Nombre: Gabriel Muñoz Marcelo Callisaya
CI: 9873103

DAT 261 - Procesamiento del lenguaje natural

Tarea1.

1. Encontrar una forma de tokenizar una frase sílaba por sílaba.

```
[^()'"‘’¿¡aeiouAEIOUáéíóúüÁÉÍÓÚ\s]*(?(?=[íú][aeo])
[aeiouAEIOUáéíóüúÁÉÍÓÚ]|(?(?=[aeo]{2})[aeiouAEIOUáéíóüúÁÉÍÓÚ]|(?(?
=y)y|[aeiouAEIOUáéíóüúÁÉÍÓÚ]+)))(?(?=[dhlnmrsyz]\b)(?(?=.\b[áéíóú])|
\w)|)(?(?=[^aeiouAEIOUáéíóúüÁÉÍÓÚ\s]{2,})(?(?=ch|ll|rr|pr|br|tr|tl|
dr|cr|kr|gr|fr|pl|bl|cl|kl|gl|fl)|(?(?=[^aeiouAEIOUáéíóúüÁÉÍÓÚ\s]
{4})\w\w|\w))|)
```

En pseudocódigo:

```
atrapa consonantes seguidas
if ([íú][aeo]) {
  atrapa vocal
} else if ([aeo]{2}){
  atrapa vocal
} else if (hay una y sola){
  atrapa y
} else {
  atrapa todas las vocales seguidas
}

if (termina en dlnmrsyz){
  if (al terminar hay áéíóú){

  } else {
    atrapa letra
}

if (choque de dos o más consonantes){
   if (choque de ch | ll | rr | pr | br | tr | tl | dr | cr | kr | gr
| fr | pl | bl | cl | kl | gl | fl){

   } else {
    if (choque de cuatro consonantes) {
      atrapa dos letras
    } else {
      atrapa letra
    }
   }
}
```

Explicación:

Para separar por sílabas, primero se analiza cómo se construyen las palabras, se deben agarrar las consonantes con las vocales que les sigan, esto ya sirve para tokenizar palabras como `llave, palabra, pala`, se toma en cuenta que una palabra puede empezar por vocal, así que se usa el cuantificador * para indicar que las consonantes pueden aparecer cero o más veces `[^aeiouAEIOUáéíóúüÁÉÍÓÚ\s]*`.

Siguiendo las reglas de la RAE para los diptongos, se manejan todos los casos en los que se debe atrapar una o más vocales, esto se logra gracias a usa una función que replica el comportamiento de un `if`: `(?(?=condición)then|else)` Con esto, se añaden palabras como `ala, aloja, emana`.

Cuando una palabra termina en una consonante (en español solo pasa con `dlnmrsyz`), esa consonante se añade a la silaba final, para comprobar si una palabra termina en una consonante. Así, se logra atrapar la consonante final como parte de la ultima sílaba, ya que regex no considera las palabras con tilde como parte de una palabra, el posicionador `\b` no las toma en cuenta, así que para evitar atrapar letras que se deberían juntar con una vocal con tilde al final para formar la sílaba, se añade la condición de que no debe seguir ningún `áéíóú`:
`[^aeiouAEIOUáéíóúüÁÉÍÓÚ\s]*[aeiouyAEIOUYáéíóüúÁÉÍÓÚ]+(?(?=[dlnmrsyz]\b[^áéíóú])\w|)`.

Las situaciones en las que las sílabas no se dividen de forma `consonante - vocal` es cuando hay un choque de dos consonantes o más, en ese caso, también pueden tener la forma `vocal - consonante` o `vocal - consonante - vocal` (men-te, a-lam-bre, am-pa-ro), pero esta regla tiene excepciones con los choques de consonantes `ch, cl, cr, ll, tl, tr, dr, br, bl, rr`, en cuyo caso se corta la sílaba para añadir ambas consonantes a la siguiente sílaba; por eso, y usando las estructuras de if, cuando hay un choque de consonantes `(?(?=[^aeiouAEIOUáéíóúüÁÉÍÓÚ\s]{2,})then|else)` dentro del then se hace otro `if` para excluir los casos especiales `(?(?=ch|cl|cr|ll|tl|tr|dr|br|bl|rr)|else)`.

Una vez exluidos los casos especiales, hay otra observación antes de capturar la consonante extra para la estructura `consonante? - vocal - consonante`, en las únicas cuatro palabras del español que tienen cuatro consonantes seguidas (abstracto, abstraer, transplantar y substraer) se sigue la estructura `vocal - consonante - consonante`, para tratar estos casos específicos, se valida que el choque de consonantes no sea `ns` ni `bs`, si lo es, atrapa las dos consonantes para cumplir la estructura de la sílaba especial: `(?(?=ns|bs)\w\w|else))`. Finalmente, por parte del `else`, solo queda el caso en el que sí se siga la estructura `consonante? - vocal - consonante`, para la cual solo se tiene que capturar la consonante extra: `\w`.

Juntando todas las validaciones según la sintaxis de `(?(?=condición)then|else)`, se da con el código final:

```
[^()'"‘’¿¡aeiouAEIOUáéíóúüÁÉÍÓÚ\s]*(?(?=[íú][aeo])
[aeiouAEIOUáéíóüúÁÉÍÓÚ]|(?(?=[aeo]{2})[aeiouAEIOUáéíóüúÁÉÍÓÚ]|(?(?
=y)y|[aeiouAEIOUáéíóüúÁÉÍÓÚ]+)))(?(?=[dhlnmrsyz]\b)(?(?=.\b[áéíóú])|
\w)|)(?(?=[^aeiouAEIOUáéíóúüÁÉÍÓÚ\s]{2,})(?(?=ch|ll|rr|pr|br|tr|tl|
dr|cr|kr|gr|fr|pl|bl|cl|kl|gl|fl)|(?(?=[^aeiouAEIOUáéíóúüÁÉÍÓÚ\s]
{4})\w\w|\w))|)
```

■