Vectorización de Textos o Contenidos Conceptuales

José Carlos Machicao

GestioDinámica

Presentación Personal

- **José Carlos Machicao**
  - Ing. Mecánico PUCP
  - Master en Energía, Universidad de Cardiff, UK
  - Especialista en Modelamiento Complejo para Gestión
  - Gestor de Portafolios y Proyectos
  - Ciencia de Datos: MIT, Complexity Academy
  - Python: DataCamp, StackOverflow

Contexto de Comunicación

Los problemas y límites de la comunicación

Gran **volumen** de textos: Impide hacer lecturas completas por falta de tiempo

Documentos **innecesarios**: Genera lecturas que ocupan tiempo y no son útiles

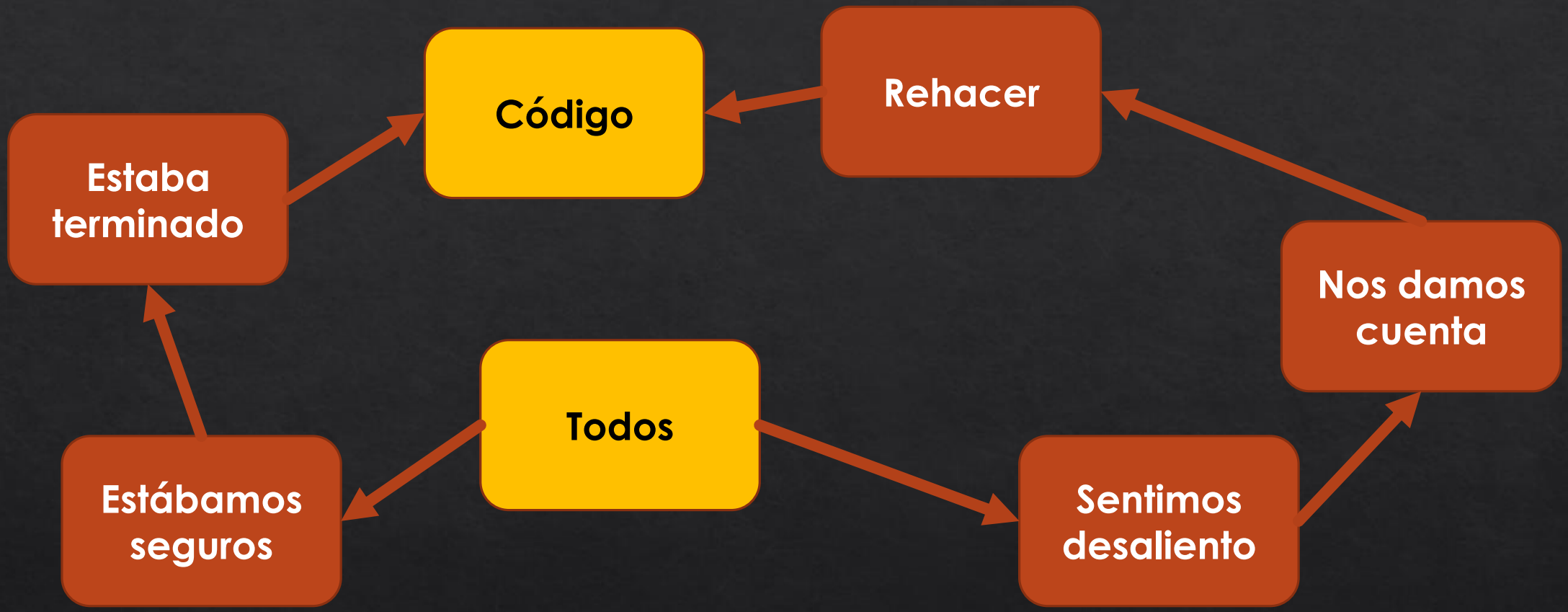
Documentos **complejos**: implican más tiempo para interpretarlos o extraer lo necesario

El sentido de una frase

- Todos sentimos un poco de desaliento cuando nos damos cuenta que hay que rehacer un código, aun cuando estábamos totalmente seguros que ya estaba terminado.

- A. Seguridad
- B. Desaliento
- C. Terminar el código

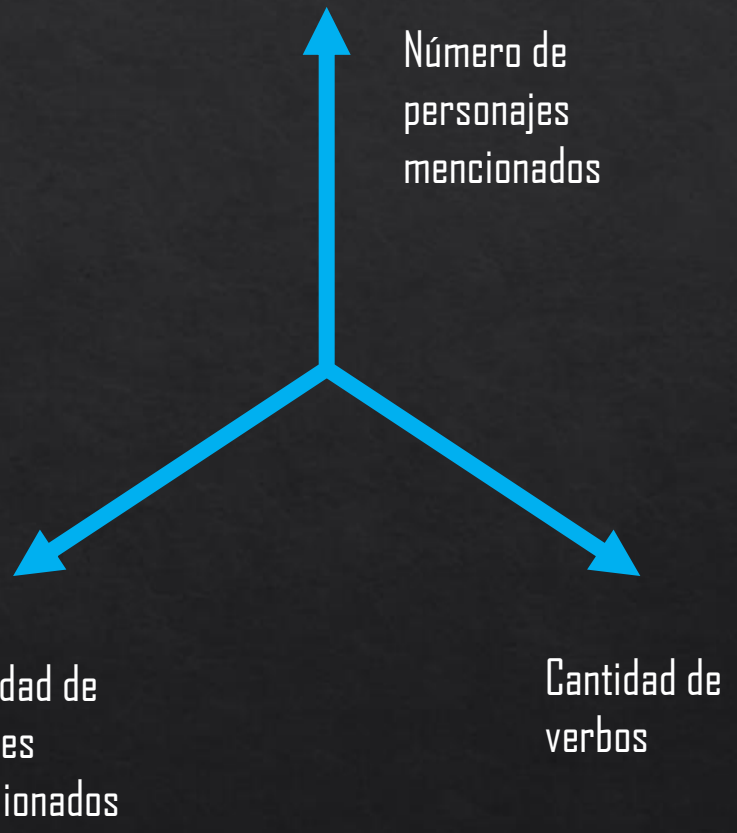
El sentido de una frase: Concatenación, Estructura



- Todos sentimos un poco de desaliento cuando nos damos cuenta que hay que rehacer un código, aun cuando estábamos totalmente seguros que ya estaba terminado.

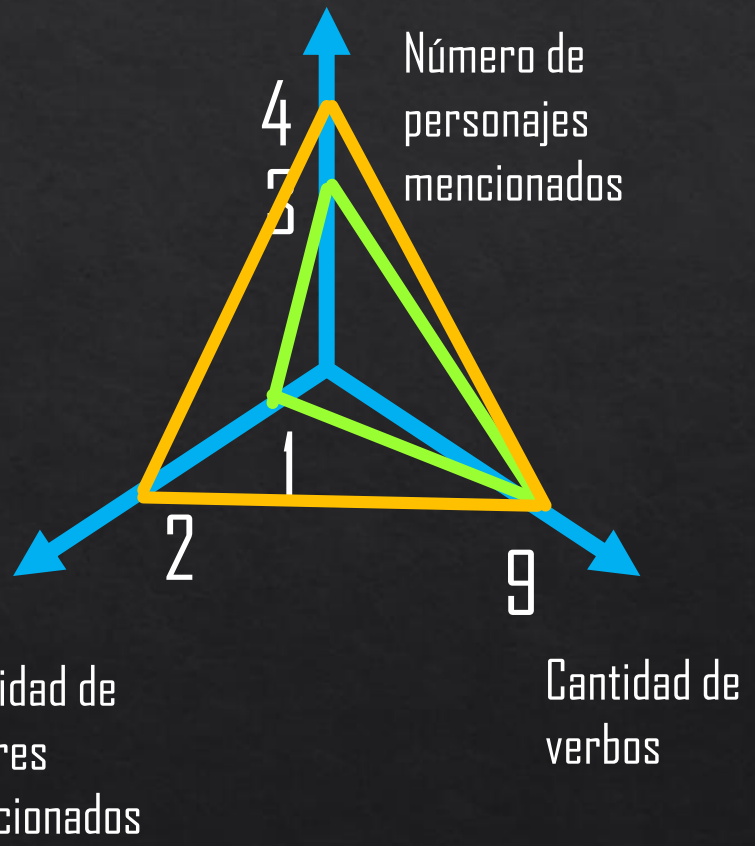
Ejemplo de Vectorización

- **Fragmento A**
  - Érase una vez una niña que era muy querida por su abuelita, a la que visitaba con frecuencia aunque vivía al otro lado del bosque. Su madre que sabía coser muy bien le había hecho una bonita caperuza roja que la niña nunca se quitaba, por lo que todos la llamaban Caperucita Roja.
- **Fragmento B**
  - Había una vez 3 cerditos que eran hermanos y vivían en lo más profundo del bosque. Siempre habían vivido felices y sin preocupaciones en aquel lugar, pero ahora se encontraban temerosos de un lobo que merodeaba la zona. Fue así como decidieron que lo mejor era construir cada uno su propia casa, que les serviría de refugio si el lobo los atacaba.



Ejemplo de Vectorización (Simplificado)

- **Fragmento A**
  - Érase una vez una niña que era muy querida por su abuelita, a la que visitaba con frecuencia aunque vivía al otro lado del bosque. Su madre que sabía coser muy bien le había hecho una bonita caperuza roja que la niña nunca se quitaba, por lo que todos la llamaban Caperucita Roja.
- **Fragmento B**
  - Había una vez 3 cerditos que eran hermanos y vivían en lo más profundo del bosque. Siempre habían vivido felices y sin preocupaciones en aquel lugar, pero ahora se encontraban temerosos de un lobo que merodeaba la zona. Fue así como decidieron que lo mejor era construir cada uno su propia casa, que les serviría de refugio si el lobo los atacaba.



Ejemplo: Word2Vec

- **Frases:**
  - Érase una vez una niña que era muy querida por su abuelita, a la que visitaba con frecuencia aunque vivía al otro lado del bosque.
  - El lobo cruzó el bosque, entró a la casa, atacó a la abuelita y se puso su capa y su gorro.

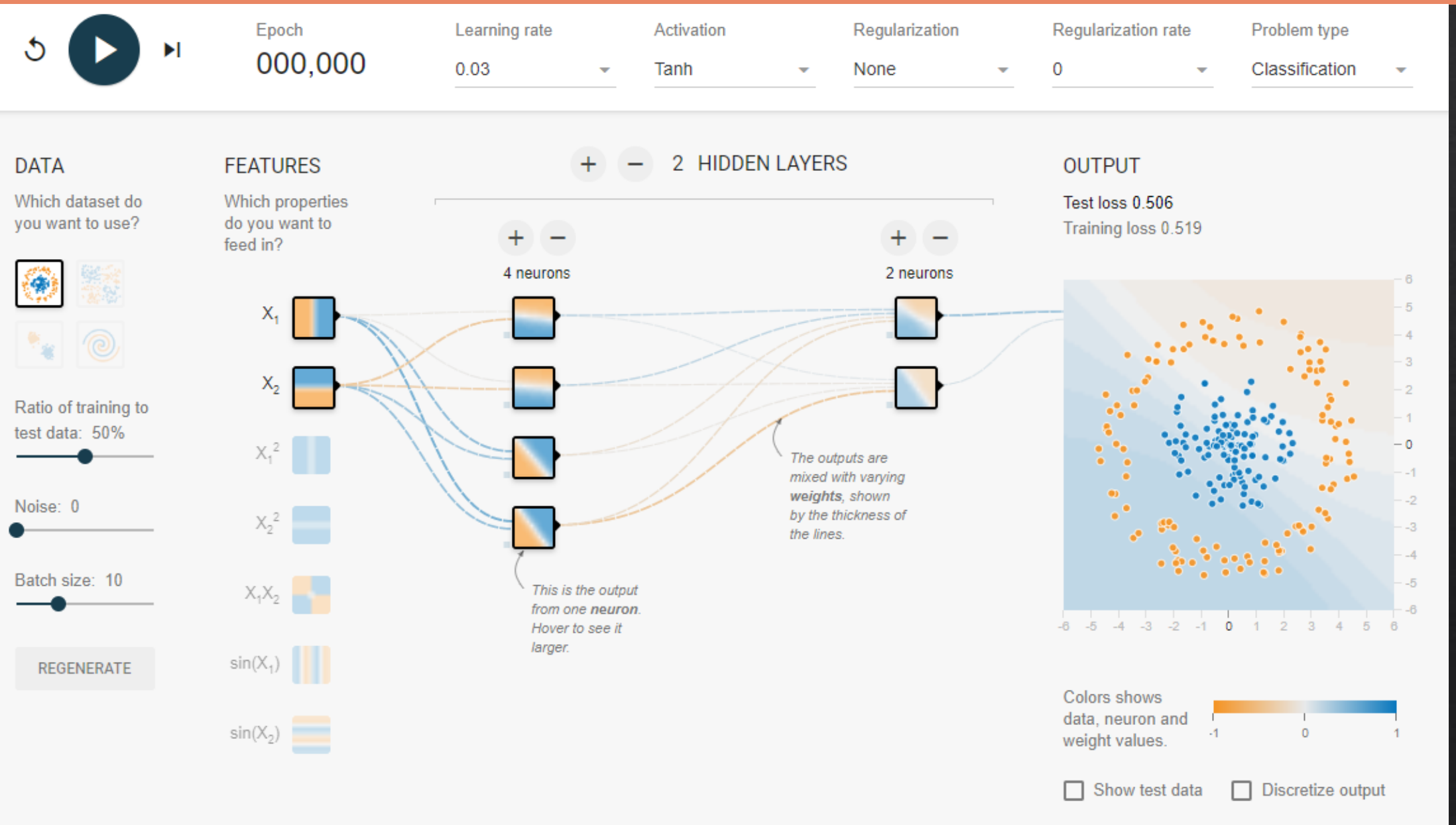
Distancia: 12 palabras hacia adelante
Distancia: 8 palabras hacia atrás
Distancia: ...

Fuente: Doc2Vec

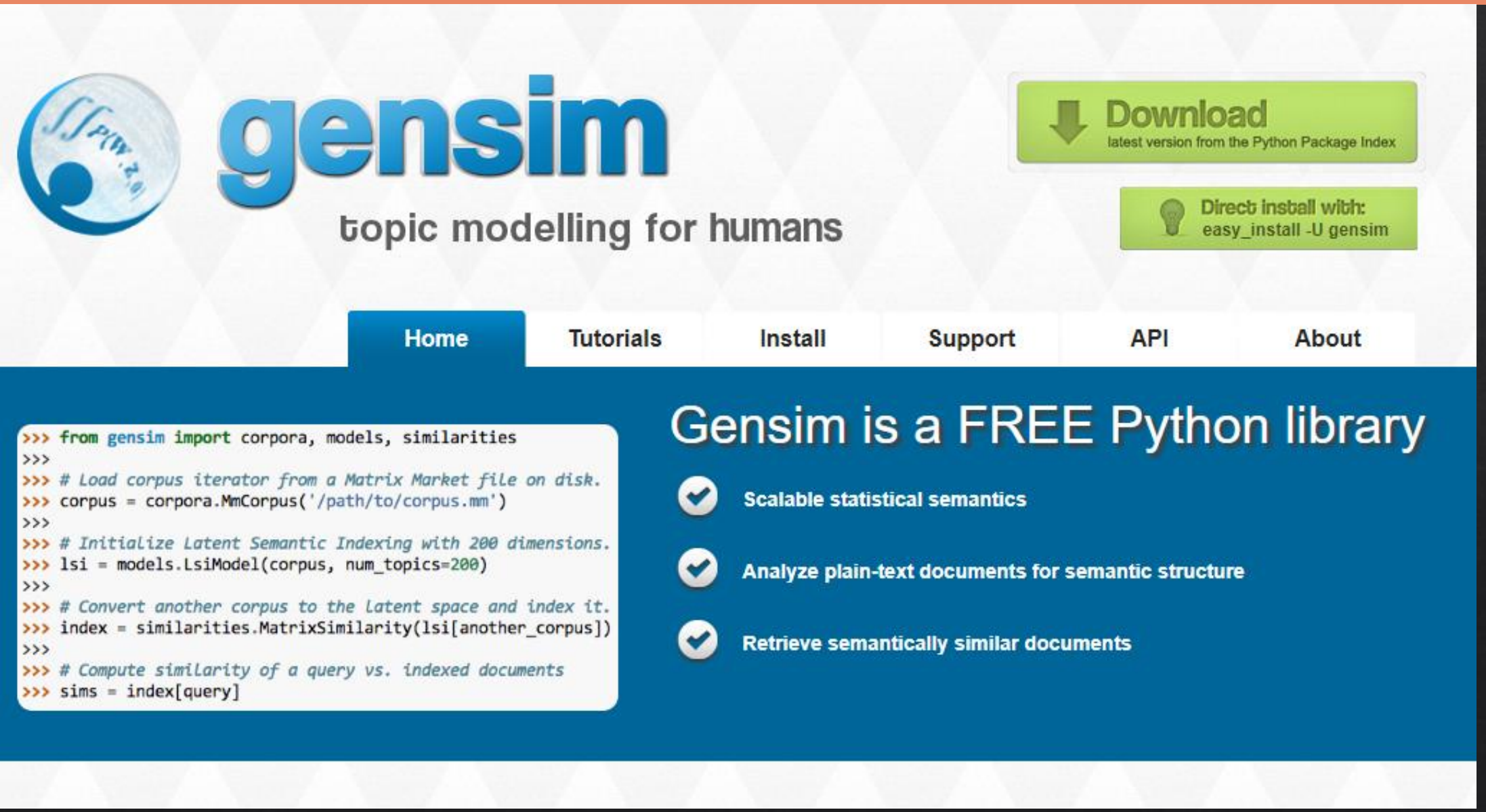
Distributed Representations of Sentences and Documents

Quoc Le QVL@GOOGLE.COM
Tomas Mikolov TMIKOLOV@GOOGLE.COM
Google Inc, 1600 Amphitheatre Parkway, Mountain View, CA 94043

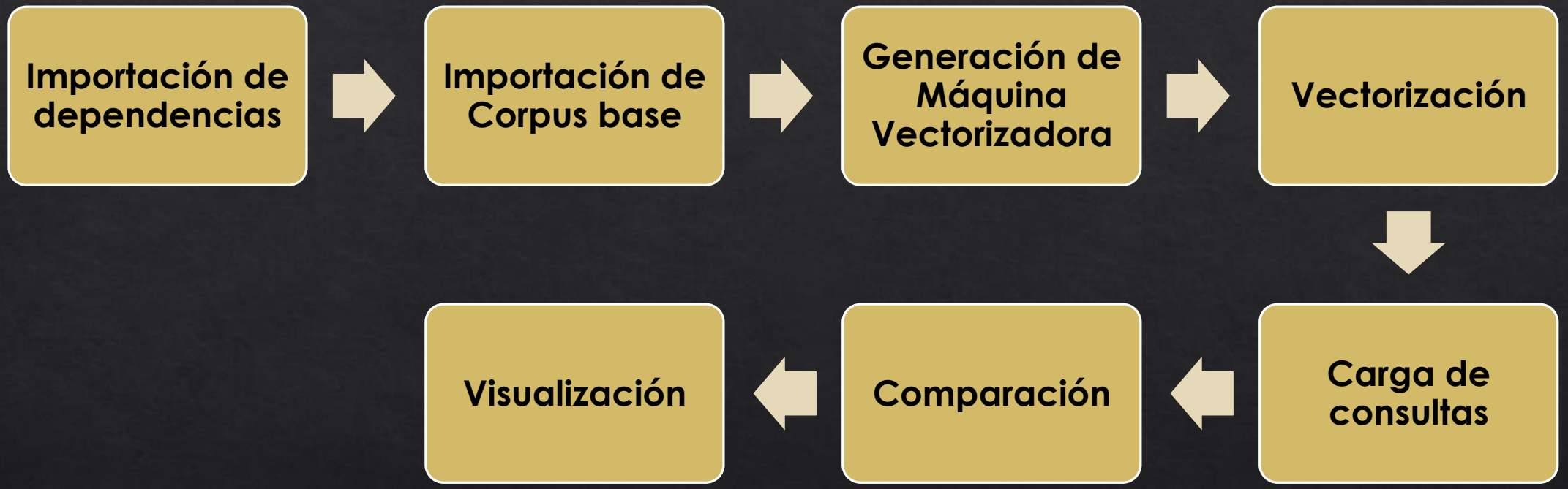
Fuente: Redes Neuronales, TensorFlow



Fuente: Doc2Vec, Word2Vec



Estructura de las soluciones



gensim
topic modelling for humans

GESTIO Dinámica

Dependencias

```
import warnings
warnings.filterwarnings(action='ignore', category=UserWarning, module='gensim')
import numpy as np
import nltk
from gensim.models.doc2vec import Doc2Vec, TaggedDocument
from nltk.tokenize import word_tokenize
import pandas as pd
import sys
import matplotlib.pyplot as plt
```

Código Python Gensim (Doc2Vec)

```
max_epochs = 100

for epoch in range(max_epochs):
    #print('iteration {0}'.format(epoch))
    model_g.train(tagg_glos,
                  total_examples=model_g.corpus_count,
                  epochs=model_g.iter)
    # decrease the learning rate
    model_g.alpha -= 0.0002
    # fix the learning rate, no decay
    model_g.min_alpha = model_g.alpha
```


```
model_g = Doc2Vec(
                  vector_size = 30,
                  alpha = 0.025,
                  min_alpha = 0.00025,
                  min_count = 1,
                  dm = 1
                  )
```

Código Python Gensim (Word2Vec)

```python
procesados = []
for elem in documentos['texto']:
    elem_p = gensim.utils.simple_preprocess(elem)
    procesados.append( list(set(elem_p)-set(sw_es)) )
```

```python
modelo = gensim.models.Word2Vec(
                                procesados,
                                size=50,
                                window=10,
                                min_count=2,
                                workers=10
                                )
```

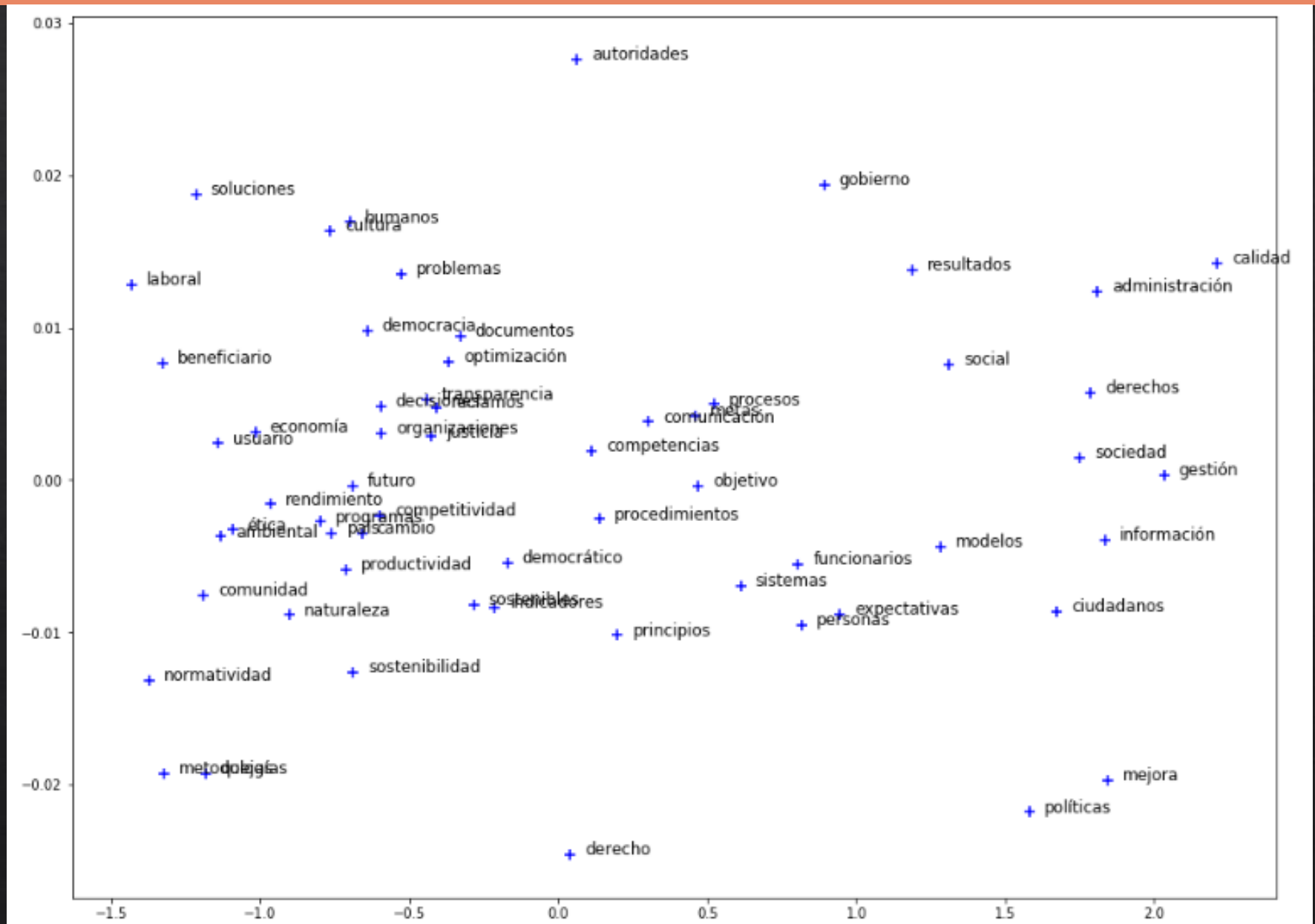
```python
modelo.train(
             procesados,
             total_examples=len(procesados),
             epochs=20
             )
```

gensim
topic modelling for humans

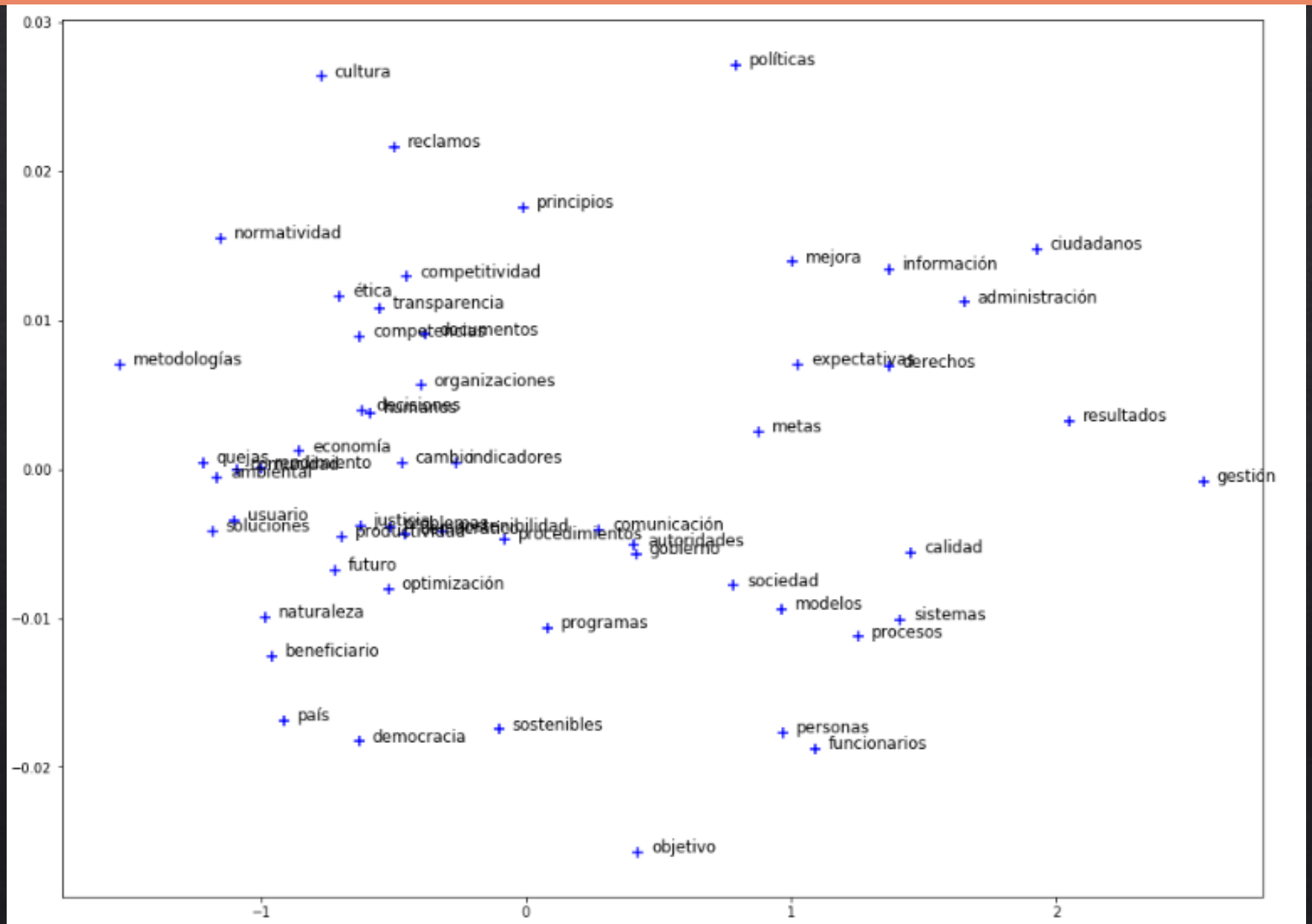
jupyter

GESTIO Dinámica

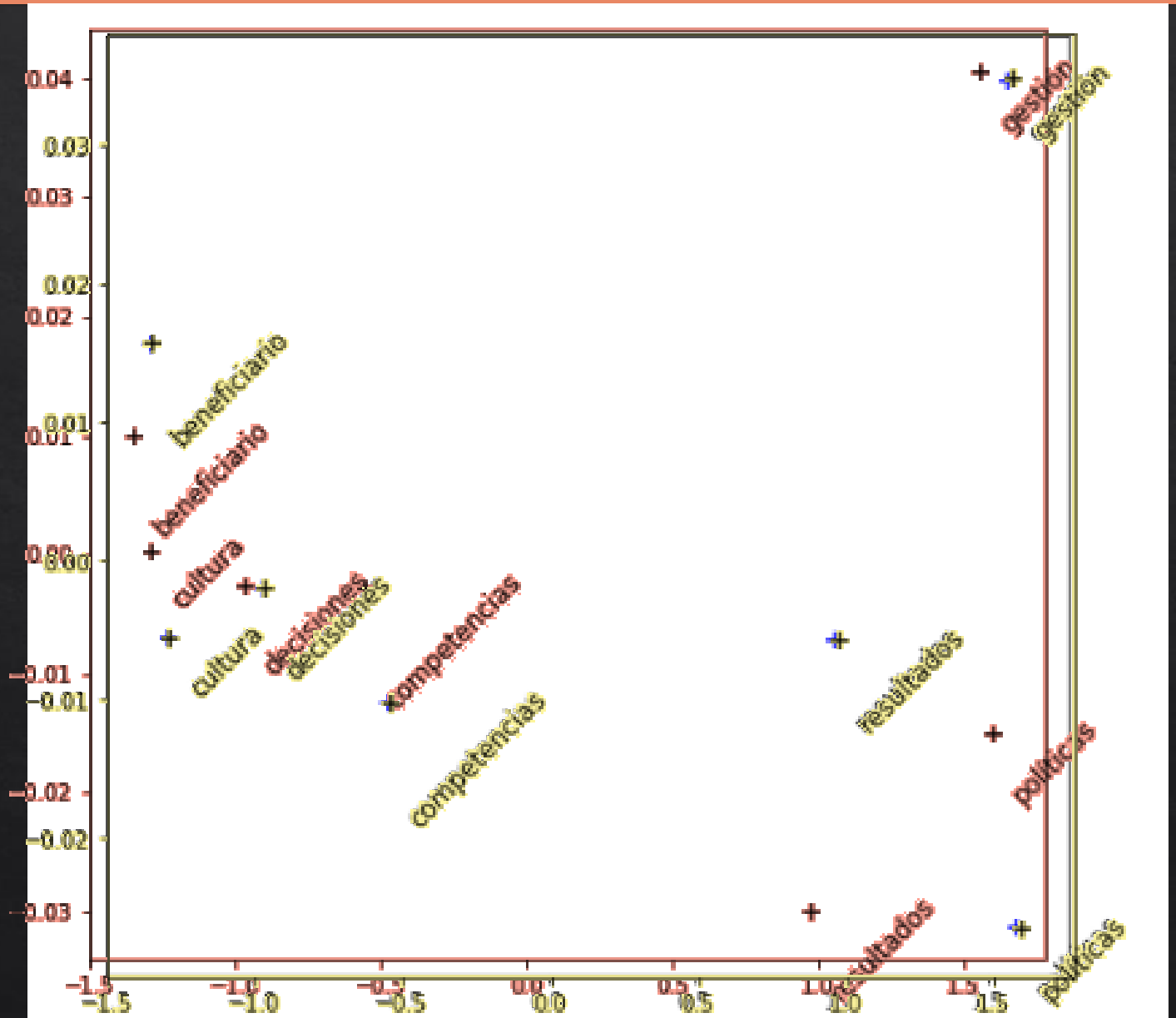
Resultados: Carta Iberoamericana (QGP)



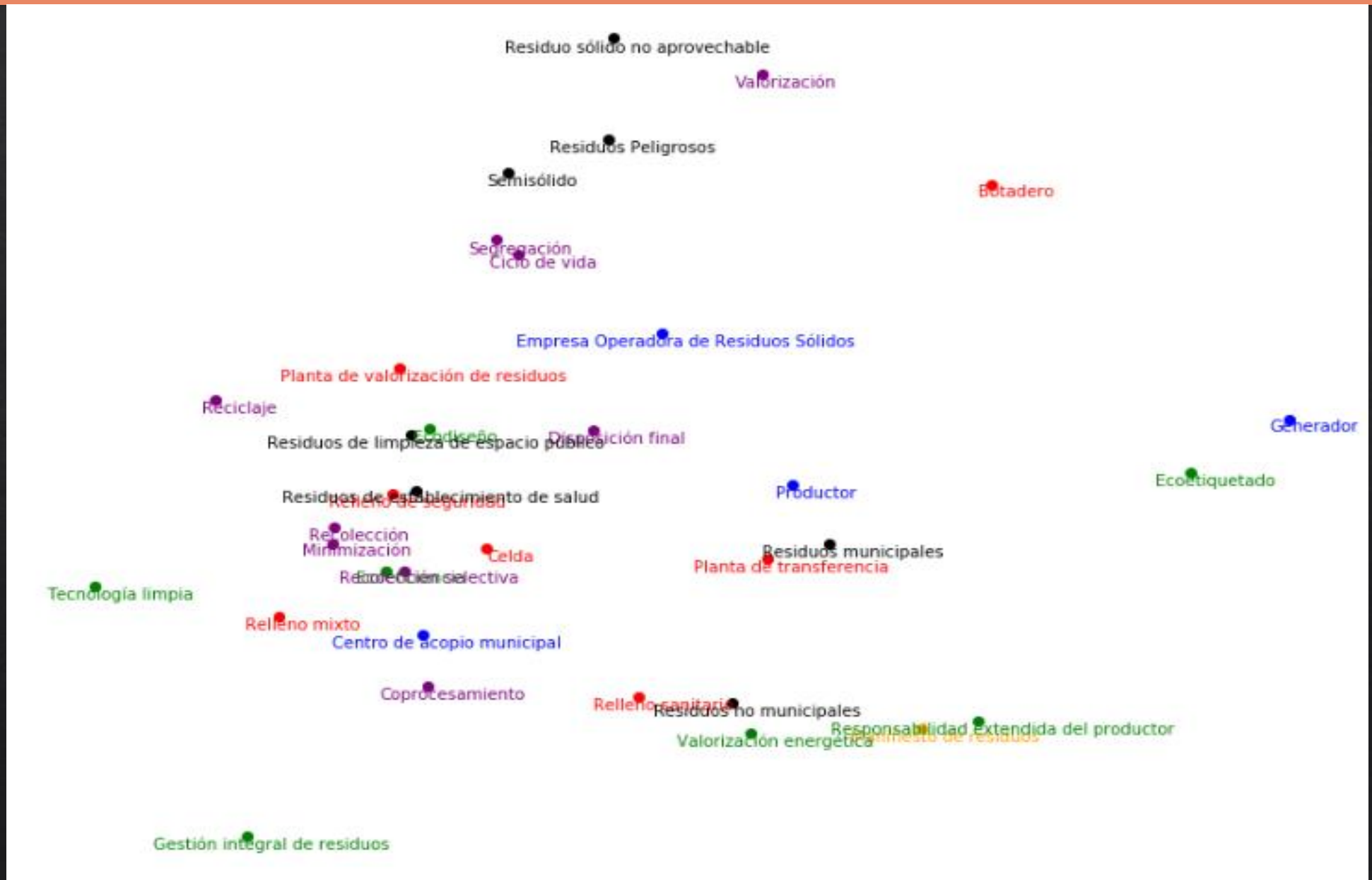
Resultados: Carta Iberoamericana (QGP)



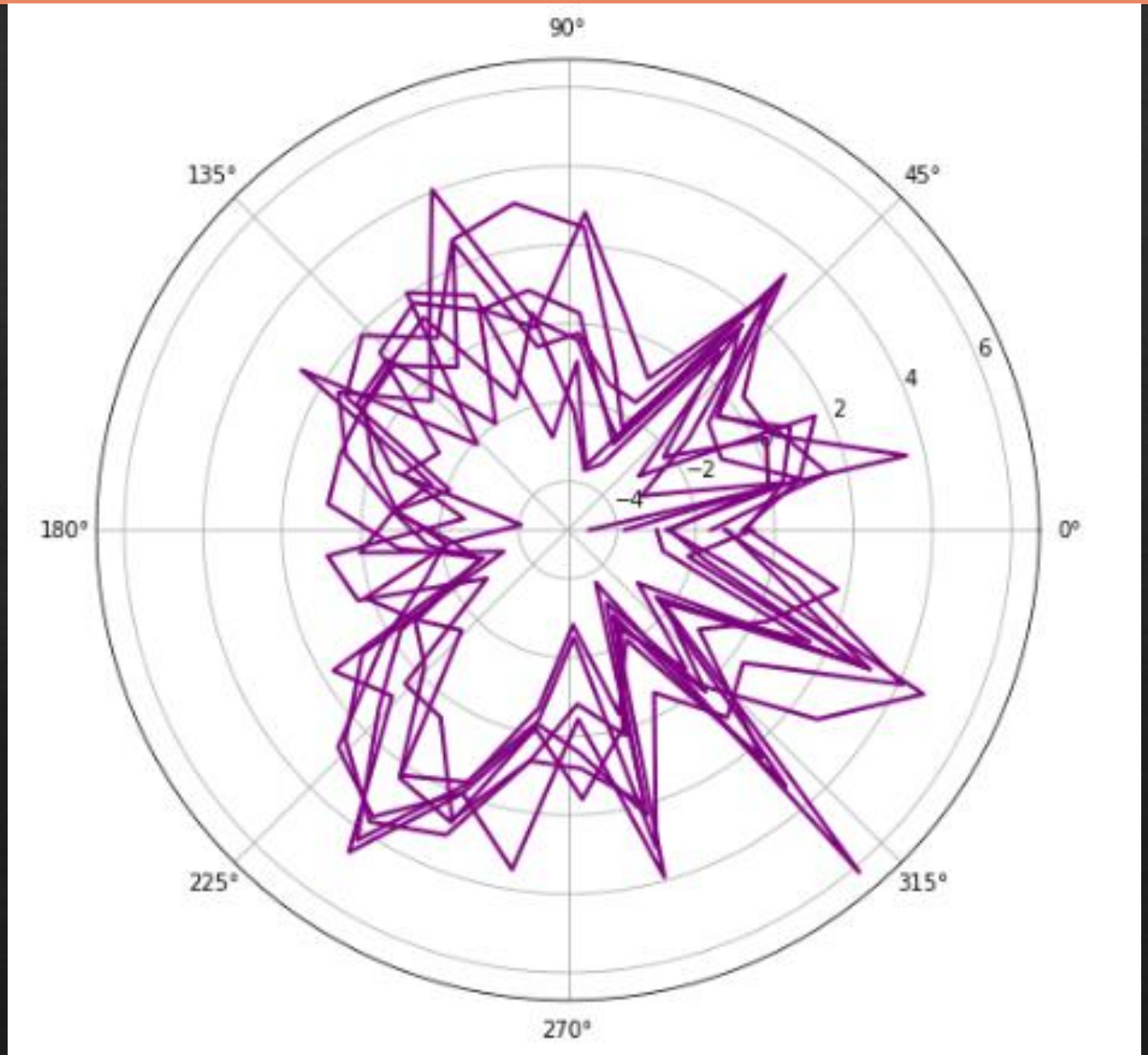
Resultados: Carta Iberoamericana (QGP)
Comparación



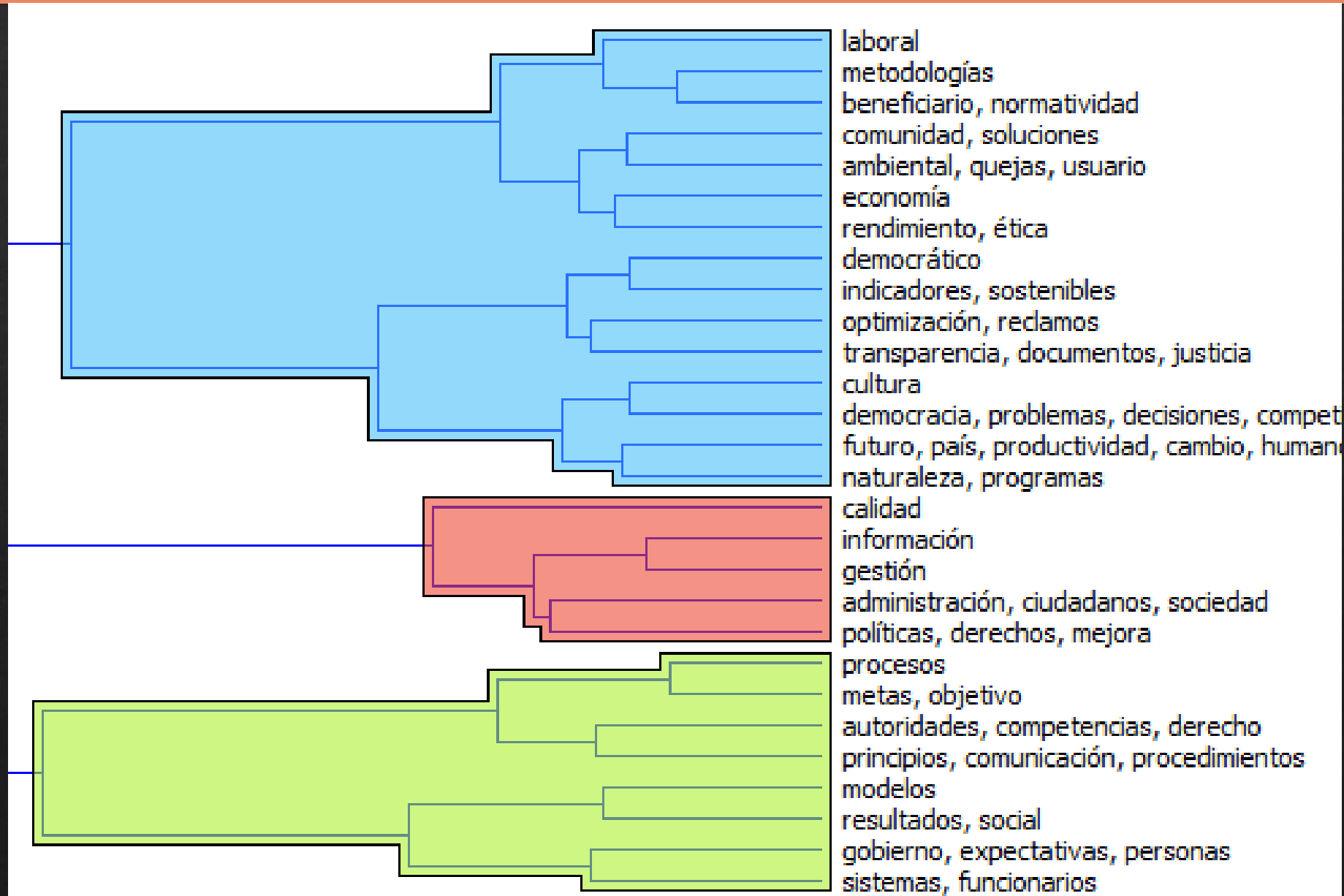
Resultados: Glosario Residuos Sólidos



Resultados: Glosario Residuos Sólidos



Resultados: Carta Iberoamericana (QGP)



Momento Eureka

Conclusión sobre Modelamiento Ad-Hoc

Comprar un Servicio en la Web

Desarrollar un Modelo Ad-Hoc

Potencial

- Qué distancia conceptual hay entre un documento y otro, o cómo se distribuyen documentos
- Chatbots especializados en relación a un conjunto específico de documentos (Conversar con mi libro o ley)
- Clasificación automática de contenidos en función de su propia estructura
- Calificación académica de ensayos
- Análisis de discursos para evaluar formas de pensar

Muchas gracias

@jcmachicao
@gestiodinamica
jcmachicao@gestiodinamica.com