Looking for similar patients

COVID-CaseExplore

FrancoArgentina Team: Paul Rognon, Diego Torres Dho •
10.01.2021

The Challenge

- A lot of medically relevant information is hidden in the huge mass of clinical cases.
- Finding clinically similar cases can help prognosis, diagnosis and decision making.
- Task: find similarities in a collection of clinical cases of COVID-19 and non COVID-19 patients.

Thanks to


#bitsxatató


BSC
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación


hackers
@UPC


50 ANYS
UPC
UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH
sense limits
FIB


Plan TL

Our idea: finding groups of similar cases with Deep Learning

Clusters to find patterns and decide

- Beyond pairwise similarity, clusters help characterizing the similarity of cases.
- New cases can be assigned to a group for decision-making.

Deep Learning to represent the cases

- Neural networks are currently the best tools to model text and language.
- Autoencoder like networks can learn a numerical representation of the text of our clinical cases.

Possible applications

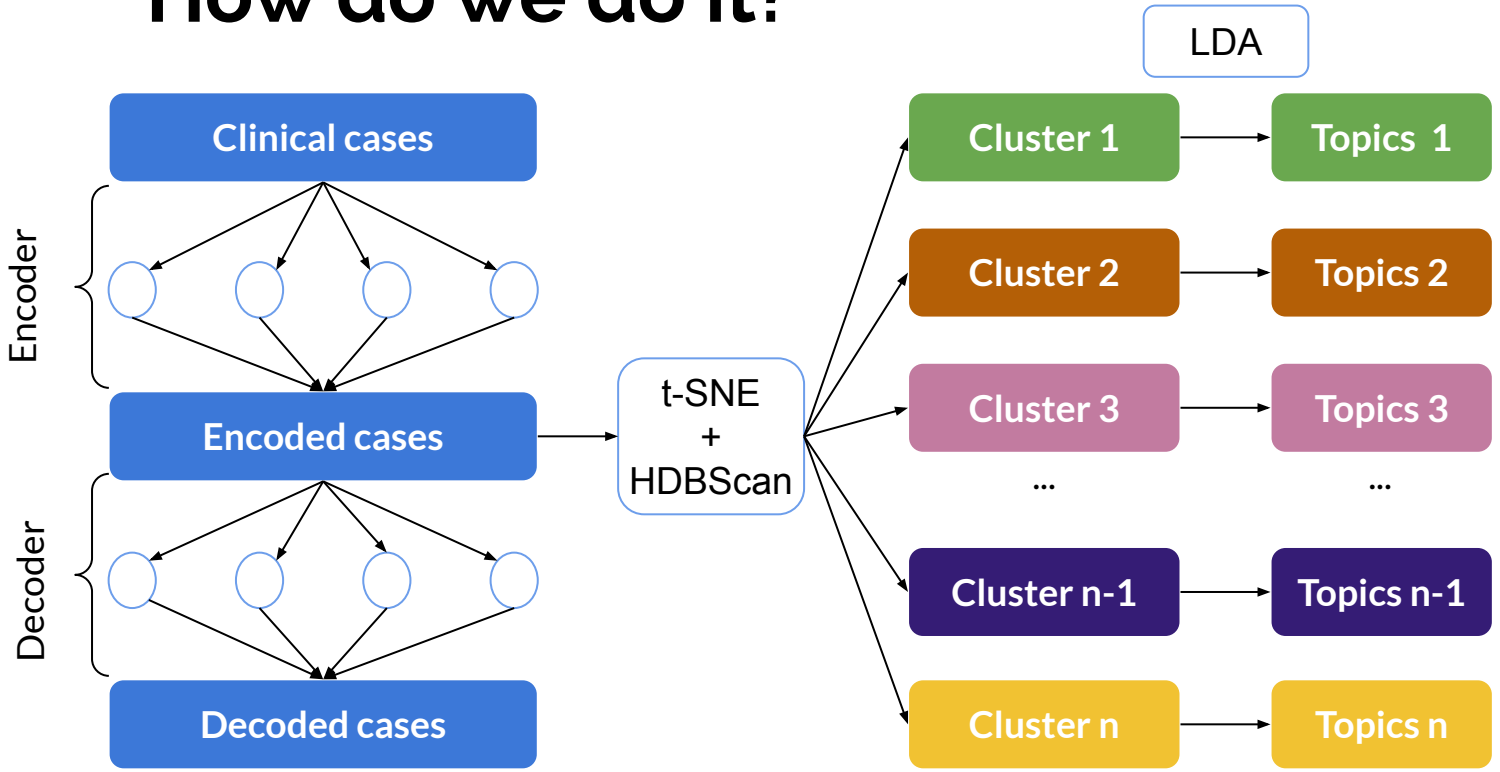
Knowledge discovery

- Find comorbidities as other types of disease can appear to characterize a cluster of COVID-19 cases.
- Find subtypes of COVID-19 cases that can be characterized by severity, progression speed, symptomes or other traits.

Taking decisions

- Assist medical professionals in triage by assigning a new case to a cluster characterized by the severity.
- Assist medical professionals in designing an adequate response to a new patient case.

How do we do it?



How do we do it?
Some technical details

Autoencoder

The encoder has 1 LSTM layer followed by a dense layer. The decoder has a bidirectional LSTM layer followed by a dense layer. We set the encoder output dimension at 64.

Clustering

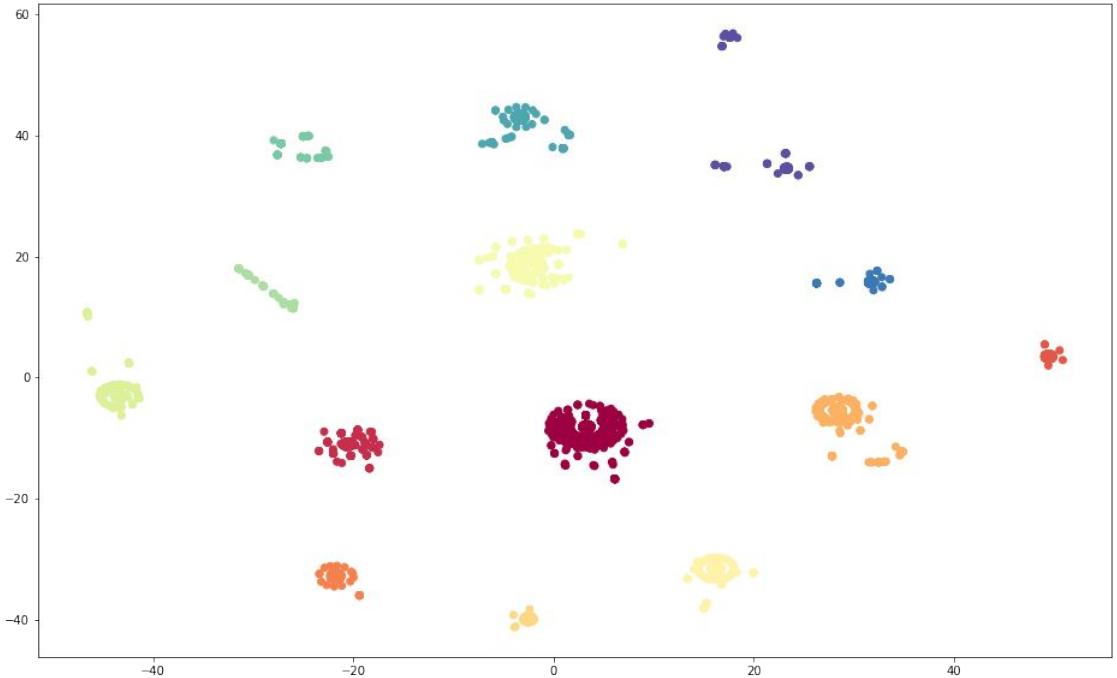
We first obtain a two dimensional t-SNE representation of the encoder output using cosine similarity as metric and Barnes-Hut method. We then run HDBScan.

Latent Dirichlet Allocation (LDA)

We first pre-process our data by tokenizing, stemming, removing stop words, very uncommon and common words. We then extract five 15-word dominant topics.

Some initial results

Clearly defined clusters



Two dimensional t-SNE representation of encoded cases

Informative clusters

- Our algorithm differentiates COVID-19 related clusters of cases from non-COVID-19 related clusters. In 3 clusters, COVID-19 is the most dominant or second most dominant topic.
- In 1 of those clusters, kidney diseases and male gender are two other salient traits in the dominant topics.
- 1 of those clusters is highly related with cancer terms along with COVID-19.

Let’s try it!

What's next for COVID-CaseExplore?

1. Build a strong pre-processing of case text to avoid clustering on non-significant word.
2. Use pre-trained word embeddings to improve the encoding of the cases.
3. Extract more significant knowledge from the clusters with the help of medical professionals.

About us

Paul Rognon

github.com/polrng

@PaulRgnn

linkedin.com/in/paul-rognon

Diego Torres Dho

github.com/dtorresdho

@dtorresdho

linkedin.com/in/dtorresdho