# NEURACULUS

Alejandro Vaca Serrano

SPAIN AI

iic
instituto
de ingeniería
del conocimiento

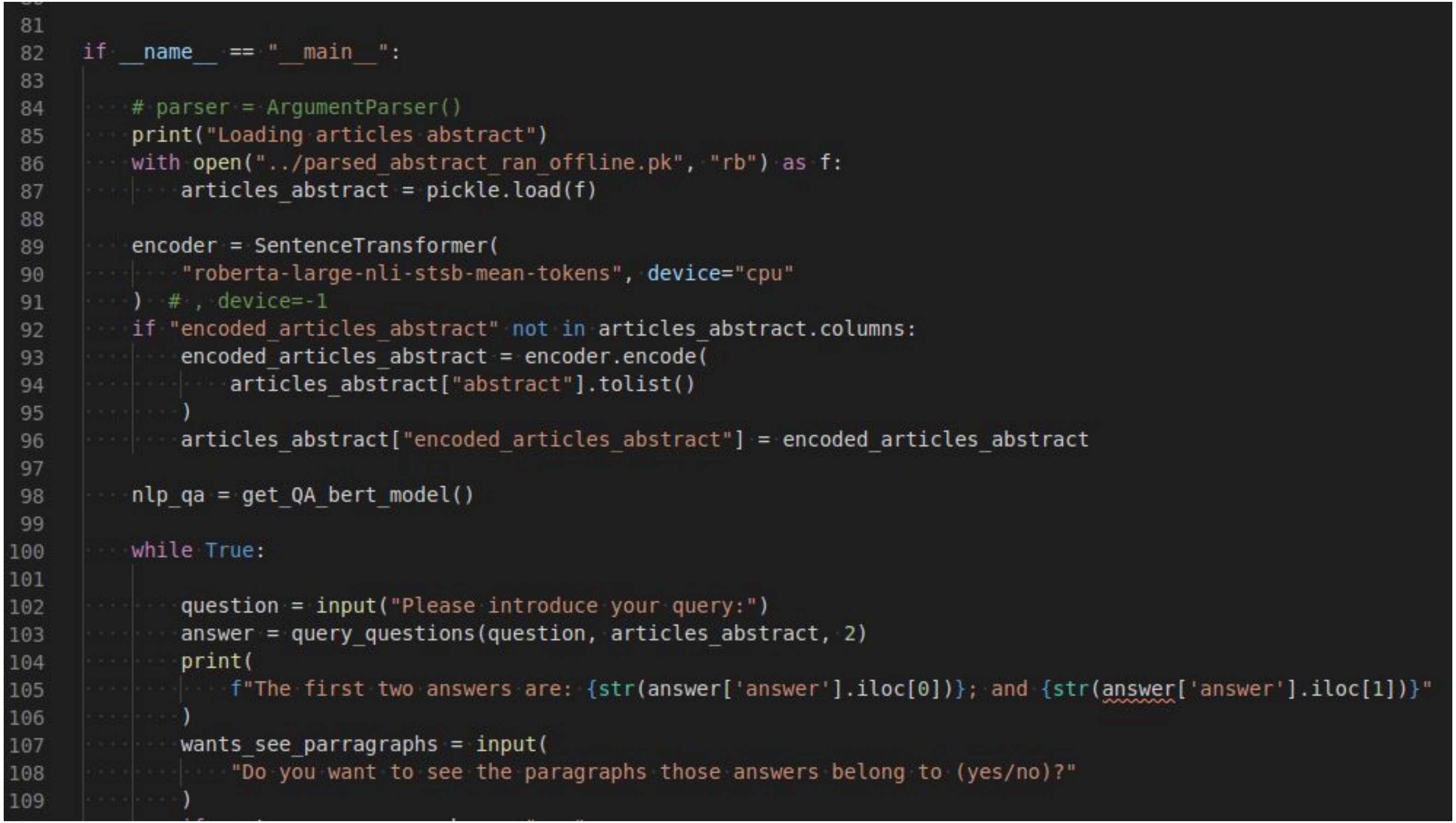
# ¿Quién soy?

- ADE Bilingüe CUNEF; TFG ⇒ ***Introduction to a Trading Strategy for Cryptocurrencies: A Machine Learning Approach.***
- Master Data Science & Big Data AFI; TFM ⇒ ***Un TraderBot Inteligente para la Gestión Automática de Carteras de Criptomonedas.***
- Data Scientist Departamento de Ingeniería Algorítmica del Instituto de Ingeniería del Conocimiento (**IIC**).
- **1er Premio** en la competición de analítica de datos más grande de España: **Cajamar UniversityHack 2020** - Minsait Land Classification.

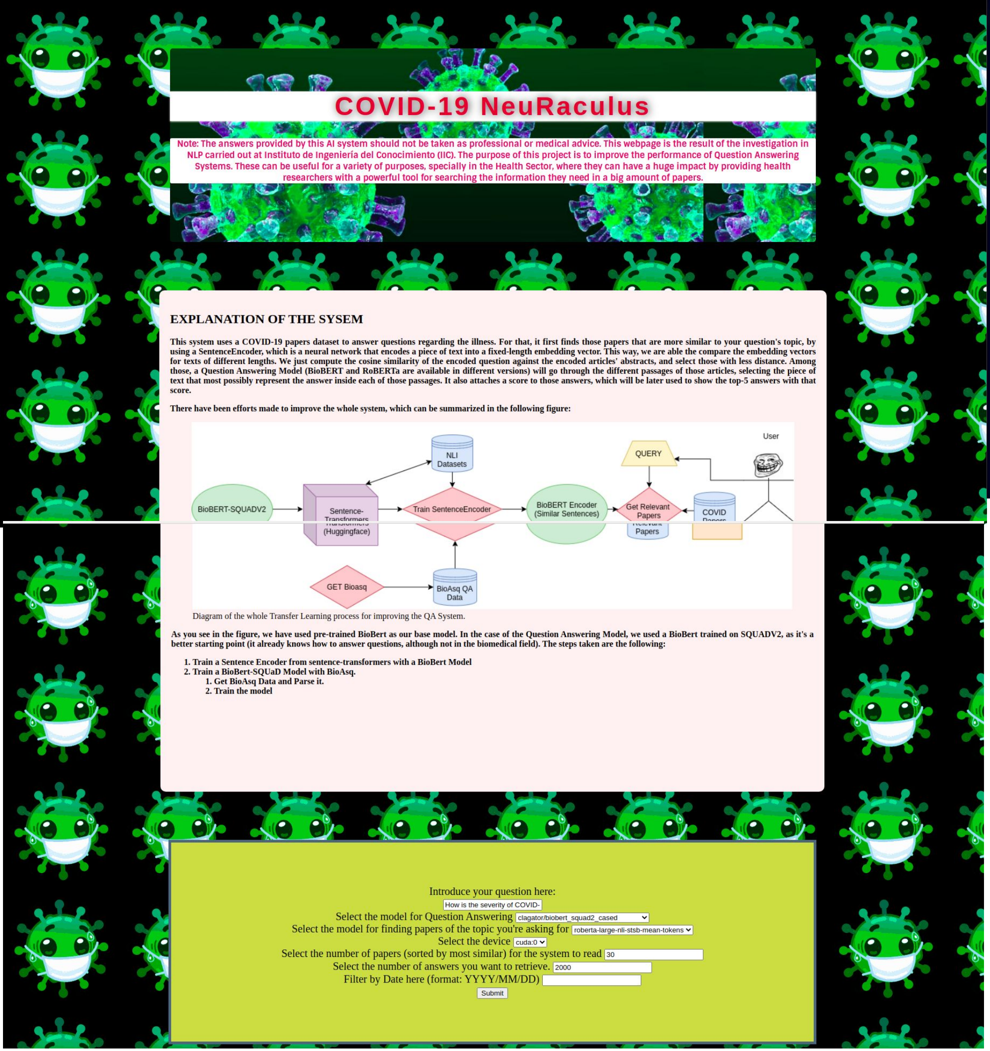
# Índice

1. ¿Qué es el Neuraculus?: Evolución del Neuraculus.
2. Objetivos
3. Demo (I)
4. Estructura  y Componentes
5. Retos
6. Demo (II)
7. Conclusión

# ¿Qué es el Neuraculus?



```python

if __name__ == "__main__":

    # parser = ArgumentParser()
    print("Loading articles abstract")
    with open("../parsed_abstract_ran_offline.pk", "rb") as f:
        articles_abstract = pickle.load(f)

    encoder = SentenceTransformer(
        "roberta-large-nli-stsb-mean-tokens", device="cpu"
    )  # , device=-1
    if "encoded_articles_abstract" not in articles_abstract.columns:
        encoded_articles_abstract = encoder.encode(
            articles_abstract["abstract"].tolist()
        )
        articles_abstract["encoded_articles_abstract"] = encoded_articles_abstract

    nlp_qa = get_QA_bert_model()

    while True:

        question = input("Please introduce your query:")
        answer = query_questions(question, articles_abstract, 2)
        print(
            f"The first two answers are: {str(answer['answer'].iloc[0])}; and {str(answer['answer'].iloc[1])}"
        )
        wants_see_parragraphs = input(
            "Do you want to see the paragraphs those answers belong to (yes/no)?"
        )
```



# COVID-19 NeuRaculus

Note: The answers provided by this AI system should not be taken as professional or medical advice. This webpage is the result of the investigation in NLP carried out at Instituto de Ingeniería del Conocimiento (IIC). The purpose of this project is to improve the performance of Question Answering Systems. These can be useful for a variety of purposes, specially in the Health Sector, where they can have a huge impact by providing health researchers with a powerful tool for searching the information they need in a big amount of papers.

## EXPLANATION OF THE SYSEM

**This system uses a COVID-19 papers dataset to answer questions regarding the illness. For that, it first finds those papers that are more similar to your question's topic, by using a SentenceEncoder, which is a neural network that encodes a piece of text into a fixed-length embedding vector. This way, we are able the compare the embedding vectors for texts of different lengths. We just compute the cosine similarity of the encoded question against the encoded articles' abstracts, and select those with less distance. Among those, a Question Answering Model (BioBERT and RoBERTa are available in different versions) will go through the different passages of those articles, selecting the piece of text that most possibly represent the answer inside each of those passages. It also attaches a score to those answers, which will be later used to show the top-5 answers with that score.**

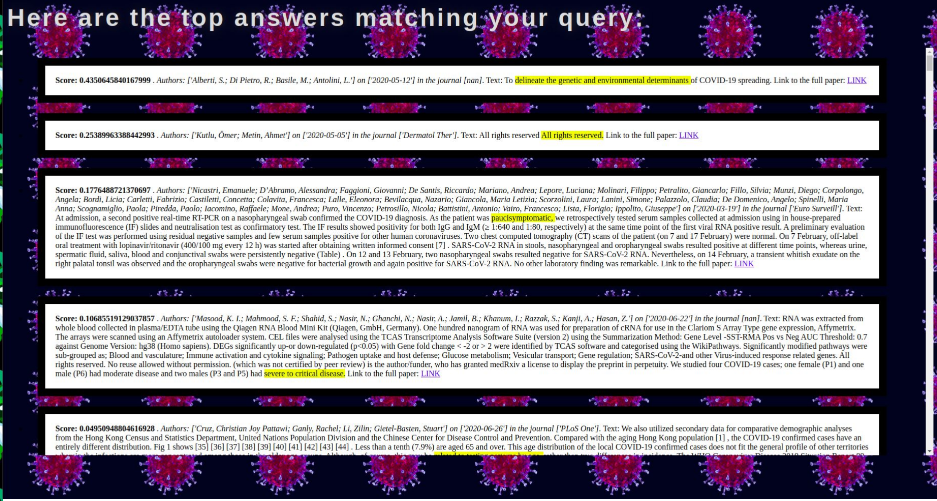
**There have been efforts made to improve the whole system, which can be summarized in the following figure:**

Diagram of the whole Transfer Learning process for improving the QA System.

**As you see in the figure, we have used pre-trained BioBert as our base model. In the case of the Question Answering Model, we used a BioBert trained on SQUADV2, as it's a better starting point (it already knows how to answer questions, although not in the biomedical field). The steps taken are the following:**

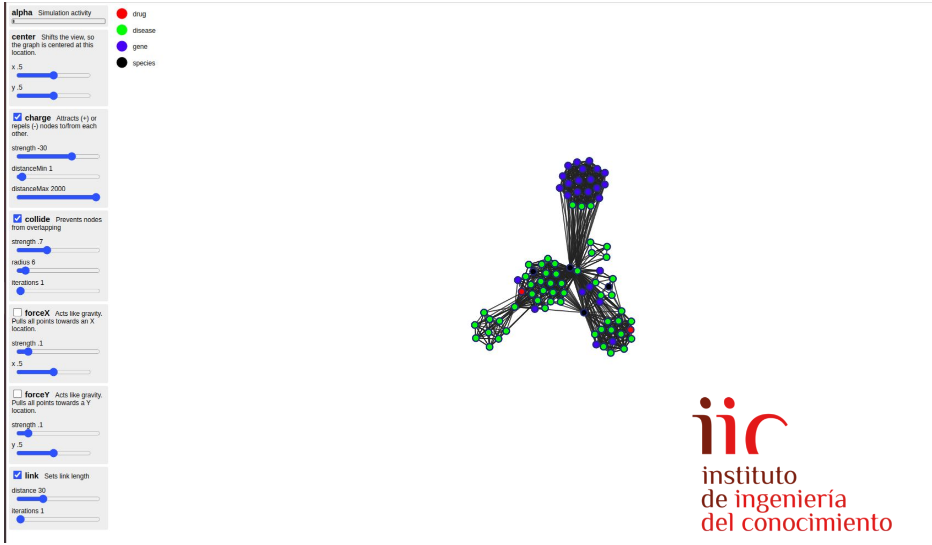
1. **Train a Sentence Encoder from sentence-transformers with a BioBert Model**
2. **Train a BioBert-SQUaD Model with BioAsq.**
    1. **Get BioAsq Data and Parse it.**
    2. **Train the model**

Introduce your question here:
How is the severity of COVID-
Select the model for Question Answering clagator/biobert_squad2_cased
Select the model for finding papers of the topic you're asking for roberta-large-nli-stsb-mean-tokens
Select the device cuda:0
Select the number of papers (sorted by most similar) for the system to read 30
Select the number of answers you want to retrieve. 2000
Filter by Date here (format: YYYY/MM/DD)



## Here are the top answers matching your query:

- **Score: 0.4350645840167999** . *Authors: ['Alberti, S.; Di Pietro, R.; Basile, M.; Antolini, L.'] on ['2020-05-12'] in the journal [nan]*. Text: To delineate the genetic and environmental determinants of COVID-19 spreading. Link to the full paper: LINK
- **Score: 0.25389963388442993** . *Authors: ['Kutlu, Ömer; Metin, Ahmet'] on ['2020-05-05'] in the journal ['Dermatol Ther']*. Text: All rights reserved All rights reserved. Link to the full paper: LINK
- **Score: 0.1776488721370697** . *Authors: ['Nicastri, Emanuele; D'Abramo, Alessandra; Faggioni, Giovanni; De Santis, Riccardo; Mariano, Andrea; Lepore, Luciana; Molinari, Filippo; Petralito, Giancarlo; Fillo, Silvia; Munzi, Diego; Corpolongo, Angela; Bordi, Licia; Carletti, Fabrizio; Castiletti, Concetta; Colavita, Francesca; Lalle, Eleonora; Bevilacqua, Nazario; Giancola, Maria Letizia; Scorzolini, Laura; Lanini, Simone; Palazzolo, Claudia; De Domenico, Angelo; Spinelli, Maria Anna; Scognamiglio, Paola; Piredda, Paolo; Iacomino, Raffaele; Mone, Andrea; Puro, Vincenzo; Petrosillo, Nicola; Battistini, Antonio; Vairo, Francesco; Lista, Florigio; Ippolito, Giuseppe'] on ['2020-03-19'] in the journal ['Euro Surveill']*. Text: At admission, a second positive real-time RT-PCR on a nasopharyngeal swab confirmed the COVID-19 diagnosis. As the patient was paucisymptomatic, we retrospectively tested serum samples collected at admission using in house-prepared immunofluorescence (IF) slides and neutralisation test as confirmatory test. The IF results showed positivity for both IgG and IgM (≥ 1:640 and 1:80, respectively) at the same time point of the first viral RNA positive result. A preliminary evaluation of the IF test was performed using residual negative samples and few serum samples positive for other human coronaviruses. Two chest computed tomography (CT) scans of the patient (on 7 and 17 February) were normal. On 7 February, off-label oral treatment with lopinavir/ritonavir (400/100 mg every 12 h) was started after obtaining written informed consent [7] . SARS-CoV-2 RNA in stools, nasopharyngeal and oropharyngeal swabs resulted positive at different time points, whereas urine, spermatic fluid, saliva, blood and conjunctival swabs were persistently negative (Table) . On 12 and 13 February, two nasopharyngeal swabs resulted negative for SARS-CoV-2 RNA. Nevertheless, on 14 February, a transient whitish exudate on the right palatal tonsil was observed and the oropharyngeal swabs were negative for bacterial growth and again positive for SARS-CoV-2 RNA. No other laboratory finding was remarkable. Link to the full paper: LINK
- **Score: 0.10685519129037857** . *Authors: ['Masood, K. I.; Mahmood, S. F.; Shahid, S.; Nasir, N.; Ghanchi, N.; Nasir, A.; Jamil, B.; Khanum, I.; Razzak, S.; Kanji, A.; Hasan, Z.'] on ['2020-06-22'] in the journal [nan]*. Text: RNA was extracted from whole blood collected in plasma/EDTA tube using the Qiagen RNA Blood Mini Kit (Qiagen, GmbH, Germany). One hundred nanogram of RNA was used for preparation of cRNA for use in the Clariom S Array Type gene expression, Affymetrix. The arrays were scanned using an Affymetrix autoloader system. CEL files were analysed using the TCAS Transcriptome Analysis Software Suite (version 2) using the Summarization Method: Gene Level -SST-RMA Pos vs Neg AUC Threshold: 0.7 against Genome Version: hg38 (Homo sapiens). DEGs significantly up-or down-regulated (p<0.05) with Gene fold change < -2 or > 2 were identified by TCAS software and categorised using the WikiPathways. Significantly modified pathways were sub-grouped as; Blood and vasculature; Immune activation and cytokine signaling; Pathogen uptake and host defense; Glucose metabolism; Vesicular transport; Gene regulation; SARS-CoV-2-and other Virus-induced response related genes. All rights reserved. No reuse allowed without permission. (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. We studied four COVID-19 cases; one female (P1) and one male (P6) had moderate disease and two males (P3 and P5) had severe to critical disease. Link to the full paper: LINK
- **Score: 0.04950948804616928** . *Authors: ['Cruz, Christian Joy Pattawi; Ganly, Rachel; Li, Zilin; Gietel-Basten, Stuart'] on ['2020-06-26'] in the journal ['PLoS One']*. Text: We also utilized secondary data for comparative demographic analyses from the Hong Kong Census and Statistics Department, United Nations Population Division and the Chinese Center for Disease Control and Prevention. Compared with the aging Hong Kong population [1] , the COVID-19 confirmed cases have an entirely different distribution. Fig 1 shows [35] [36] [37] [38] [39] [40] [41] [42] [43] [44] . Less than a tenth (7.9%) are aged 65 and over. This age distribution of the local COVID-19 confirmed cases does not fit the general profile of other territories



**alpha** Simulation activity

**center** Shifts the view, so the graph is centered at this location.
x .5
y .5

☑ **charge** Attracts (+) or repels (-) nodes to/from each other.
strength -30
distanceMin 1
distanceMax 2000

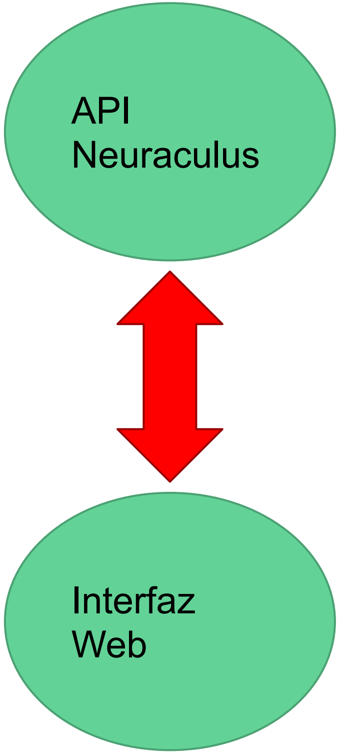
☑ **collide** Prevents nodes from overlapping
strength .7
radius 6
iterations 1

□ **forceX** Acts like gravity. Pulls all points towards an X location.
strength .1
x .5

□ **forceY** Acts like gravity. Pulls all points towards a Y location.
strength .1
y .5

☑ **link** Sets link length
distance 30
iterations 1

instituto
de ingeniería
del conocimiento


API Neuraculus

Interfaz Web

# ¿Por qué este proyecto?

VS

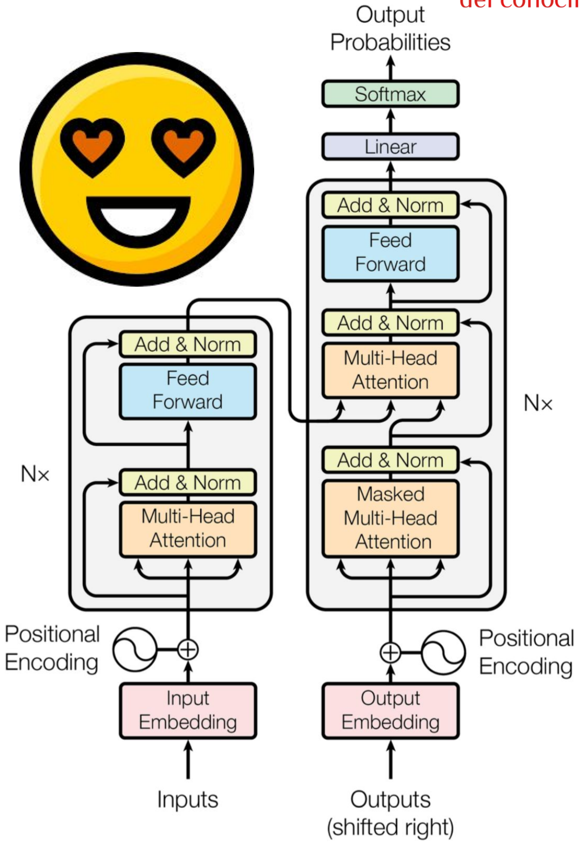


Figure 1: The Transformer - model architecture.

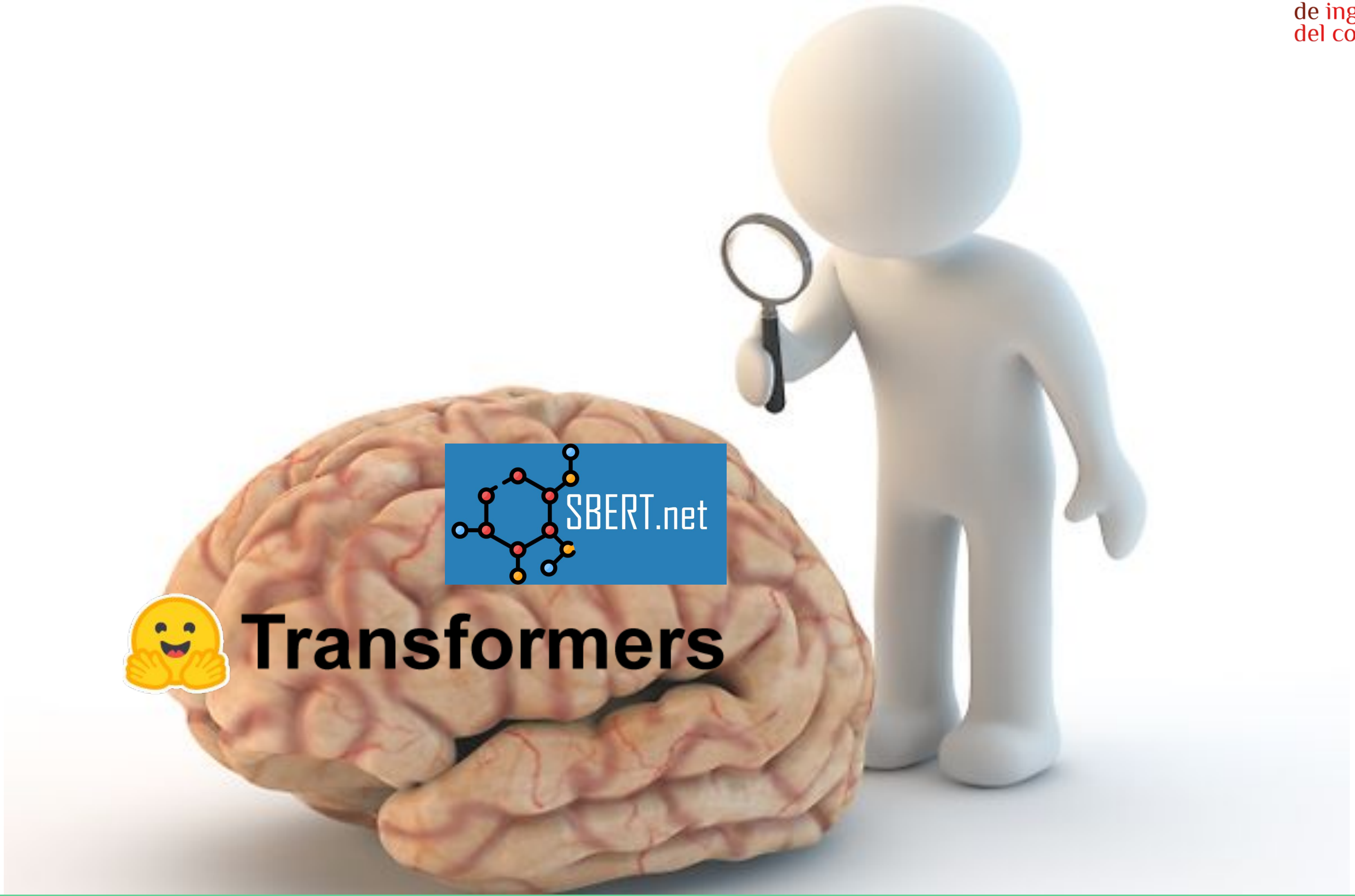
# Objetivos

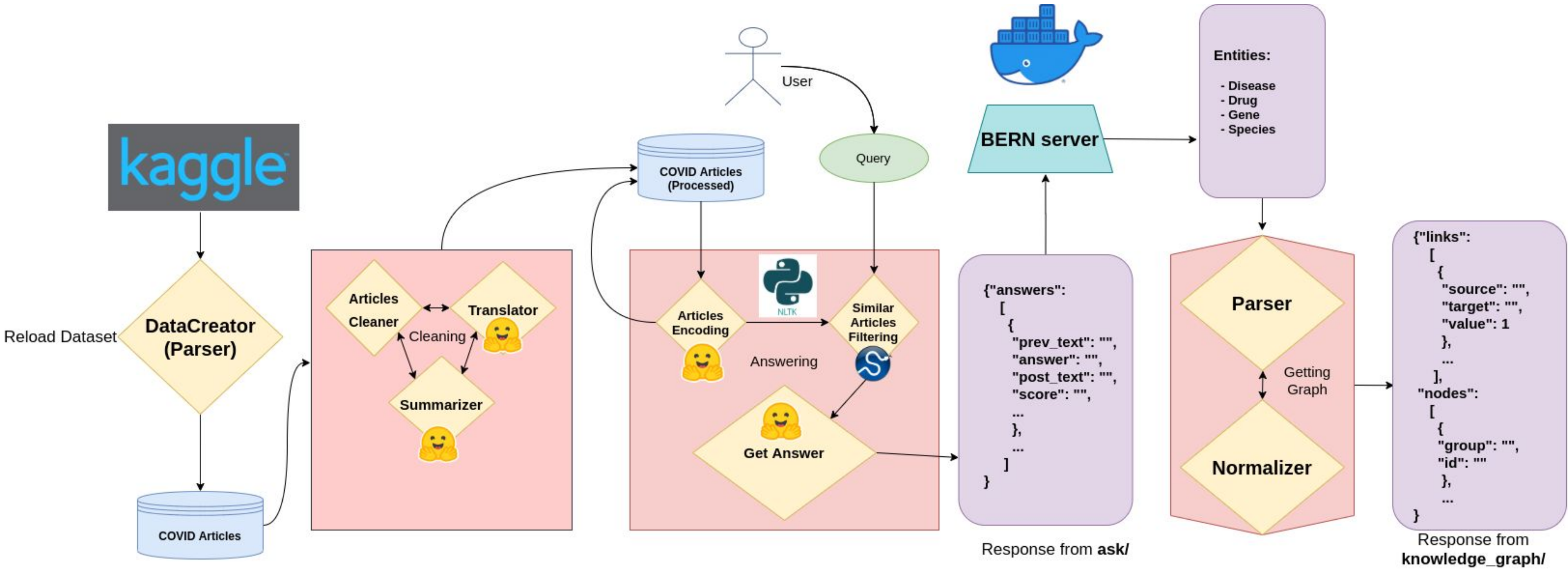
1. Investigar el alcance y la utilidad de técnicas vanguardistas de NLP.
2. Facilitar la búsqueda de información científica en artículos ⇒ Apoyo a investigadores.
3. Estructurar la información y enriquecerla.

# Demo (I)

http://localhost:4200/

# Estructura





kaggle

Reload Dataset

DataCreator (Parser)

COVID Articles

Articles Cleaner

Translator

Cleaning

Summarizer

COVID Articles (Processed)

User

Query

Articles Encoding

NLTK

Similar Articles Filtering

Answering

Get Answer

BERN server

Entities:

- Disease
- Drug
- Gene
- Species

```
{"answers":
   [
    {
     "prev_text": "",
     "answer": "",
     "post_text": "",
     "score": "",
     ...
    },
    ...
   ]
}
```

Response from **ask/**

Parser

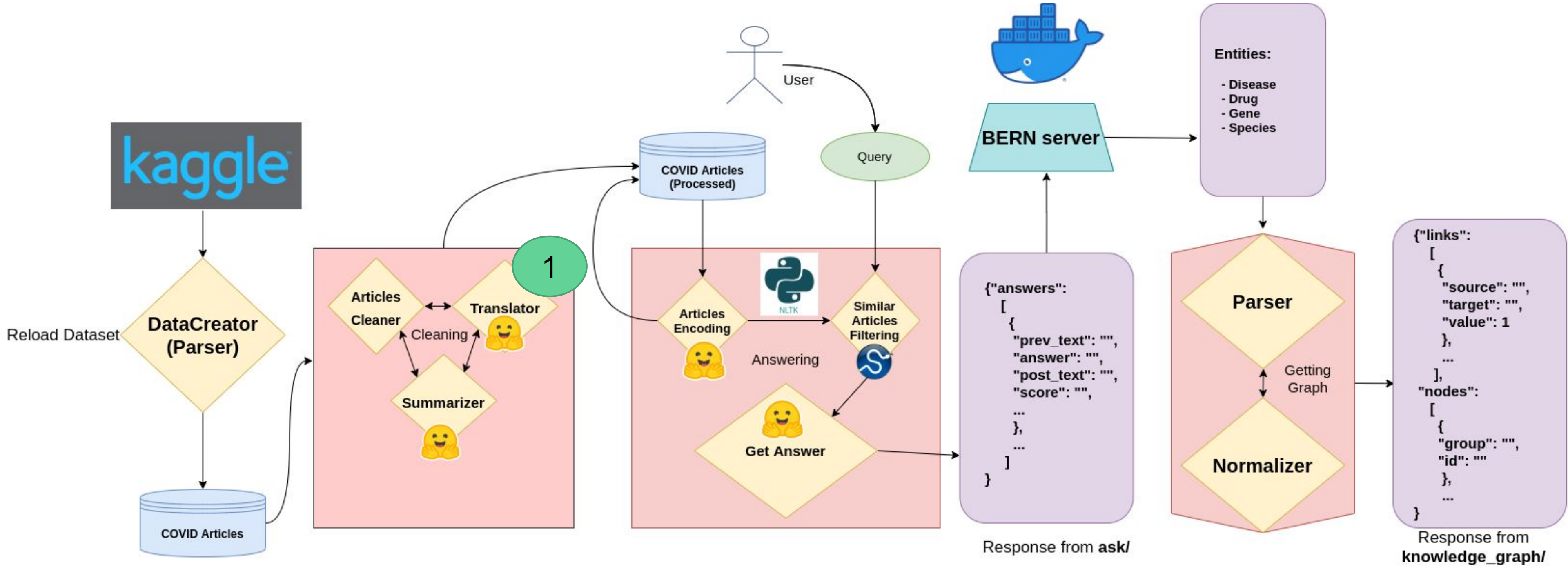
Getting Graph

Normalizer

```
{"links":
   [
    {
     "source": "",
     "target": "",
     "value": 1
     },
     ...
    ],
  "nodes":
    [
     {
     "group": "",
     "id": ""
     },
     ...
}
```

Response from **knowledge_graph/**

# Estructura Neuraculus ⇒ Translation



Entities:

- Disease
- Drug
- Gene
- Species

```
{"answers":
  [
   {
     "prev_text": "",
     "answer": "",
     "post_text": "",
     "score": "",
     ...
   },
   ...
  ]
}
```

Response from **ask/**

```
{"links":
  [
   {
     "source": "",
     "target": "",
     "value": 1
   },
   ...
  ],
 "nodes":
  [
   {
    "group": "",
    "id": ""
   },
   ...
}
```

Response from **knowledge_graph/**

# Traducción del Español

**Original:** El COVID-19 entra al cuerpo humano principalmente a través de aerosoles, cuando las partículas de SARS-CoV-2 entran en contacto con una proteína llamada ACE2, en un proceso conocido como endocitosis.

**Google Translator:** COVID-19 enters the human body mainly through aerosols, when the SARS-CoV-2 particles they come into contact with a protein called ACE2, in a process known as endocytosis.

**Nuestro sistema:** COVID-19 enters the human body primarily through aerosols, when SARS-CoV-2 particles come into contact with a protein called ACE2, in a process known as endocytosis.

# Traducción del Francés

**Original:** Le 28 octobre 2020, le président de la République a décidé de prendre des mesures pour réduire à leur plus strict minimum les contacts et déplacements sur l'ensemble du territoire en établissant un confinement du 30 octobre au 1er décembre minimum.
Les déplacements sont interdits sauf dans les cas suivants et sur attestation uniquement pour :
Les déplacements entre le domicile et le lieu d'exercice de l'activité professionnelle ou les universités (ou établissements d'enseignement supérieur) pour les étudiants ou les centres de formation pour adultes et les déplacements professionnels ne pouvant être différés.
Les déplacements pour effectuer des achats de fournitures nécessaires à l'activité professionnelle, des achats de première nécessité dans des établissements dont les activités demeurent autorisées (liste sur gouvernement.fr) et les livraisons à domicile.

**Google Translator:** On October 28, 2020, the President of the Republic decided to take measures to reduce to contact and travel throughout the country to their strict minimum by establishing a confinement from October 30 to December 1 minimum. Travel is prohibited except in the following cases and upon certification only for: Travel between home and the place of exercise of professional activity or universities (or higher education institutions) for students or adult training centers and business travel cannot be postponed. Travel to purchase supplies necessary for professional activity, essential purchases in establishments whose activities remain authorized (list on government.fr) and home deliveries.

**Nuestro Sistema:** On 28 October 2020, the President of the Republic decided to take measures to minimise contact and travel throughout the territory by establishing a confinement from 30 October to 1 December at a minimum. Travel is prohibited except in the following cases and on attestation only for: Movements between the home and the place of exercise of the professional activity or universities (or higher education institutions) for students or adult training centres and professional travel which cannot be deferred. Travel to make purchases of supplies necessary for the professional activity, purchases of first necessity in establishments whose activities remain authorized (list on government.fr) and deliveries to home.

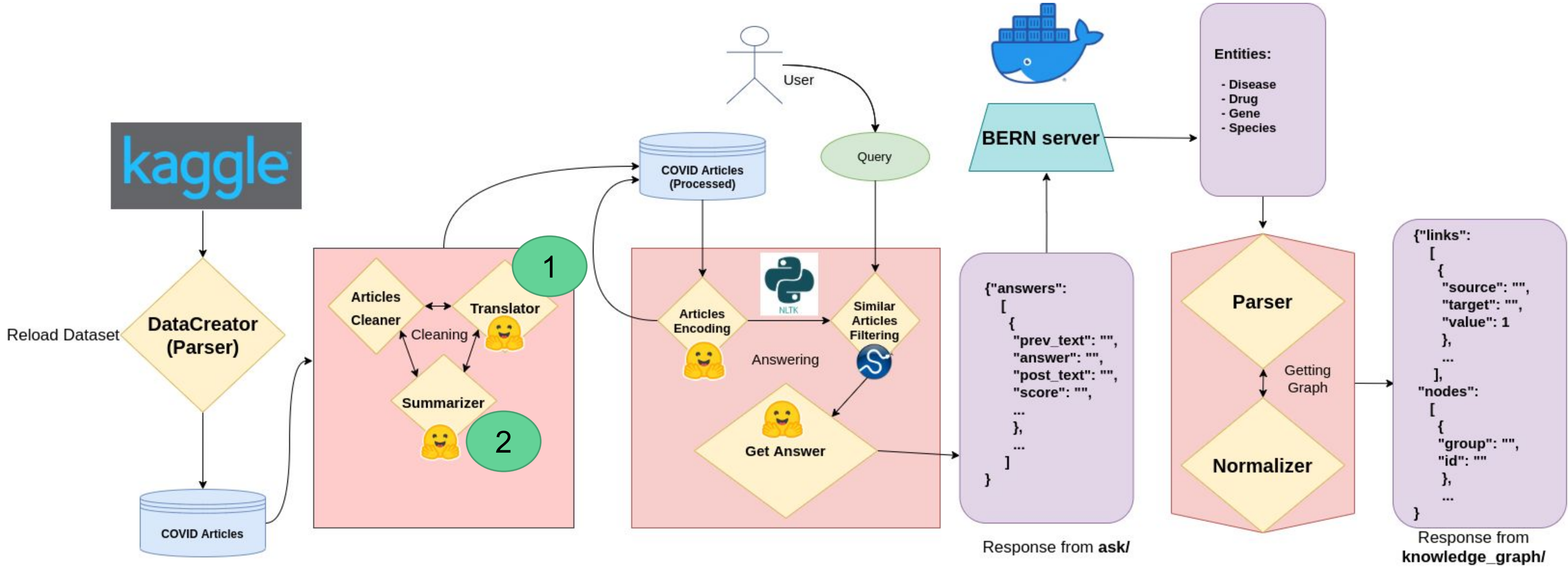
# Traducción del Alemán

**Original:** Prof. Janssens, der Chef der deutschen Intensivmediziner ist, die im Ernstfall so eine Triage durchführen müssen, sagte der „Rheinischen Post“: „Herr Drosten ist ein erstklassiger Virologe und einer der wichtigsten Experten, die wir derzeit bei der Pandemiebekämpfung haben. Seine Äußerungen zu einer möglicherweise drohenden Triage in Deutschland kann ich jedoch nicht nachvollziehen und halte sie für unverantwortlich. Indem er auf diese Weise davor warnt, macht er den Menschen unnötige Angst.“ Und weiter sagte der Intensivmediziner: Man sei von solchen Zuständen trotz Personalknappheit weit entfernt. Janssens wörtlich: „Herr Drosten sollte sich aus der Diskussion um Kapazitätsengpässe auf Intensivstationen heraushalten.

**Google Translator:** Prof. Janssens, the head of the German intensive care physicians who have to carry out such a triage in an emergency, told the “Rheinische Post”: “Mr. Drosten is a first-class virologist and one of the most important experts we currently have in combating pandemics. However, I cannot understand his statements about a possible threat of triage in Germany and I consider them to be irresponsible. By warning against it in this way, he scares people unnecessarily. "And the intensive care doctor went on to say: Despite the shortage of staff, it is a long way from such conditions. Janssens literally: "Mr. Drosten should stay out of the discussion about capacity bottlenecks in intensive care units."

**Nuestro sistema:** I'm not sure if this is the case, but I'm not sure if this is the case, and I'm not sure if this is the case.

# Estructura Neuraculus ⇒ Summarization



Reload Dataset

DataCreator (Parser)

COVID Articles

Articles Cleaner

Translator

Cleaning

Summarizer

1

2

COVID Articles (Processed)

User

Query

Articles Encoding

Similar Articles Filtering

Answering

Get Answer

BERN server

Entities:

- Disease
- Drug
- Gene
- Species

```
{"answers":
   [
    {
      "prev_text": "",
      "answer": "",
      "post_text": "",
      "score": "",
      ...
    },
    ...
   ]
}
```

Response from **ask/**

Parser

Getting Graph

Normalizer

```
{"links":
   [
    {
      "source": "",
      "target": "",
      "value": 1
    },
    ...
   ],
 "nodes":
   [
    {
      "group": "",
      "id": ""
    },
    ...
}
```

Response from **knowledge_graph/**

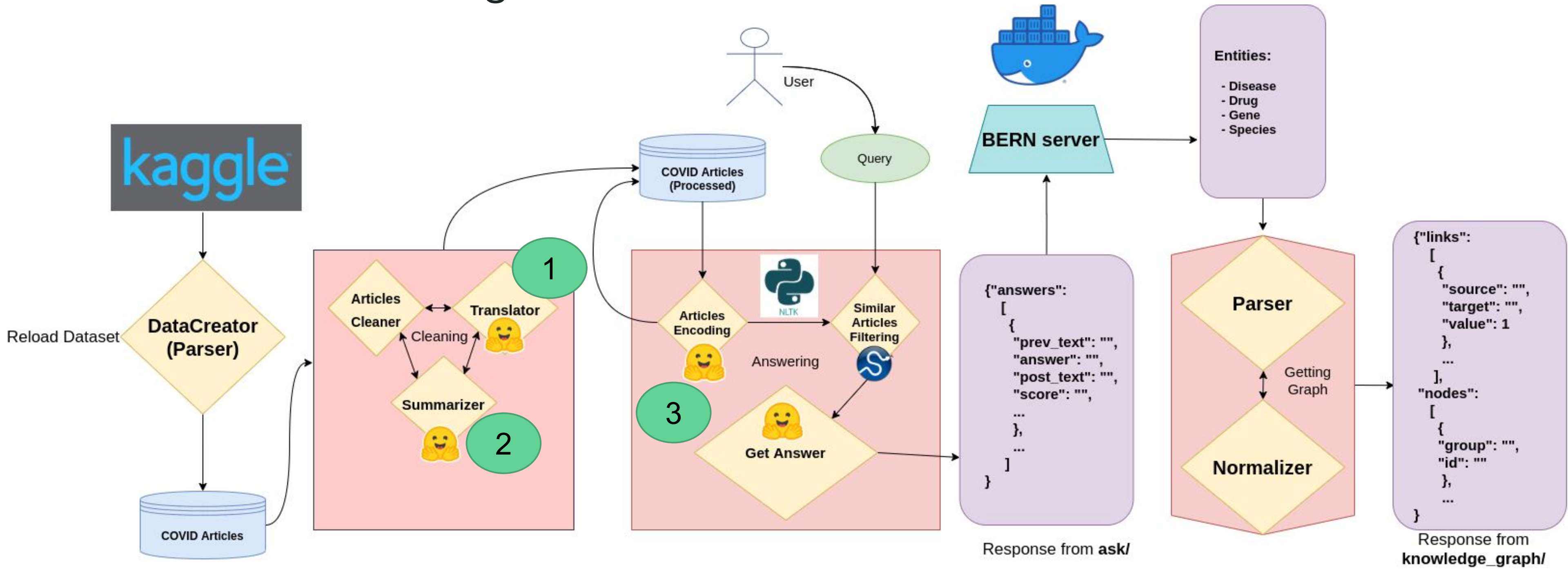
# Summarizer: Ejemplo

**Original:** The exploration vs. exploitation trade-off has been most thoroughly studied through the multi-armed bandit problem and for finite state space MDPs in Burnetas and Katehakis (1997).[5] Reinforcement learning requires clever exploration mechanisms; randomly selecting actions, without reference to an estimated probability distribution, shows poor performance. The case of (small) finite Markov decision processes is relatively well understood. However, due to the lack of algorithms that scale well with the number of states (or scale to problems with infinite state spaces), simple exploration methods are the most practical. One such method is {\displaystyle \varepsilon }\varepsilon -greedy, where {\displaystyle 0<\varepsilon <1}{\displaystyle 0<\varepsilon <1} is a parameter controlling the amount of exploration vs. exploitation. With probability {\displaystyle 1-\varepsilon }1-\varepsilon, exploitation is chosen, and the agent chooses the action that it believes has the best long-term effect (ties between actions are broken uniformly at random). Alternatively, with probability {\displaystyle \varepsilon }\varepsilon , exploration is chosen, and the action is chosen uniformly at random. {\displaystyle \varepsilon }\varepsilon is usually a fixed parameter but can be adjusted either according to a schedule (making the agent explore progressively less), or adaptively based on heuristics.[6]

**Nuestro Resumen:** The exploration vs. exploitation trade-off has been studied through the multi-armed bandit problem and for finite state space MDPs in Burnetas and Katehakis (1997). Reinforcement learning requires clever exploration mechanisms. Simple exploration methods are the most practical.

instituto
de ingeniería
del conocimiento

# Estructura Neuraculus ⇒ Similar Texts Filtering + Question Answering



Entities:

- Disease
- Drug
- Gene
- Species

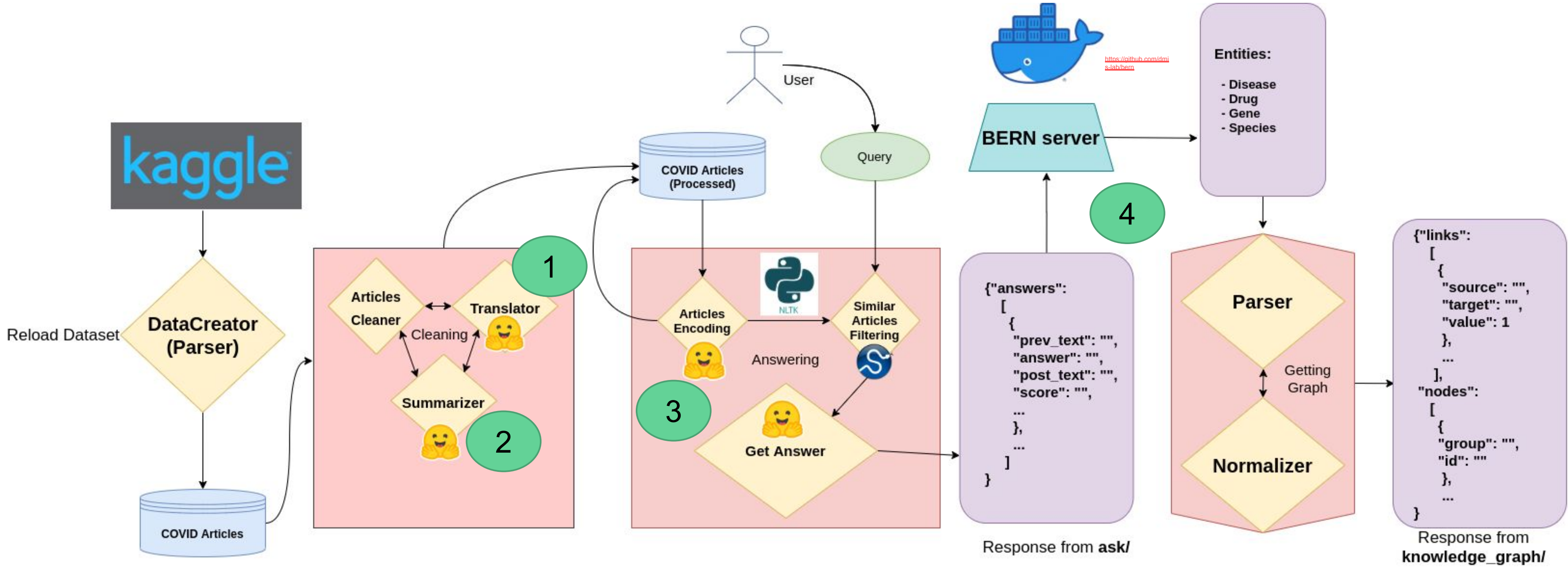
```
{"answers":
  [
    {
      "prev_text": "",
      "answer": "",
      "post_text": "",
      "score": "",
      ...
    },
    ...
  ]
}
```

Response from **ask/**

```
{"links":
  [
    {
      "source": "",
      "target": "",
      "value": 1
    },
    ...
  ],
 "nodes":
  [
    {
      "group": "",
      "id": ""
    },
    ...
}
```

Response from **knowledge_graph/**

# Estructura Neuraculus ⇒ NER + Normalization



Entities:

- Disease
- Drug
- Gene
- Species

```
{"answers":
  [
    {
      "prev_text": "",
      "answer": "",
      "post_text": "",
      "score": "",
      ...
    },
    ...
  ]
}
```

Response from **ask/**

```
{"links":
  [
    {
      "source": "",
      "target": "",
      "value": 1
    },
    ...
  ],
 "nodes":
  [
    {
     "group": "",
     "id": ""
    },
    ...
}
```

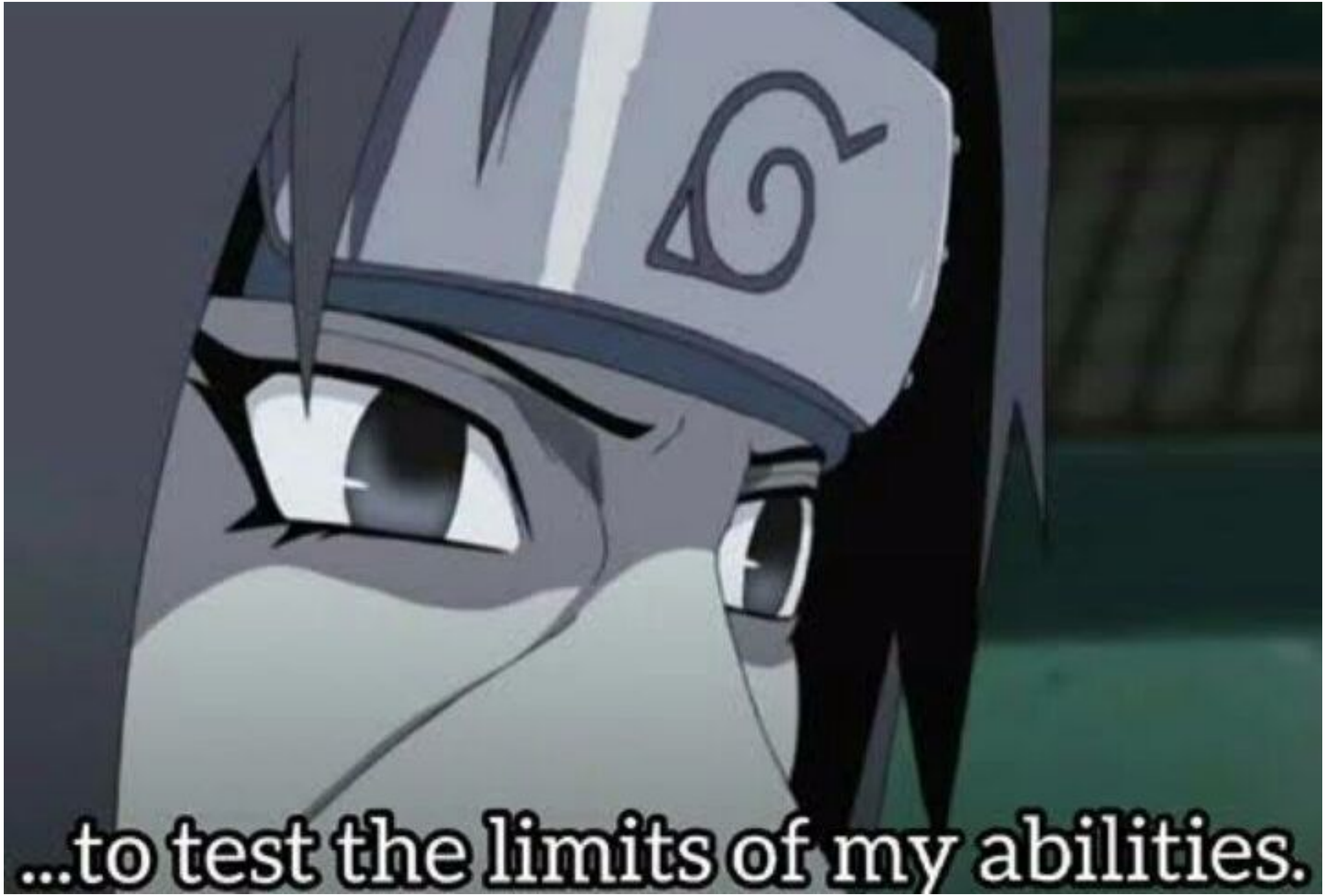
Response from **knowledge_graph/**

# Retos del sistema

1. Grandes requisitos de hardware: approx. 32GB free RAM, 200GB free Disk, 1 GPU w/ 16GB Memory recomendable.
2. Modelos no entrenados para este objetivo en concreto: puro transfer learning, sin fine-tuning.
3. Muchos artículos sin abstract: difícil encontrar los artículos del tema correcto (pese al Summarizer) + Artículos incompletos / mal parseados en BD original.
4. Textos muy largos (>> 512 / 1024 de longitud de los modelos) ⇒ Los batches son de párrafos, no de artículos ⇒ Ralentiza el proceso.

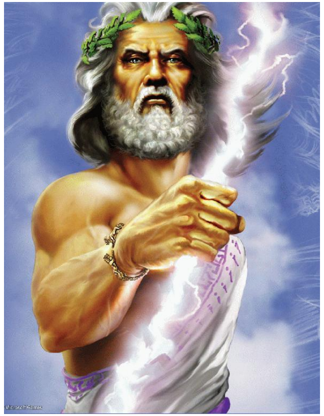
# Demo (II)

[http://localhost:4200/](http://localhost:4200/)



instituto
de ingeniería
del conocimiento

# Conclusión

-  vs.
- Modelos Transformers ⇒ Presente y Futuro.
- Apoyo investigación ⇒ Especial necesidad desde inicio de pandemia.
- Lucha contra la desinformación.

Muchas gracias por vuestro tiempo, y especialmente por vuestra atención.

♡