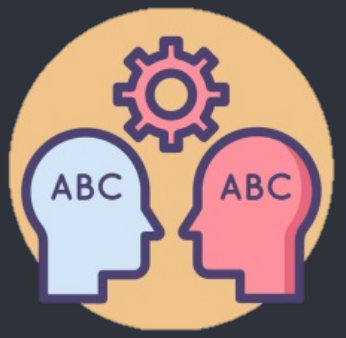
Procesamiento de Lenguaje Natural{



[NLP]

<N-Gramas>

}

Ivan Rojas Gonzalez – Universidad Panamericana

La Agenda de hoy {

01 N-Grams

<Que son y para que se usan?>

02 Bigrama, trigrama ... pentagrama?

<Analizando textos usando distintos valores de N>

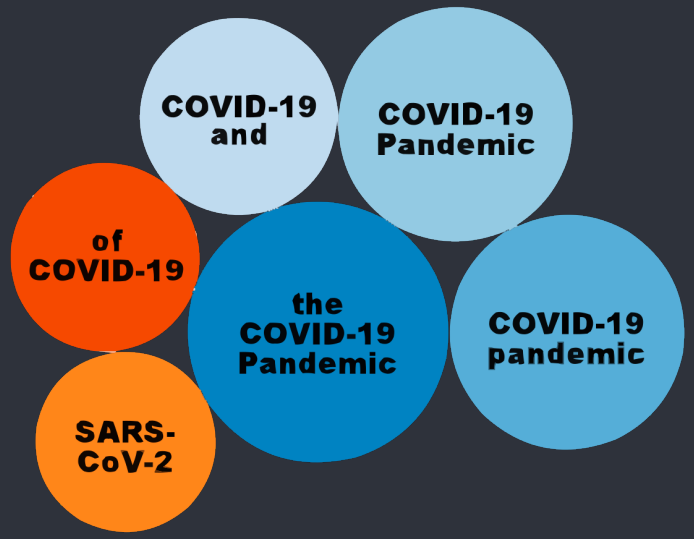
03 Visualizando N Gramas

<Nubes y contexto>

}

N-Gramas{

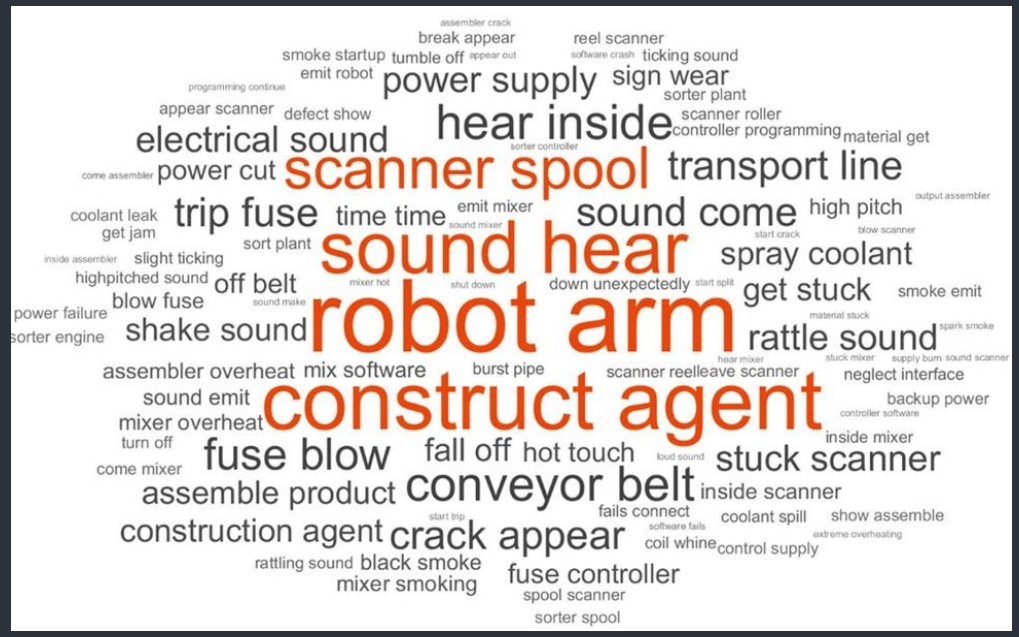
<Colección de n elementos sucesivos en un documento de texto . Son útiles en muchas aplicaciones de análisis de texto donde las secuencias de palabras son relevantes>



}

N-Gramas{

<Son utiles para preservar parte de la relacion o semantica de las palabras presentes en el texto>



}

N-Gramas{

}

Texto original

<"Un niño floto sobre mi y volo un auto con su rasho laser">

Tokenizacion{

<Dividir una oración en las palabras que la componen. >

```
texto = '''Mi Homero no es comunista. Podrá ser
          mentiroso, puerco, idiota, comunista,
          pero nunca una estrella de porno'''
```

Unigramas

```
texto.split()
✓ 0.1s

['Mi',
 'Homero',
 'no',
 'es',
 'comunista.',
 'Podrá',
 'ser',
 'mentiroso,',
 'puerco,',
 'idiota,',
 'comunista,',
 'pero',
 'nunca',
 'una',
 'estrella',
 'de',
 'porno']
```

N-Gramas{

}

Unigrama (N=1)

<(Un) (niño) (floto) (sobre) (mi) (y), (volo) (un) (auto) (con) (su) (rasho) (laser)>

N-Gramas{

}

Bigrama (N=2)

```
<('Un', 'niño') ('niño', 'floto')
('floto', 'sobre') ('sobre', 'mi')
('mi', 'y') ('y', 'volo') ('volo',
'un') ('un', 'auto') ('auto', 'con')
('con', 'su') ('su', 'rasho')
('rasho', 'laser')>
```

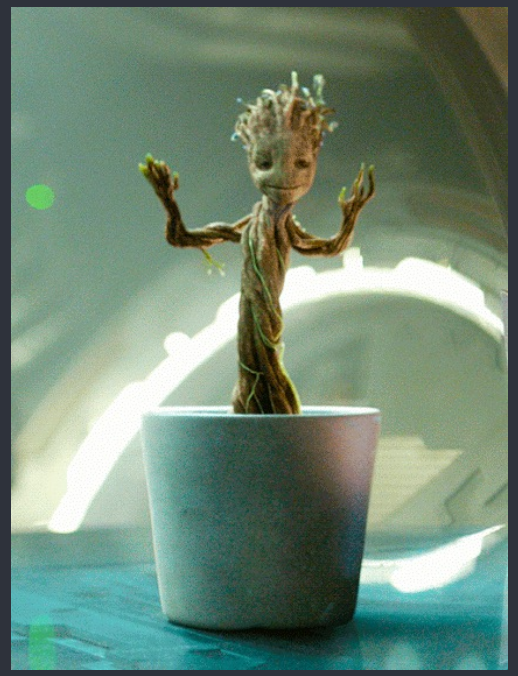
N-Gramas{

}

Trigrama (N=3)

```
<('Un', 'niño', 'floto') ('niño',
'floto', 'sobre') ('floto', 'sobre',
'mi') ('sobre', 'mi', 'y') ('mi', 'y',
'volo') ('y', 'volo', 'un') ('volo',
'un', 'auto') ('un', 'auto', 'con')
('auto', 'con', 'su') ('con', 'su',
'rasho') ('su', 'rasho', 'laser')>
```

N-Gramas{

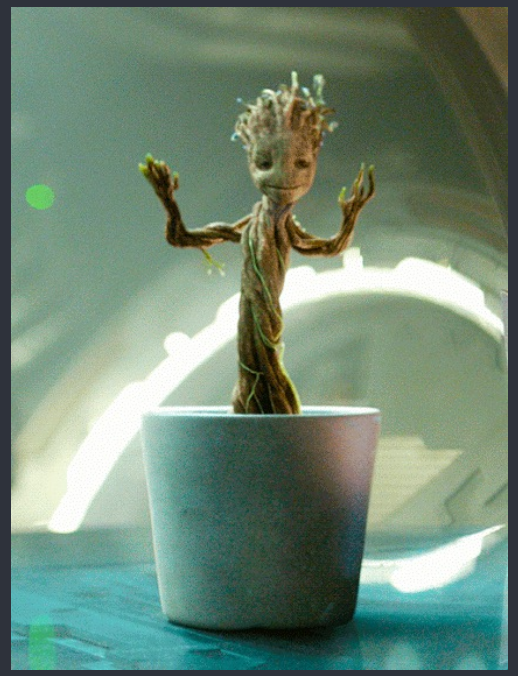


}

Bigrama (N=2)

```
('Yo', 'soy')
('soy', 'Groot')
```

N-Gramas{



}

Pentagrama (N=5)

Tokenizacion{

<Dividir una oración en las palabras que la componen. >

```
import nltk
nltk.download('punkt')
sentence = 'Mi hijo no es comunista podrá ser tonto, estúpido, inútil, comunista pero nunca una estrella porno'
tokens = nltk.word_tokenize(sentence)
print(tokens)
```

✓ 0.0s

```
['Mi', 'hijo', 'no', 'es', 'comunista', 'podrá', 'ser', 'tonto', ',', 'estúpido', ',', 'inútil', ',', 'comunista', 'pero', 'nunca', 'una', 'estrella', 'porno']
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\ivani\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

Las comas y puntuaciones se toman como su propio token

}

N-Gramas{

```
from nltk import ngrams
n = 2
n_grama = ngrams(tokens,n)
for grama in n_grama:
    print(grama)
```

✓ 0.0s

```
('Mi', 'hijo')
('hijo', 'no')
('no', 'es')
('es', 'comunista')
('comunista', 'podrá')
('podrá', 'ser')
('ser', 'tonto')
('tonto', ',')
(',', 'estúpido')
('estúpido', ',')
(',', 'inútil')
('inútil', ',')
(',', 'comunista')
('comunista', 'pero')
('pero', 'nunca')
('nunca', 'una')
('una', 'estrella')
('estrella', 'porno')
```

Bigrama

```
from nltk import ngrams
n = 3
n_grama = ngrams(tokens,n)
for grama in n_grama:
    print(grama)
```

✓ 0.0s

```
('Mi', 'hijo', 'no')
('hijo', 'no', 'es')
('no', 'es', 'comunista')
('es', 'comunista', 'podrá')
('comunista', 'podrá', 'ser')
('podrá', 'ser', 'tonto')
('ser', 'tonto', ',')
('tonto', ',', 'estúpido')
(',', 'estúpido', ',')
('estúpido', ',', 'inútil')
(',', 'inútil', ',')
('inútil', ',', 'comunista')
(',', 'comunista', 'pero')
('comunista', 'pero', 'nunca')
('pero', 'nunca', 'una')
('nunca', 'una', 'estrella')
('una', 'estrella', 'porno')
```

Trigrama

```
from nltk import ngrams
n = 6
n_grama = ngrams(tokens,n)
for grama in n_grama:
    print(grama)
```

✓ 0.0s

```
('Mi', 'hijo', 'no', 'es', 'comunista', 'podrá')
('hijo', 'no', 'es', 'comunista', 'podrá', 'ser')
('no', 'es', 'comunista', 'podrá', 'ser', 'tonto')
('es', 'comunista', 'podrá', 'ser', 'tonto', ',')
('comunista', 'podrá', 'ser', 'tonto', ',', 'estúpido')
('podrá', 'ser', 'tonto', ',', 'estúpido', ',')
('ser', 'tonto', ',', 'estúpido', ',', 'inútil')
('tonto', ',', 'estúpido', ',', 'inútil', ',')
(',', 'estúpido', ',', 'inútil', ',', 'comunista')
('estúpido', ',', 'inútil', ',', 'comunista', 'pero')
(',', 'inútil', ',', 'comunista', 'pero', 'nunca')
('inútil', ',', 'comunista', 'pero', 'nunca', 'una')
(',', 'comunista', 'pero', 'nunca', 'una', 'estrella')
('comunista', 'pero', 'nunca', 'una', 'estrella', 'porno')
```

Hexagrama

N-GRAMA

}

Q&A