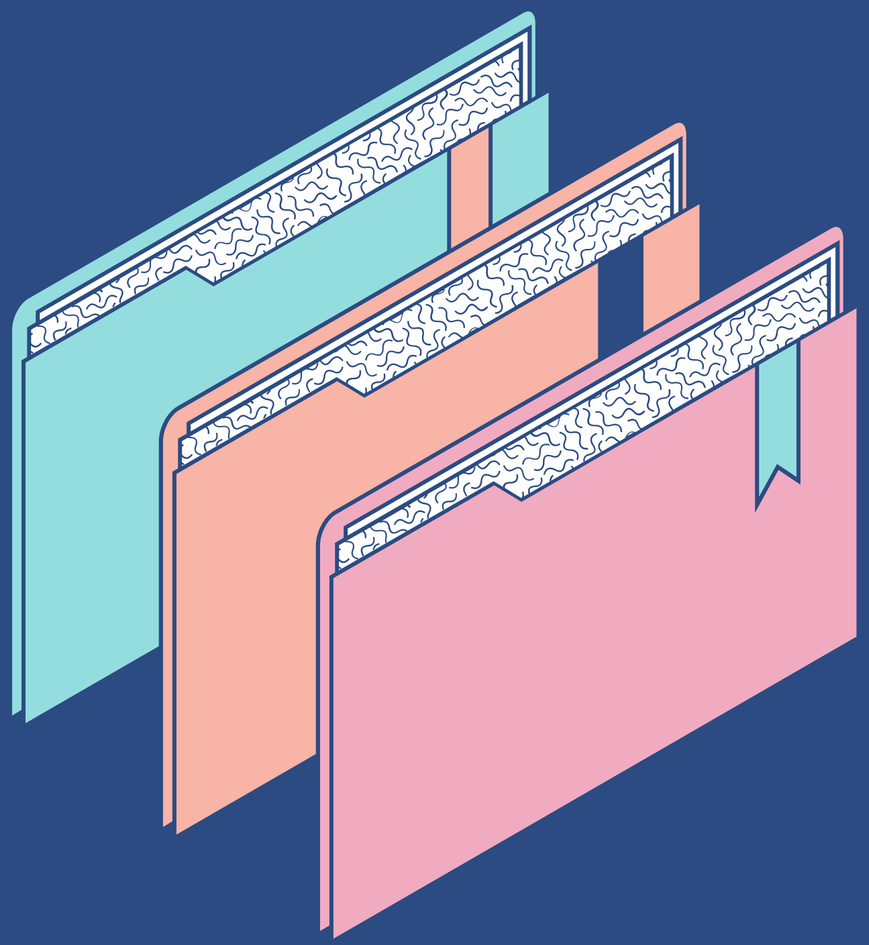
UNIVERSIDAD DE COSTA RICA
FACULTAD DE INGENIERÍAS
ESCUELA DE CIENCIAS DE LA COMPUTACIÓN E INFORMÁTICA

Grupo ZZZ:
Nathalie Alfaro Quesada - B90221
Diego Bolaños Villalobos - C21256
José Pablo Mora Cubillo - B75044

CI - 0124 Computabilidad y Complejidad
II Semestre, 2024



TAREA PROGRAMADA 1
AVANCE 2

Analizador léxico (Lexer)

```
# List of token names
tokens = (

    # XML Header
    'TAG_XML',
    'XML_TAG_CLOSURE',
    'ATTRIBUTE_VERSION',
    'ATTRIBUTE_ENCODING',

    # DOCTYPE tag
    'TAG_DOCTYPE',

    # Health-topics tag
    'TAG_HEALTH_TOPICS',
    'TAG_HEALTH_TOPICS_CLOSURE',
    'ATTRIBUTE_TOTAL',
    'ATTRIBUTE_DATE_GENERATED',

    # Health-topic tag
    'TAG_HEALTH_TOPIC',
    'TAG_HEALTH_TOPIC_CLOSURE',
    'ATTRIBUTE_ID',
    'ATTRIBUTE_DATE_CREATED',
    'ATTRIBUTE_LANGUAGE',
    'ATTRIBUTE_META_DESC',
    'ATTRIBUTE_URL',
    'ATTRIBUTE_TITLE',
```

```
    # Tags under <health-topic>
    'TAG_LANGUAGE_MAPPED_TOPIC',
    'TAG_LANGUAGE_MAPPED_TOPIC_CLOSURE',
    'TAG_FULL_SUMMARY',
    'TAG_FULL_SUMMARY_CLOSURE',
    'TAG_ALSO_CALLED',
    'TAG_ALSO_CALLED_CLOSURE',
    'TAG_SEE_REFERENCE',
    'TAG_SEE_REFERENCE_CLOSURE',
    'TAG_GROUP',
    'TAG_GROUP_CLOSURE',
```

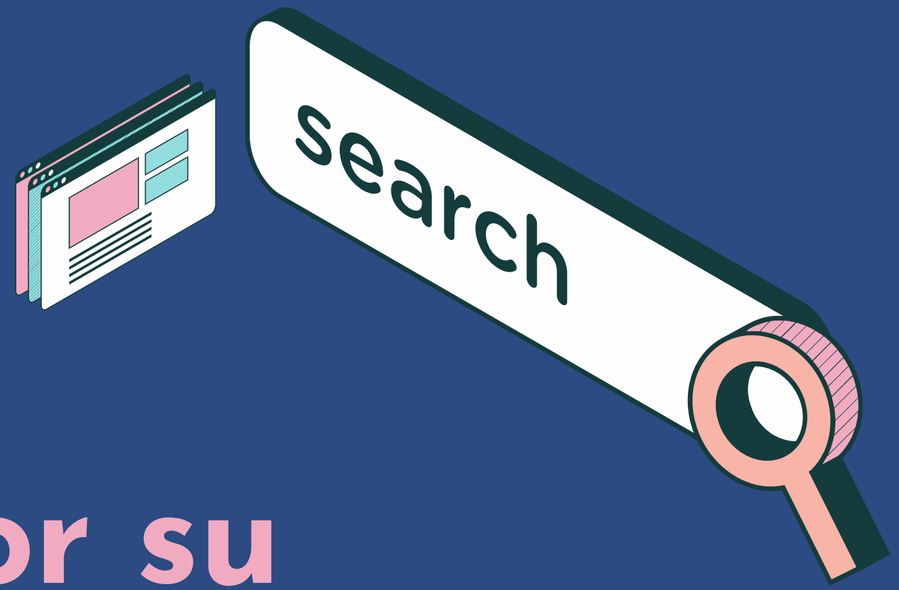
```
# Mesh heading tags
'TAG_MESH_HEADING',
'TAG_MESH_HEADING_CLOSURE',
'TAG_DESCRIPTOR',
'TAG_DESCRIPTOR_CLOSURE',

# Related topic tags
'TAG_RELATED_TOPIC',
'TAG_RELATED_TOPIC_CLOSURE',

# Other language tags
'TAG_OTHER_LANGUAGE',
'TAG_OTHER_LANGUAGE_CLOSURE',
'ATTRIBUTE_VERNACULAR_NAME',

# Primary institute tags
'TAG_PRIMARY_INSTITUTE',
'TAG_PRIMARY_INSTITUTE_CLOSURE',

# Site tags
'TAG_SITE',
'TAG_SITE_CLOSURE',
'TAG_INFORMATION_CATEGORY',
'TAG_INFORMATION_CATEGORY_CLOSURE',
'TAG_ORGANIZATION',
'TAG_ORGANIZATION_CLOSURE',
'TAG_STANDARD_DESCRIPTION',
'TAG_STANDARD_DESCRIPTION_CLOSURE',
'ATTRIBUTE_LANGUAGE_MAPPED_URL',
```

```
# HTML tags
'TAG_P',
'TAG_P_CLOSURE',
'ATTRIBUTE_CLASS',
'TAG_UL',
'TAG_UL_CLOSURE',
'TAG_LI',
'TAG_LI_CLOSURE',
'TAG_A_HREF',
'TAG_A_HREF_CLOSURE',
'TAG_H3',
'TAG_H3_CLOSURE',
'TAG_STRONG',
'TAG_STRONG_CLOSURE',

# '>' token
'TAG_CLOSURE',

# Tokens corresponding to texts
'TEXT_OF_ATTRIBUTE',
'TEXT_OF_TAG',
```

```
# Regular expression rules for Health topic and Health topics tags
t_TAG_HEALTH_TOPICS = r'<health-topics'
t_TAG_HEALTH_TOPICS_CLOSURE = r'</health-topics>'
t_TAG_HEALTH_TOPIC = r'<health-topic'
t_TAG_HEALTH_TOPIC_CLOSURE = r'</health-topic>'

# Regular expression rules for Mesh heading tags
t_TAG_MESH_HEADING = r'<mesh-heading>'
t_TAG_MESH_HEADING_CLOSURE = r'</mesh-heading>'
t_TAG_DESCRIPTOR = r'<descriptor'
t_TAG_DESCRIPTOR_CLOSURE = r'</descriptor>'
```

```
# Other language tags
def t_ATTRIBUTE_VERNACULAR_NAME(t):
    r'vernacular-name='
    return t
```

```python
# Define a rule so we can track line numbers
def t_newline(t):
    r'\n+'
    t.lexer.lineno += len(t.value)

# A string containing ignored characters (spaces and tabs)
"""
t_ignore matches a string containing ignored characters (spaces and tabs)
"""
t_ignore  = ' \t'

def t_error(t):
    """Error handling rule"""
    print(f"Illegal character '{str(t.value[0])}'")
    t.lexer.skip(1)

def tokenize(data):
    """Tokenize the input string and print each token."""
    # Build the lexer
    lexer = lex.lex()

    # Give the lexer some input
    lexer.input(data)

    # Tokenize
    while True:
        tok = lexer.token()
        if not tok:
            # No more input
            break
        print(tok)
```



Gracias por su atención