# Generación de Resúmenes de Textos Médicos

Procesamiento de Lenguaje Natural 2020-1

Edwin Mahecha

Jimmy Pulido

UNIVERSIDAD NACIONAL DE COLOMBIA

# Contenido

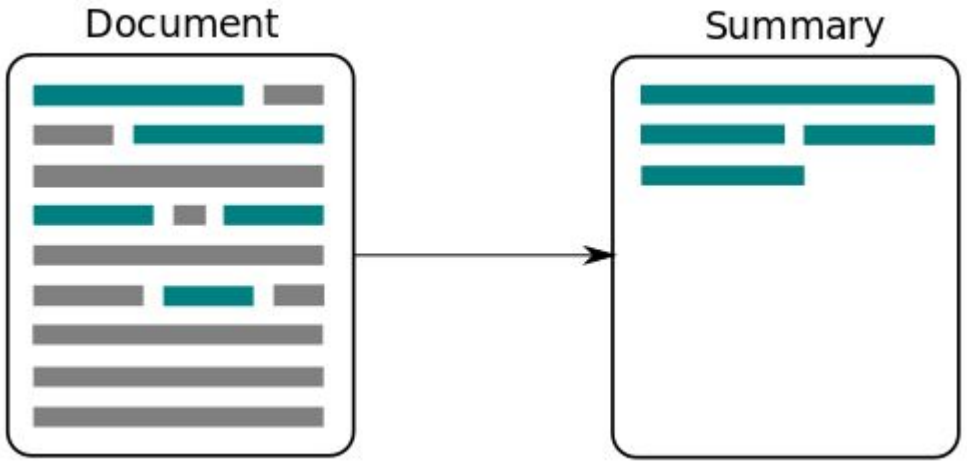
Sumarización de Textos Médicos Relacionados a **COVID-19**

1. Introducción
2. Objetivos
3. Alcance
4. Marco Teórico
5. Desarrollo
    a. Obtención de Corpus.
    b. Preprocesamiento
    c. Normalización
    d. Transformers & Attention Models
    e. Obtención de Resúmenes
6. API
7. Resultados

# Resumen de texto

Factores a tener en cuenta:

- Estructura del texto.
- Temática del texto.
- Palabras clave
- Términos específicos del contexto

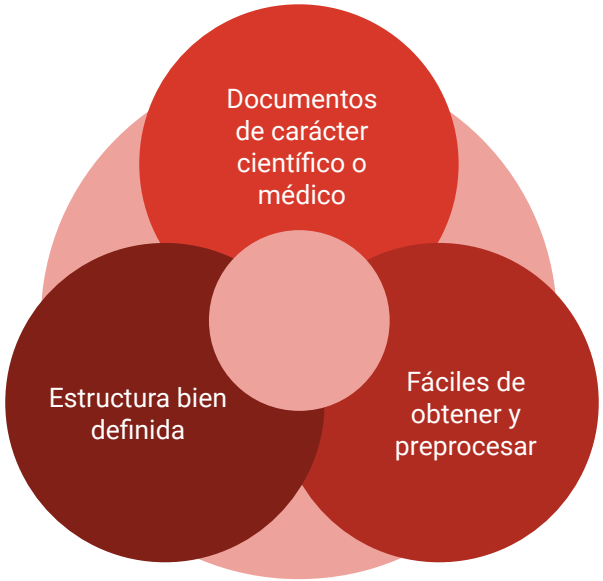


# Objetivos

- Aplicar técnicas de sumarización de textos orientados al área de la salud, enfocándonos principalmente en aquellos relacionados con COVID-19.
- Realizar una revisión de los métodos y tipos de sumarización existentes en la actualidad para determinar cuáles se ajustan mejor a nuestras necesidades y restricciones de tiempo.
- Generar api para que sea consumida y se puedan visualizar datos.

# Alcance

Los documentos que pretendemos analizar deben seguir cierto tipo de estructura, por lo que preferiblemente deseamos documentos con las siguientes características:



# Marco Teórico

## Métodos Extractivos

Los sistemas resumidores se divide en 3 tareas independientes

- **Representación Intermedia**
- **Scoring de frases**
- **Selección de frases para el resumen**

## Métodos Abstractivos

Los métodos basados en abstracción generan nuevas sentencias a partir de la información dada por el texto.

En los  últimos años los métodos abstractivos se han decantado más hacia métodos basados en redes neuronales y  deep  learning.

# Ténicas



## Métodos Extractivos:

| Enfoques y Técnicas Extractiva | |
|---|---|
| Topic Representation | Topic Words<br>Frequency-driven Approaches<br>Latent Semantic Analysis<br>Bayesian Topic Models<br>Sentence Clustering and<br>Domain-dependent Topic |
| Influence of Context | Web Summarization<br>Summarization of Scientific Articles<br>Query-focused Summarization<br>Email Summarization |
| Indicator Representations and Machine Learning for Summarization | Graph Methods for Sentence Importance<br>Machine Learning for Summarization |
| Selecting Summary Sentences | Greedy Approaches:Maximal Marginal Relevance<br>Global Summary Selection |

# Técnicas

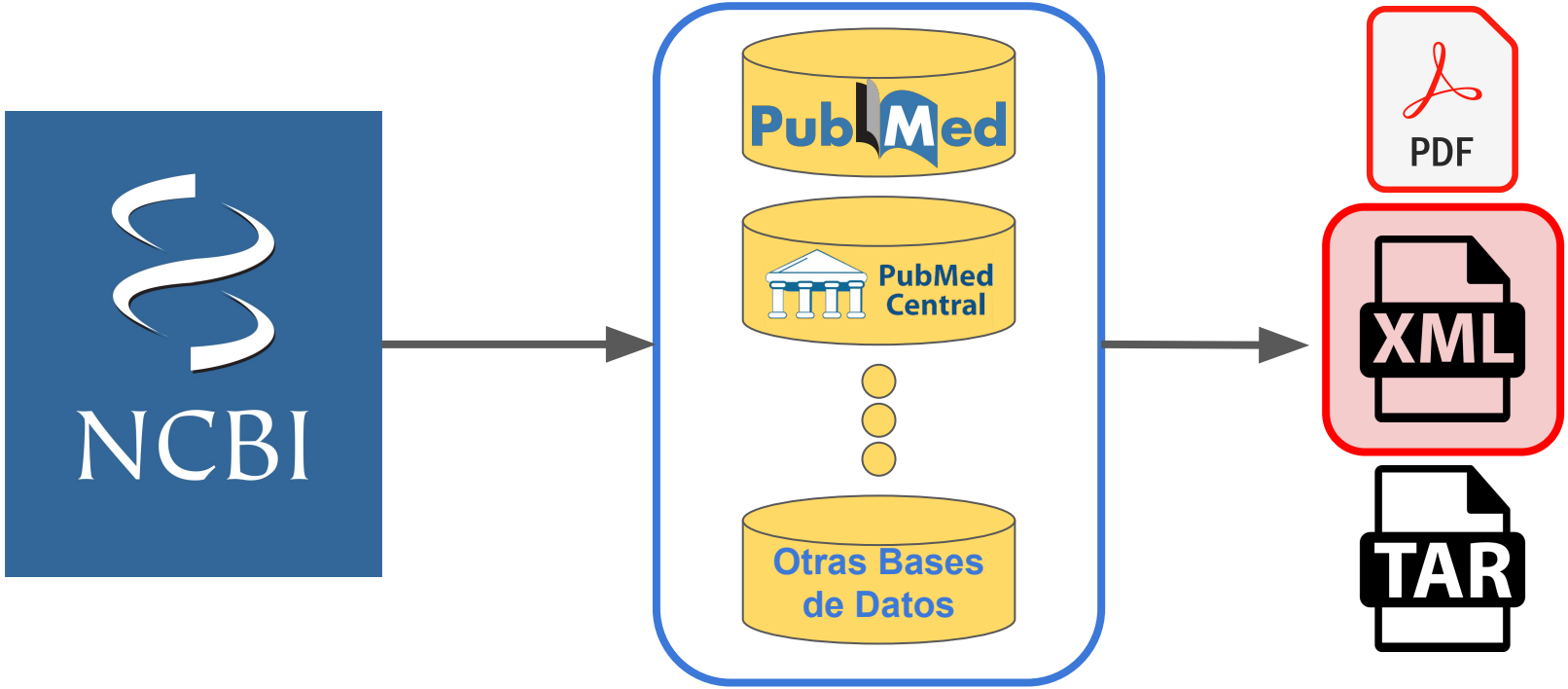
Enfoques para sumarización de textos

## Métodos Abstractos:

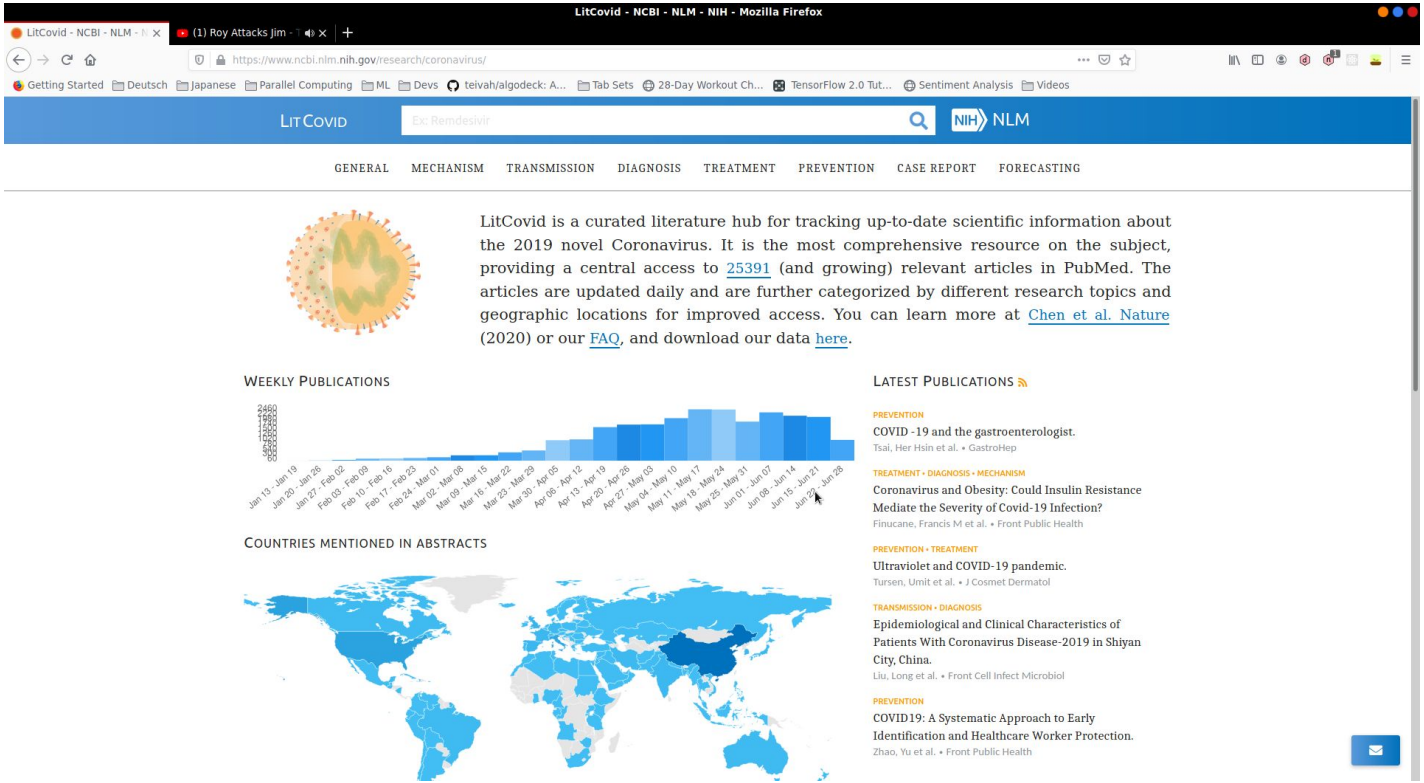
<table>
<tr><th colspan="3">Enfoques y Técnicas abstractas</th></tr>
<tr><td rowspan="2">Neural Attention Model</td><td>RNN encoder-decoder</td><td>• LSTM<br>• GRU<br>• NAS</td></tr>
<tr><td colspan="2">Attention models</td></tr>
<tr><td colspan="3">Large Vocabulary Trick and Feature-rich Encoder</td></tr>
<tr><td colspan="3">Hierarchical Attention</td></tr>
<tr><td colspan="3">Pointer Generator Network</td></tr>
<tr><td colspan="3">Coverage Mechanism (avoid repetition)</td></tr>
<tr><td>Reinforcement Learning</td><td colspan="2">Policy Gradient-based Reinforcement Learning</td></tr>
</table>

# Desarrollo

# Obtención de Datasets - Open Access



# Obtencion de Datasets - LitCovid



# Preprocesamiento y Limpieza

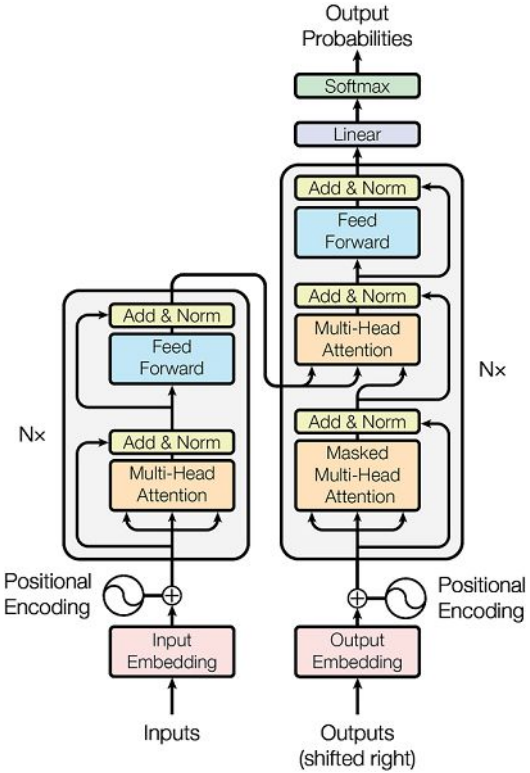
**Tags que se pueden eliminar**

<fig>
<code>
<graphic>
...

```
−<body>
  −<p>
      <italic>Background:</italic>
      The coronavirus disease 2019 (COVID-19) pandemic originated in China in late 2019 and continues to spread globally (
      <xref rid="r1-1169" ref-type="bibr">1</xref>
      ). At the time of writing, there were nearly 2 million COVID-19 cases causing approximately 110 000 deaths across more than 200 affected countries and territories (
      <xref rid="r2-1169" ref-type="bibr">2</xref>
      ). As some health care systems approach collapse, a pressing need exists for tools modeling the capacity of acute and critical care systems during the COVID-19 pandemic.
    </p>
  −<p>
      <italic>Objective:</italic>
      To develop an online tool to estimate the maximum number of COVID-19 cases that could be managed per day within the catchment area served by a health care system, given acute and critical care resource availability.
    </p>
  −<p>
      <italic>Methods:</italic>
      We modeled steady-state patient-flow dynamics constrained by the number of acute care beds, critical care beds, and mechanical ventilators available for COVID-19–infected patients seeking health care during the pandemic. Parameters for patient-flow dynamics were extracted from evolving data on COVID-19 and assumptions based on expert guidance. We used the package
      <italic>shiny</italic>
      within R, version 3.5.3 (R Foundation for Statistical Computing), to create the interactive tool.
    </p>
```

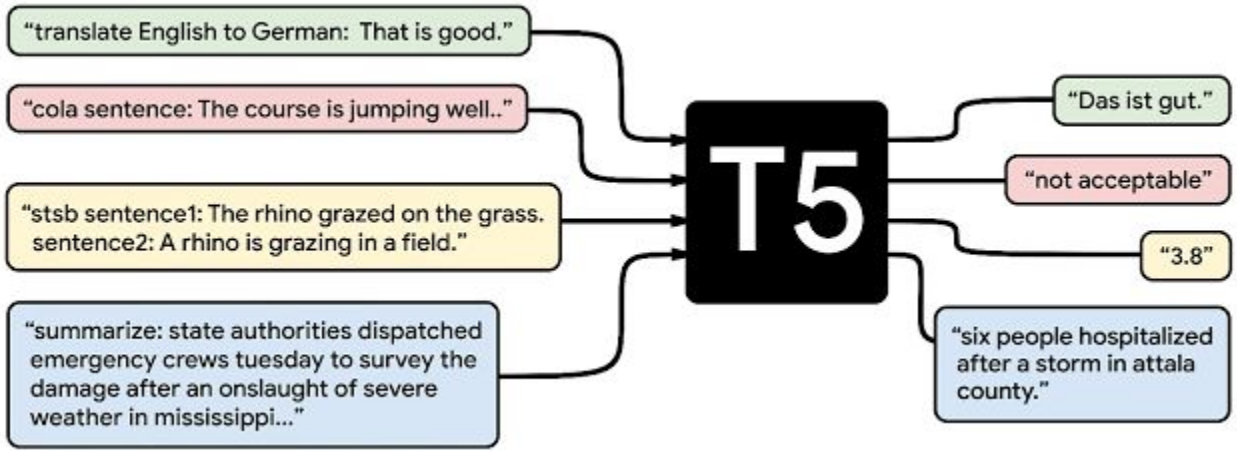
# Transformers & Attention Models

# Transformers & Attention Models

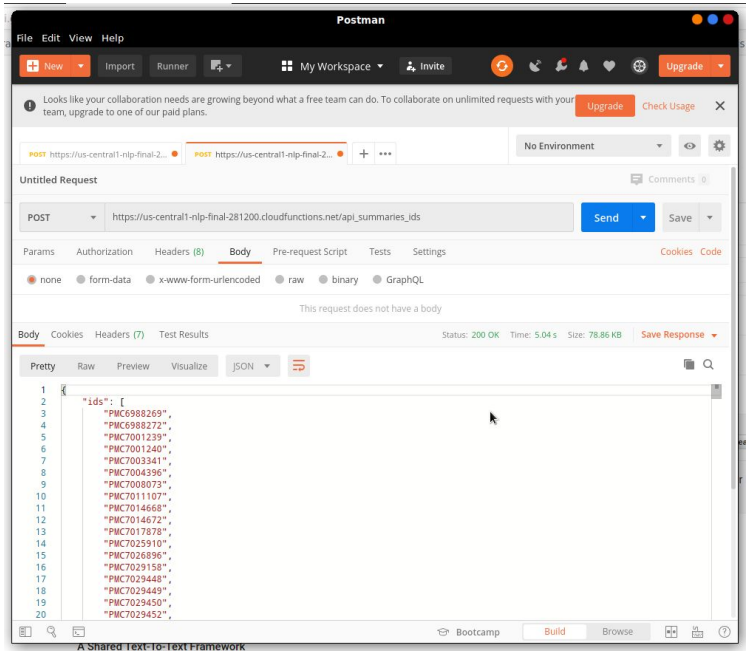


# Obtención de Resúmenes

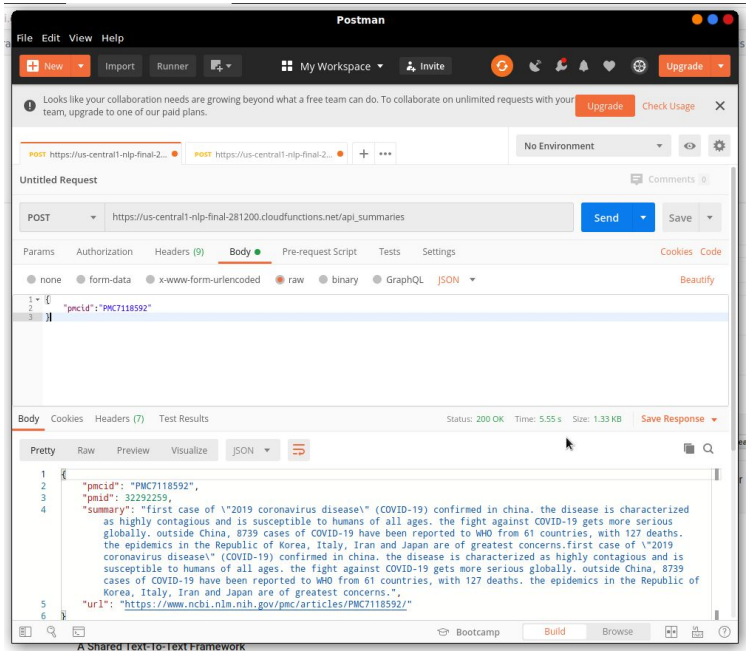
# Google T5


"translate English to German: That is good."
"cola sentence: The course is jumping well.."
"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi..."
T5
"Das ist gut."
"not acceptable"
"3.8"
"six people hospitalized after a storm in attala county."

# API




Obtener todos los PMCID con resúmen disponible




Obtener resúmen por PMCID

# Resultados

# Ejemplos de resúmenes

**What's New With the Old Coronaviruses?**

A coronavirus was first isolated as a causative agent of bronchitis in birds in 1937. it was originally discovered in humans during studies that evaluated the common cold. the novel coronavirus SARS-CoV-2 is spreading worldwide, causing anxiety, disease, and mortality. limited pediatric data are available regarding HCoV infections in children in the pre-coronavirus disease 2019 (COVID-19) era.

**Surgical Strategy During the COVID-19 Pandemic in in Milan, Italy**

The global pandemic of coronavirus disease 2019 started as an atypical pneumonia. the disease leads to severe acute respiratory syndrome and resulted in thousands of deaths. in Italy, the hardest hit region has been Lombardy, where the first case was reported on February 20. important changes in surgical activities have been introduced by hospitals in response to COVID-19-related challenges, says dr. san pellegrini.

**The potential impact of COVID-19-related disruption on tuberculosis burden**

Before the COVID-19 pandemic, over 4000 people were dying from tuberculosis (TB) every day. the impact of COVID-19 on TB outcomes is a serious cause for concern but is currently unknown. physical distancing interventions could also limit Mycobacterium tuberculosis transmission outside of households, where most transmission occurs. this has not been adequately explored in existing work. cdc: impact of COVID-19 on TB outcomes is a

**COVID-19 critical illness pathophysiology driven by diffuse pulmonary thrombi and pulmonary endothelial dysfunction responsive to thrombolysis**

Patients with severe COVID-19-induced respiratory failure demonstrate gas exchange abnormalities including shunt and dead-space ventilation. their syndrome is atypical in that the majority have relatively well-preserved lung mechanics. high rate of venous thromboembolism (VTE) in critically ill COVID-19 patients. d-dimer levels have also been noted to be elevated. rapid rises presage cardiopulmonary decompensation.

# Referencias

*[1]Entrez Programming Utilities Help[Internet]. Bethesda (MD):National Center for Biotechnology Information (US); 2010-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK25501/*

*[2]Text Mining Collections[Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2010-. Available from: https://www.ncbi.nlm.nih.gov/pmc/tools/textmining/*

*[3]Global Research On Coronavirus Disease (COVID-19) [Internet]. Who.int; 2020. Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov*