distancias de alguna forma “mejorada” que reciba un umbral `threshold` de modo que se pueda dejar de calcular cualquier distancia que supere dicho umbral. Si la distancia calculada es mayor a `threshold` el método devolverá `threshold+1` (incluso al final).

Se trata de:

- A) Establecer límites en el recorrido para que solamente se calculen aquellas partes del grafo de dependencias que tengan sentido para dicho umbral. Por ejemplo: zonas relativamente cercanas a la diagonal principal de la matriz.
- B) Detener el algoritmo si, tras calcular una etapa (fila o columna según sea tu algoritmo) se puede asegurar que el coste superará el umbral.

Debes implementar las mismas variantes que antes:

- Distancia de Levenshtein (inserción, borrado y sustitución tienen coste 1).
- Distancia de Damerau-Levenstein restringida.
- Distancia de Damerau-Levenstein “intermedia” (es decir, considerar las operaciones de edición $aub \rightarrow bva$ cuando $|u| + |v| \leq cte$) para $cte = 1$.

Nota: En cualquier caso, al final hay que devolver le mínimo entre la distancia calculada y `threshold+1`.

Tareas a realizar en el proyecto (tarea 2)

Se os proporciona un fichero `test_tarea2_plantilla.py`. El resultado esperado es:

```
thresholds: 1
algoritmo   algortimo      levenshtein 2 restricted 1 intermediate 1
algoritmo   algortximo     levenshtein 2 restricted 2 intermediate 2
algoritmo   lagortimo      levenshtein 2 restricted 2 intermediate 2
algoritmo   agaloritom     levenshtein 2 restricted 2 intermediate 2
algoritmo   algormio       levenshtein 2 restricted 2 intermediate 2
acb         ba             levenshtein 2 restricted 2 intermediate 2
thresholds: 2
algoritmo   algortimo      levenshtein 2 restricted 1 intermediate 1
algoritmo   algortximo     levenshtein 3 restricted 3 intermediate 2
algoritmo   lagortimo      levenshtein 3 restricted 2 intermediate 2
algoritmo   agaloritom     levenshtein 3 restricted 3 intermediate 3
algoritmo   algormio       levenshtein 3 restricted 3 intermediate 2
acb         ba             levenshtein 3 restricted 3 intermediate 2
thresholds: 3
algoritmo   algortimo      levenshtein 2 restricted 1 intermediate 1
algoritmo   algortximo     levenshtein 3 restricted 3 intermediate 2
algoritmo   lagortimo      levenshtein 4 restricted 2 intermediate 2
algoritmo   agaloritom     levenshtein 4 restricted 4 intermediate 3
algoritmo   algormio       levenshtein 3 restricted 3 intermediate 2
acb         ba             levenshtein 3 restricted 3 intermediate 2
```



Tareas a realizar en el proyecto (tarea 3)

En esta etapa, para calcular el conjunto de palabras del diccionario que son próximas ortográficamente a un término escrito por el usuario no tendremos más remedio que calcular la distancia entre dicho término y **cada una** de las palabras del diccionario.

Una vez completada la tarea 2 debes implementar una clase Python que reciba una referencia del diccionario y que tenga un método `suggest` con el que obtener aquellas palabras del diccionario a una distancia menor que `threshold` del término buscado.

Para generar un diccionario con el que hacer pruebas se ha dejado el fichero `"quijote.txt"`. Se trata de: cargar ese fichero, limpiar el texto (convertir a minúsculas y tokenizar) y extraer el vocabulario (palabras únicas). Para hacer esto, se os proporciona el siguiente código, con el objetivo de que todos los alumnos compartáis el mismo preproceso y obtener resultados comparables:

```
def build_vocab(vocab_file):  
    clean_re = re.compile('W+')  
    with open(vocab_file, "r") as fr:  
        vocab = set(clean_re.split(fr.read()).lower())  
    return vocab
```

Tareas a realizar en el proyecto (tarea 3)

Para que evaluéis vuestras implementaciones de las distancias de edición, se os proporcionan los ficheros *result_levenshtein_quijote.txt*, *result_restricted_damerau-levenshtein_quijote.txt* y *result_intermediate_damerau-levenshtein_quijote.txt*, cuyas líneas contienen el término buscado, el threshold usado, el número de resultados obtenido tras la búsqueda y una enumeración de las palabras más similares ortográficamente al término buscado. Cada uno de los elementos está separado del resto mediante tabuladores. A continuación se muestran dos líneas de ejemplo extraídas del fichero *result_restricted_damerau-levenshtein_quijote.txt*:

```
casa      1    31  0:casa 1:basa 1:caba 1:cada 1:caja 1:cala 1:cama ...
quixot    2     6  2:quieto 2:quijo 2:quijote 2:quinto 2:quiso 2:quito
```

Los términos que incluyen los ficheros son: “casa”, “senor”, “jabón”, “constitución”, “savaedra”, “vicios”, “quixot”, “s3afg4ew” y “ancho”, con thresholds desde 1 hasta 5. No es necesario que automaticéis la evaluación (aunque sí recomendable), es suficiente con probar a mano varios de estos ejemplos con varios thresholds y ver que los resultados coinciden con los ficheros de resultados proporcionados.

Tareas a realizar en el proyecto (tarea 3)

Está claro que comparar una cadena contra todas las cadenas de un diccionario es muy **costoso**, por eso en la tarea 3, además de crear una clase con el método `suggest`, contempla:

- 3 Evitar la ejecución de la distancia de edición entre aquellas cadenas en las que algún tipo de cota optimista nos permita saber que dicha distancia será mayor al umbral establecido.

Se trata de evitar, cuando se buscan sugerencias a un término dado, el cálculo de algunas distancias de edición. Obviamente, el cálculo que determine si esa distancia puede ser evitada deberá ser mucho más rápido que la propia distancia, ya que en otro caso sería contraproducente.

De manera **obligatoria** se debe probar la diferencia entre las longitudes de ambas cadenas. Ver el siguiente enlace ([pinchar aquí](#)).

Tareas a realizar en el proyecto (tarea 3)

De manera **optativa** (ampliación) y *únicamente para Levenshtein* se puede probar la siguiente cota optimista basada en contar el nº de veces que aparece cada letra en cada palabra. Es una cota optimista porque se ha perdido el orden en el que aparecen las letras (como ocurre también con los modelos *bag of words* que puede que vieras en SAR). Veamos un ejemplo:

- Si se compara "casa" con "saca", el nº veces que aparece cada letra es el mismo, así que una cota basada en el conteo devolvería 0.
- Si se compara "casa" con "saco", vemos que hay una "a" de más en la primera y una "o" de más en la segunda. Una cota optimista basándonos en el conteo (ignorando el orden) sólo podría concluir que la distancia es ≥ 1 por una posible sustitución de la "a" por la "o".
- En general, si sacáramos en un vector v_x el nº veces que aparece cada letra en la cadena x y en v_y el nº veces que aparece cada letra en la cadena y , podríamos obtener una cota optimista restando $v = v_x - v_y$ y luego sumando los elementos positivos de v , por una parte, y los negativos, por otra. La cota sería $\max(\text{positivos}, -\text{negativos})$. En numpy los positivos son `np.sum(v[v>0])`.

Tareas a realizar en el proyecto (tarea 3)

(continuación de la explicación de la ampliación optativa)

- Observa que, a pesar de cierta similitud con la distancia (L1 o Manhattan) entre ambos vectores **NO** se trata de lo mismo, por ejemplo, la distancia entre "casa" y "abad" nos genera estos 2 vectores (asumimos que los índices de los vectores corresponden a las letras a, b, c, d y s):

- Para $x = \text{"casa"}$ sería $v_x = [2, 0, 1, 0, 1]$.

- Para $y = \text{"abad"}$ sería $v_y = [2, 1, 0, 1, 0]$.

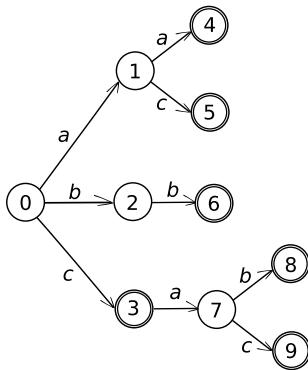
La resta sería $v_x - v_y = [0, -1, 1, -1, 1]$ (los positivos suman 2, los negativos suman -2). La norma L1 es 4, pero **no sería una cota optimista** (la distancia de Levenshtein es 3). Sin embargo, si nos quedamos con el máximo entre la suma de positivos (suman $1+1=2$) y -negativos (los negativos suman -2) obtenemos $\max(2, 2)=2$ que sí es una cota optimista.

- Se aconseja trabajar directamente con todas las letras que vayan apareciendo haciendo uso de diccionarios python (en lugar de vectores).

Tareas a realizar en el proyecto (tarea 4)

- 4 Extender esta distancia, originalmente definida entre pares de cadenas, para calcular la distancia entre una cadena y un **trie**.

Un *trie* es un árbol de prefijos que permite representar un conjunto de cadenas. El siguiente trie que representa las cadenas ["aa", "ac", "bb", "c", "cab", "cac"]:





Tareas a realizar en el proyecto (tarea 4)

- Observa que todo prefijo de cualquier cadena aparece como un nodo o estado del *trie* incluyendo el prefijo vacío que es la raíz.
- Salvo el nodo raíz, a todo nodo del *trie* se llega con una única arista que está etiquetada con un símbolo del alfabeto. De alguna manera podemos asumir que ese símbolo podría ser un atributo de dicho nodo.
- Aquellos estados/prefijos que corresponden a una palabra aparecen marcados como finales. Vemos estados finales internos (no hojas) cuando una palabra puede ser prefijo de otra.
- Un *trie* puede representar un diccionario de una manera posiblemente más compacta que guardando las cadenas por separado.

*¿Cuánto más compacta? Depende de hasta qué punto esas cadenas comparten prefijos. De hecho, a veces podríamos plantearnos hacer el *trie* con las cadenas al revés (para que compartan sufijos).*



Tareas a realizar en el proyecto (tarea 4)

¿Cómo podemos representar una clase `Trie` en Python?

Lo primero que podemos plantearnos sería crear una clase `TrieNode` con campos como:

- Etiqueta o símbolo del alfabeto con el que se llega a dicho nodo.
- Lista de nodos hijos.
- El nodo padre del que proviene.
- Si es un estado terminal o no.

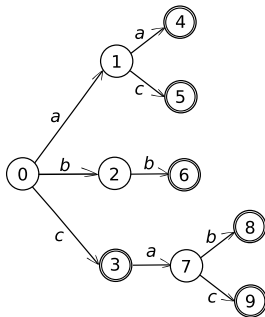
Con esto, la clase `Trie` tendría como principal atributo un `TrieNode` asociado al nodo raíz.

El problema de esta representación es que consumiría mucho espacio en memoria (comparado con la solución alternativa propuesta).

Vamos a proponer una representación más eficiente en espacio. Para ello observemos qué sucede si numeramos los nodos del trie siguiendo un **recorrido por niveles**.

Tareas a realizar en el proyecto (tarea 4)

El siguiente trie que representa las cadenas ["aa", "ac", "bb", "c", "cab", "cac"] y tiene los estados numerados siguiendo un **recorrido por niveles**:



Los hijos de cada nodo ocupan **posiciones contiguas**. Esto nos permite poder recorrer los hijos de cualquier nodo sabiendo únicamente dónde está su primer hijo (observa que hay un estado ficticio nº 10 que sirve de centinela):

state	0	1	2	3	4	5	6	7	8	9	10
first_child	1	4	6	7	8	8	8	8	10	10	10

Tareas a realizar en el proyecto (tarea 4)

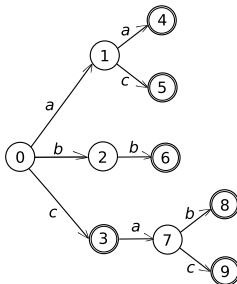
No necesitamos guardar el nº de hijos ni dónde está el último hijo puesto que después de él viene el primer hijo del siguiente nodo.

Para recorrer todos los hijos de un determinado nodo i basta con hacer:

```
# recorreremos los nodos hijos del nodo i
for child_of_i in range(self.first_child[i], self.first_child[i+1]):
    # process node child_of_i
```

Observa el valor `first_child` cuando un nodo es una hoja:

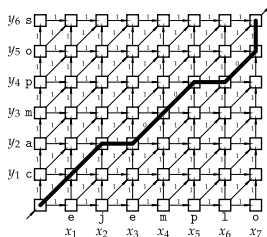
state	0	1	2	3	4	5	6	7	8	9	10
first_child	1	4	6	7	8	8	8	8	10	10	10



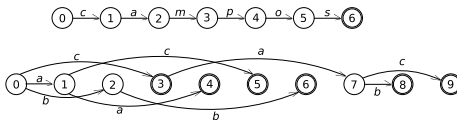
Tareas a realizar en el proyecto (tarea 4)

- 4 Extender esta distancia, originalmente definida entre pares de cadenas, para calcular la distancia entre una cadena y un **trie**.

Se trata sustituir una de las 2 cadenas por el Trie. Observa el grafo de dependencias de programación dinámica entre "campos" y "ejemplo":



Realmente la cadena "campos" es equivalente al siguiente autómata de $1 + \text{len}(\text{"campos"})$ estados que lo sustituimos por el Trie:



Tareas a realizar en el proyecto (tarea 4)

Por tanto:

- **Antes:** cada estado de la matriz de programación dinámica corresponde a haber alcanzado un prefijo en la cadena "campos" y a otro prefijo en "ejemplo".
- **Ahora:** cada estado de la matriz de programación dinámica corresponde a haber llegado a un nodo del trie y a haber alcanzado un determinado prefijo de la cadena "ejemplo".

La salida de dicho método será un diccionario que a cada palabra de salida le asocie su distancia. Para ello hay que observar la última columna de la matriz de programación dinámica y ver qué estados del Trie son finales y si la distancia es menor que el `threshold`. Mismo resultado que el método `suggest` de la tarea 3. De manera **obligatoria** debes implementarlo para Levenshtein utilizando la técnica *hacia atrás* o *backwards* (la que se utiliza normalmente en teoría y que consiste en aplicar la transformación recursiva-iterativa a las ecuaciones de recurrencia). En ese caso puedes utilizar una matriz o bien 2 vectores de tipo `numpy` si aplicas la técnica de reducción del coste espacial.

Tareas a realizar en el proyecto (tarea 4)

Consejos para implementar la distancia de Levenshtein cadena-Trie:

- Cada etapa corresponde a procesar un carácter de la cadena.
- Inicialmente distancia es 0 en root y el resto se calcula recorriendo los estados del Trie (antes la 1a columna recorría la longitud de la cadena en vertical).
*Es **imprescindible** recorrer los estados del Trie de menor a mayor, ya que ese recorrido garantiza que se procesan los padres antes de los hijos.*
- Las **inserciones** de caracteres en la cadena corresponde a avanzar por el trie **en la misma etapa** por lo que has de utilizar el método `get_parent` y valores del vector que representa el estado actual. Antes, en lugar de `get_parent` se miraba el índice anterior.
*Es **imprescindible** recorrer los estados del Trie de menor a mayor, ya que ese recorrido garantiza que se procesan los padres antes de los hijos.*
- Los **borrados** de caracteres en la cadena corresponde a quedarse en el mismo estado del Trie (en lugar de ir a un hijo), lo que corresponde a “copiar sumándole 1” el valor del vector en la misma posición pero del vector de la iteración anterior.

Tareas a realizar en el proyecto (tarea 4)

- Las sustituciones/aciertos corresponden a avanzar al mismo tiempo en la cadena y en el Trie, lo que corresponde a usar `get_parent` y comparar la etiqueta (`get_label`) con el caracter de la etapa actual.
- Si al terminar una etapa el menor valor (de toda la etapa) es $> \text{threshold}$ se podría parar el cálculo y devolver directamente $\text{threshold}+1$.
- La salida final se obtiene de los estados finales (método `is_final`) con distancia $\leq \text{threshold}$ y la cadena la da el método `get_output`.

En el fondo es cambiar las siguientes cosas en la versión cadena-cadena:

- El nº filas antes era la longitud(cadena)+1, ahora es el nº estados del Trie.
- Cuando miras una dependencia **vertical** en el grafo de dependencia antes venía de la letra anterior $[i-1, j]$, ahora con $[\text{trie.get_parent}(i), j]$.
- Cuando miras una dependencia **diagonal** en el grafo de dependencia era una sustitución que venía de $[i-1, j-1]$ y ahora será de $[\text{trie.get_parent}(i), j-1]$.
Para determinar si es sustitución o acierto hay que consultar `trie.get_label(i)`.

Ampliaciones opcionales en la tarea 4

- Versiones *hacia atrás* para Damerau-Levenshtein restringida e intermedia.



Tareas a realizar en el proyecto (tarea 5)

- 5 Realizar un estudio experimental de medidas de tiempo para determinar qué versiones de las anteriores son las más eficientes para unos datos concretos.

Se trata de diseñar un experimento empírico en el que se pruebe:

- Los diversos algoritmos implementados.
- Distintas tallas de diccionario.
- Distintos valores de *threshold*.
- Opcionalmente puedes generar errores ortográficos: tomar un conjunto de palabras del diccionario y modificarlas para que tengan algún error ortográfico, ya sea manualmente o mediante una función que lo haga de manera automática. Se aconseja que exista alguna trasposición y alguna trasposición antes o después de un borrado de un solo carácter para Damerau-Levenstein.

Tareas a realizar en el proyecto (tarea 5)

Se valorará la forma correcta de realizar esta experimentación, en particular:

- Puedes medir el tiempo llamando a `time.process_time()` antes y después de lo que quieras medir, si fuese algo demasiado rápido puedes repetir para ganar precisión (ver función `measure_time` en este mismo boletín).
- Repetir muestreos para una misma talla. Puedes reportar media, mediana y desviación típica si lo consideras conveniente.
- Variar las tallas del diccionario donde se busca, para ello se aconseja ordenar las palabras del corpus `quijote.txt` de mayor a menor frecuencia y utilizar las N primeras variando esa N .
- Utilizar **los mismos datos** con todos los algoritmos.

Tareas a realizar en el proyecto (tarea 5)

Cómo medir tiempos en Python

```
def measure_time(function, arguments,
                  prepare=dummy_function, prepare_args=()):
    """ mide el tiempo de ejecutar function(*arguments)
```

IMPORTANTE: como se puede ejecutar varias veces puede que sea necesario pasarle una función que establezca las condiciones necesarias para medir adecuadamente (ej: si mides el tiempo de ordenar algo y lo deja ordenado, la próxima vez que ordenes no estará desordenado)

DEVUELVE: tiempo y el valor devuelto por la función"""

```
count, accum = 0, 0
while accum < 0.1:
    prepare(*prepare_args)
    t_ini = time.process_time()
    returned_value = function(*arguments)
    accum += time.process_time()-t_ini
    count += 1
return accum/count, returned_value
```



Tareas a realizar en el proyecto (tarea 5)

Puedes ordenar el vocabulario por frecuencia (para variar la talla del vocabulario y utilizar las n palabras más frecuentes) así:

```
import collections
import re
vocab_file_path = "./corpora/quijote.txt"
tokenizer = re.compile("\W+")
with open(vocab_file_path, "r", encoding='utf-8') as fr:
    c = collections.Counter(tokenizer.split(fr.read()).lower())
    if '' in c:
        del c['']
    reversed_c = [(freq, word) for (word, freq) in c.items()]
    sorted_reversed = sorted(reversed_c, reverse=True)
    sorted_vocab = [word for (freq, word) in sorted_reversed]
```

Atención

Una vez elegidas las n primeras palabras de `sorted_vocab` debes ordenarlas lexicográficamente para pasárselas al constructor del `Trie`.

Tareas a realizar en el proyecto (tarea 6)

- 6 Modificar el Indexador y el Recuperador de SAR para permitir la búsqueda aproximada de cadenas utilizando las versiones más eficientes según el punto anterior.

Para facilitar esto, se ha pedido crear una clase `SpellSuggest` que representa el corrector ortográfico:

- Utiliza una distancia cadena-cadena y recorre el diccionario (utilizando quizás algún tipo de técnica para evitar algunas distancias).

De esta clase hereda la siguiente `TrieSpellSuggest`:

- Calcula al menos la distancia de Levenshtein usando el algoritmo *trie*-cadena, puede que hayáis implementado (era opcional) Damerau-Levenstein restringida y/o intermedia también con *trie*-cadena.

Añade en el recuperador de noticias de SAR la posibilidad de activar la búsqueda aproximada de manera que, cuando uno de los términos a buscar no se encuentre en el diccionario, se recuperen las noticias asociadas a los términos que sugiera el corrector.



Material proporcionado

Material proporcionado

Para la realización de este proyecto se proporciona:

- `test_tarea1_plantilla.py` y `test_tarea2_plantilla.py` para comprobar las distancias de edición entre diversas palabras (tareas 1 y 2).
- Fichero "`quijote.txt`" para realizar pruebas (tareas 3 en adelante).
- Propuesta de la clase `SpellSuggest` para integrar vuestro código (tareas 3 en adelante) y luego utilizar esa clase en el proyecto de SAR.
- Referencias para la tarea 3.
- El código de la clase `Trie` (para la tarea 4).
- Código para medir tiempos (para la tarea 5).
- Ficheros de referencia para comprobar que la integración con el proyecto de SAR es correcta.

Atención

Si alguien no cursó SAR en los 2 últimos cursos debe ponerse en contacto con su profesor de prácticas:

- Grupos del miércoles contactar con Lluís Hurtado lhurtado@inf.upv.es
- Grupos del viernes contactar con Salva España sespana@dsic.upv.es



Presentación y defensa del proyecto



Presentación y defensa del proyecto

Cada grupo/equipo debe preparar:

- Un **informe** con los resultados obtenidos y justificando las decisiones que ha adoptado. Debe ser en formato pdf (o ir acompañado de una versión pdf si se se utilizan otros formatos).
- El **código** utilizado. Se valorará la organización (incluyendo la utilización de módulos/bibliotecas), estilo, eficiencia, documentación/comentarios, etc.
- Preparar una **presentación muy breve** de los resultados. En dicha presentación deben intervenir todos los miembros del equipo y una pequeña **demo**.

Distribución de las sesiones

Distribución de las sesiones

- Sesión 1: Presentación + implementar distancia de edición básica (T1).
- Sesión 2: Implementar la distancia de edición básica entre dos cadenas (T1).
- Sesión 3: Implementar la distancia “mejorada” con umbral (T2) + incluir una cota optimista (T3).
- Sesión 4: Implementar de distancia de edición entre una cadena y un trie (T4).
- Sesión 5: Estudio experimental de los tiempos de ejecución de las distintas versiones (T5).
- Sesión 6: 6 Adaptar el código de SAR para incluir la nueva funcionalidad (T6).
- Sesión 7: Presentación.
- Sesión 8: Presentación.

Nota

La asignatura tiene 10 sesiones. Las 2 últimas sesiones se dedicarán a una práctica del tema “Ramificación y poda”.

Distribución de las sesiones

22 sep 2021	miércoles	Sesión 1	presentación + Tarea 1
24 sep 2021	viernes		
29 sep 2021	miércoles	Sesión 2	Tarea 1
1 oct 2021	viernes		
6 oct 2021	miércoles	Sesión 3	Tareas 2 y 3
8 oct 2021	viernes		
13 oct 2021	miércoles	Sesión 4	Tarea 4
15 oct 2021	viernes		
20 oct 2021	miércoles	Sesión 5	Tarea 5
22 oct 2021	viernes		
27 oct 2021	miércoles	Sesión 6	Tarea 6
29 oct 2021	viernes		
24 nov 2021	miércoles	Sesión 7	Presentación
26 nov 2021	viernes		
1 dic 2021	miércoles	Sesión 8	Presentación
2 dic 2021	viernes		