Tokenizar

Como se puede observar en el diccionario existen palabras compuestas, por lo que al hacer tokenizacion normal se pierde el sentido de la expresión que esta escrita, por lo que el procedimiento es el siguiente:

Recorrer la frase completa e irle separando por ventanas (es decir grupos):

**Ejemplo**

**FRASE:** “yaku ishpa yalli jnk ninoi”

**Tamaño de la Frase:** 5

Se recorrería la frase por grupos es decir (grupo de 5 palabras, grupo de 4 palabras, grupo de 3 palabras, grupo de 2 palabras, grupo de 1 palabras)

**Grupo de 5:** (yaku ishpa yalli jnk ninoi)

**Grupo de 4:** (yaku ishpa yalli jnk) | (ishpa yalli jnk ninoi)

**Grupo de 3:** (yaku ishpa yalli) | (ishpa yalli jnk) | (yalli jnk ninoi)

**Grupo de 2:** (yaku ishpa) | (ishpa yalli) | (yalli jnk) | (jnk ninoi)

**Grupo de 1:** (yaku) | (ishpa) | (yalli ) | (jnk) | (ninoi)

Intervalo de posiciones que recorre:

**Grupo de 5:** 0-4

**Grupo de 4:** 0-3 | 1-4

**Grupo de 3:** 0-2 | 1-3 | 2-4

**Grupo de 2**: 0-1 | 1-2 | 2-3 | 3-4

**Grupo de 1:** 0 | 1 | 2 | 3 | 4

Una vez que se obtiene el recorrido por grupos de palabras podemos comparar con el diccionario y si existe dicha palabra en el diccionario se procese a unir.

**Ejemplo**

**Grupo de 2:** (yaku ishpa) | (ishpa yalli) | (yalli jnk) | (jnk ninoi)

**La palabra compuesta encontrada es “yaku ishpa”, por lo que en el vector de tokenizacion normal se procede a unir la palabra compuesta dando de resultado:**

FRASE\_TOKENIZADA\_CORRECTAMENTE = [“yaku ishpa”, “yalli”, “jnk”,”ninoi”]

**Codigo:**

**def toke(**

**conjunto: “Es la frase tokenizada normal”,**

**index: “indica el valor para saber el grupo de palabras que debe formar”,**

**etiquetas: “son los diccionarios POSITIVO,NEGATIVO,NEUTRO”**

**)**

**Funcionamiento:**

**Inicio: saber en q posición empieza el grupo**

**Fin: posición al q llega el grupo como se explico en el ejemplo arriba de posiciones**

**While: Recorre cada grupo de palabras formada**

**EJEMPLO (GRUPO DE 2 PALABRAS):** (yaku ishpa) | (ishpa yalli) | (yalli jnk) | (jnk ninoi)

1 2 3 4

“ ”.join para unir los grupos de palabras en un solo string,

Método FIND(etiqueta: “diccionarios..”,

Temp: “palabra unida por el .join”)

Devuelve la posición en donde se encontró la palabra o la palabra compuesta o -1 en el caso de q no se encontró en ningún diccionario.

If != -1 debe ser diferente de -1 porque quiere decir q si se encontró en algún diccionario, y esto “conjunto[inicio:fin+1] = [temp]” unifica la palabra en el vector, como se explico en FRASE\_TOKENIZADA\_CORRECTAMENTE