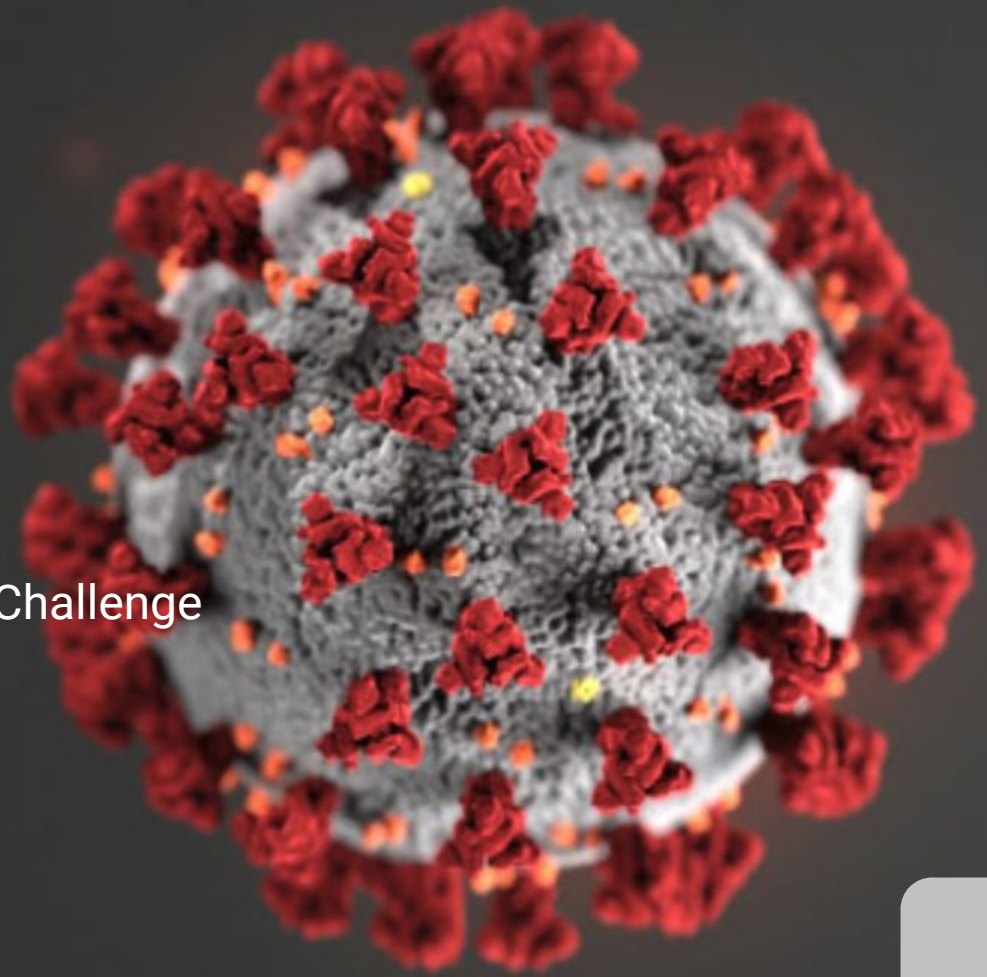
CORD-19

COVID-19 Open Research Dataset Challenge



Chao Zhou, Ruijin Jia, Matteo Bucalossi

Introduction

The White House has prepared with other partners an immense dataset of scholarly articles about coronaviruses.

The challenge is applying ML, DL, NLP tools to gain insights from the data and help the medical community face this new global emergency.

What do we want to know?

- **What is known about transmission, incubation, and environmental stability?**
- **What do we know about COVID-19 risk factors?**
- **What do we know about virus genetics, origin, and evolution?**
- **What do we know about vaccines and therapeutics?**
- **What has been published about medical care?**
- **What do we know about non-pharmaceutical interventions?**
- **What do we know about geographical distribution?**
- **What do we know about diagnostics and surveillance?**
- **What has been published about ethical and social science considerations?**
- **What has been published about information sharing and inter-sectoral collaboration?**

Transformers & Sentence Embeddings

Embeddings for Unsupervised Tasks

- BERT = bidirectional representations of language elements so that overall context is used when transforming seq2seq
- SciBERT = BERT model trained on scientific corpus of articles
- SBERT = modified BERT model for semantic sentence embeddings easier to compare for unsupervised tasks

Fine-tuned SBERT on SciBERT for domain-specific sentence embeddings for unsupervised tasks

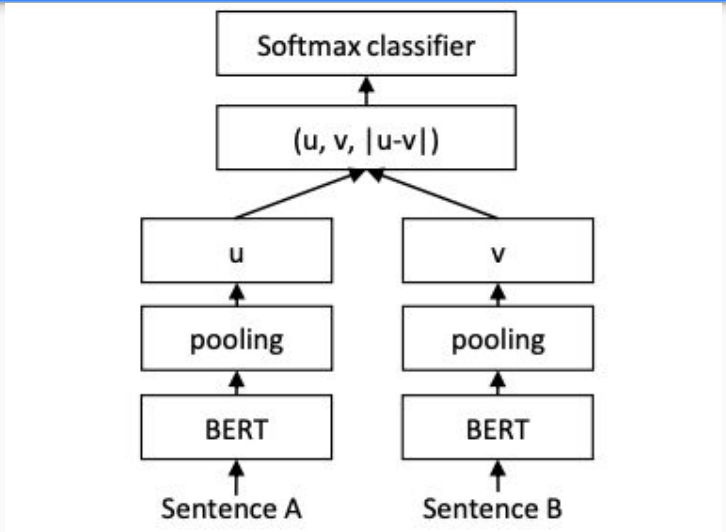


Figure 1: SBERT architecture with classification objective function, e.g., for fine-tuning on SNLI dataset. The two BERT networks have tied weights (siamese network structure).

```python
# select one Transformer
model_name = 'allenai/scibert_scivocab_uncased'

# Use sciBERT model for mapping tokens to embeddings
word_embedding_model = models.BERT(model_name)

# Apply mean pooling to get one fixed sized sentence vector
pooling_model = models.Pooling(word_embedding_model.get_word_embedding_dimension(),
                               pooling_mode_mean_tokens=True,
                               pooling_mode_cls_token=False,
                               pooling_mode_max_tokens=False)

model = SentenceTransformer(modules=[word_embedding_model, pooling_model])
```

Sequential model (sent to embedding)

Data loader

```python
# Convert the dataset to a DataLoader ready for training
train_data = SentencesDataset(nli_reader.get_examples('train.gz'), model=model)
train_dataloader = DataLoader(train_data, shuffle=True, batch_size=batch_size)
train_loss = losses.SoftmaxLoss(model=model, sentence_embedding_dimension=model.get_sentence_embedding_dimensio
n(), num_labels=train_num_labels)
```

Dev-set

```python
dev_data = SentencesDataset(examples=sts_reader.get_examples('sts-dev.csv'), model=model)
dev_dataloader = DataLoader(dev_data, shuffle=False, batch_size=batch_size)
evaluator = EmbeddingSimilarityEvaluator(dev_dataloader)
```

Training SBERT

```python
num_epochs = 2
warmup_steps = math.ceil(len(train_dataloader) * num_epochs / batch_size * 0.1)
# Train the model
model.fit(train_objectives=[(train_dataloader, train_loss)],
          evaluator=evaluator,
          epochs=num_epochs,
          evaluation_steps=1000,
          warmup_steps=warmup_steps,
          output_path=model_save_path
          )
```

Clustering

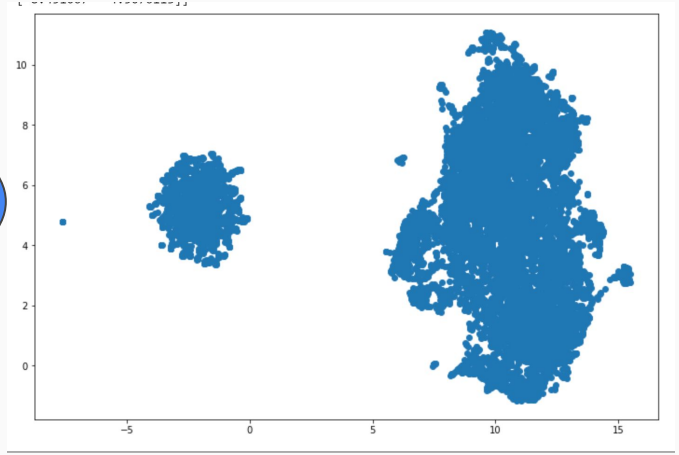
UMAP: manifold dimensionality reduction

HDBSCAN: hierarchical version of density-based clustering model

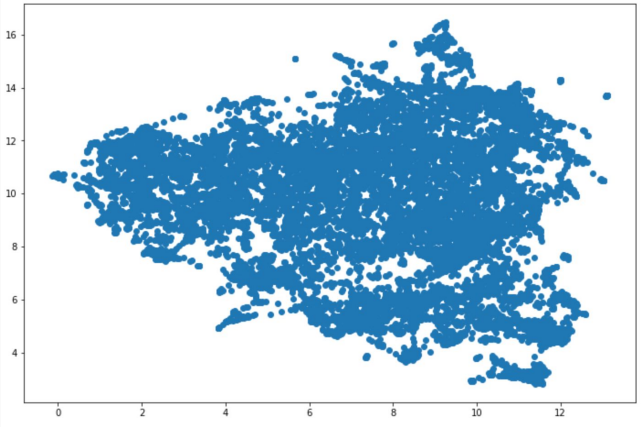
K-means: model of vectorization to identify **10** clusters by nearest mean

LDA: topic modeling by probabilistic-based model

Dimensionality Reduction using UMAP for Abstract



Dimensionality Reduction using UMAP for Body Text



Clustering by HDBSCAN

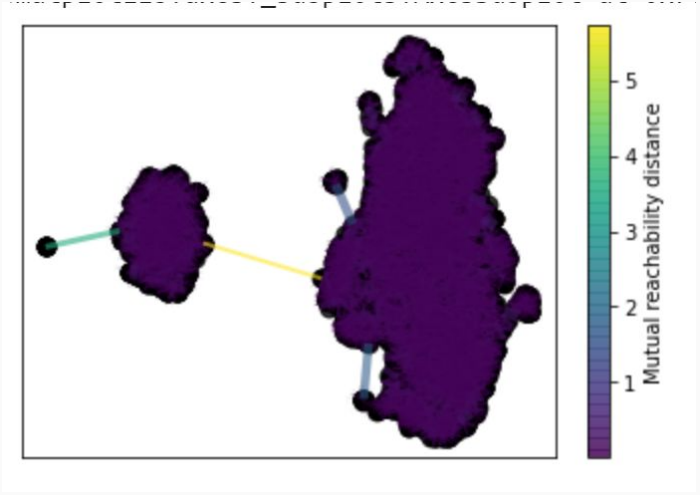
HDBSCAN: hierarchical version of density-based clustering model

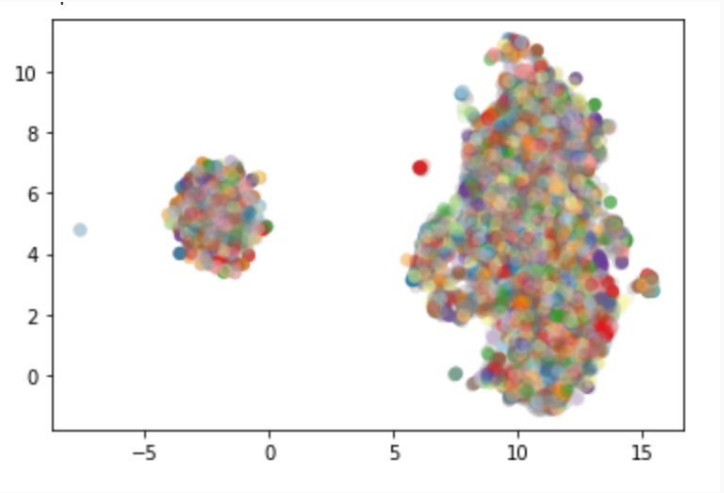
Clustering Abstract:

➢ More than 1000 labels

➢ Failed to build the cluster hierarchy

Condense, and extract the cluster tree

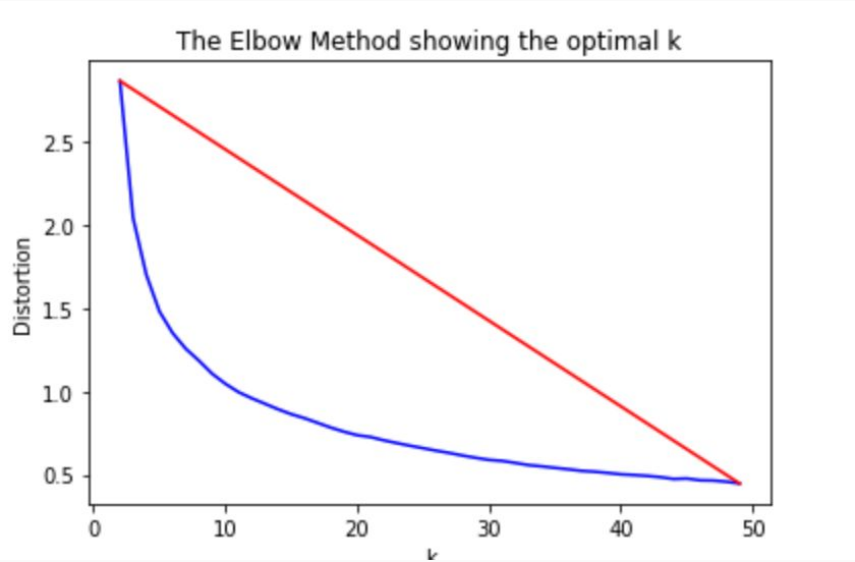




Clustering by K-Means

Clustering Abstract:

➢ Best k value 10




```
[ ] from scipy.spatial.distance import cdist
    distortions = []
    K = range(2, 50)
    for k in K:
        k_means = KMeans(n_clusters=k, random_state=42).fit(clusterable_embedding)
        k_means.fit(clusterable_embedding)
        distortions.append(sum(np.min(cdist(clusterable_embedding, k_means.cluster_centers_, 'euclidean'), axis=1)) / pd.DataFrame(data3).shape[0])
```

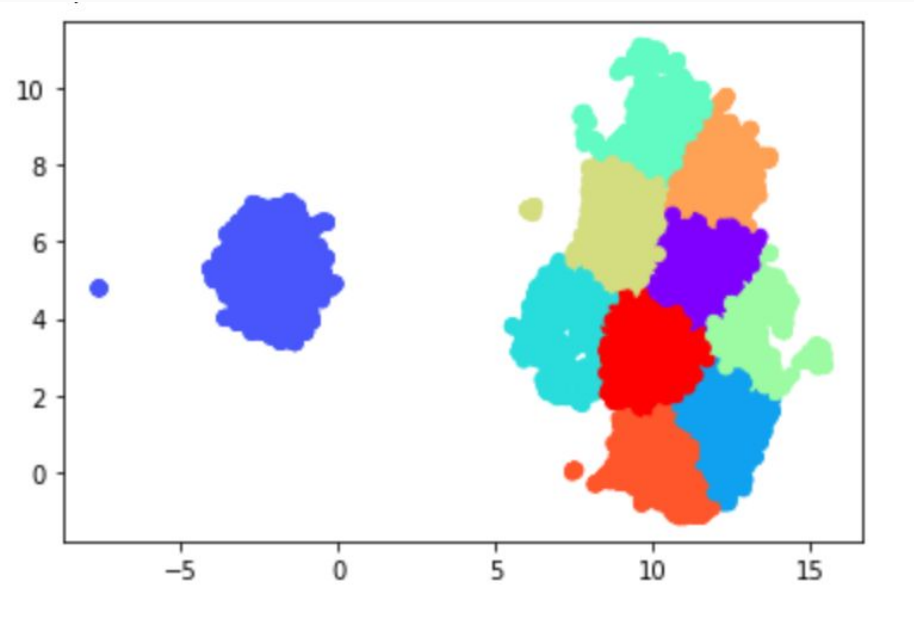
Clustering by K-Means

Why K-Means?

➢ K-Means works well for “round” or spherical

➢K-Means works well for most dense in the center of the sphere

➢Data does not contain much noise/outliers.



Future Work...

Topic Modeling by:

➤ LDA

➤ NPM

...

Embedding Ways:

➤ TFIDFVectorizer

➤ CountVectorizer

Sample Work

	0
101	Background: Air pollution has a significant he...
463	The main purpose of this study was to investig...
129	Animal viruses and bacteria are ubiquitous in ...
321	Background: Zika virus infection has recently ...
3	A survey was conducted into respiratory infect...
199	Background: The widespread forest fires in Ind...
341	Vector-borne infectious diseases, such as mala...
99	Background: Wearing a pollution mask is an eff...
2	Prevention of serious infections in pregnant m...
513	Background: Asthma is a major public health pr...

Semantic Search

Cosine similarity

We used a simple algorithm to retrieve the top 5 most relevant articles to an user query by identifying the most similar embeddings among all articles to the query embedding.

Here's an example!

System can use both the embeddings of abstracts or of the full article texts (as you prefer!)

How about transmission dynamics of the virus?

Top 5 most similar articles in corpus:

Abstract	Title	Journal	Score
a number of virologic and	Mechanisms of viral emergence	Vet Res	0.7622
the emergence of zika virus zikv	Experimental Zika virus infection of Jamaican fruit bats (Artibeus jamaicensis) and possible entry of virus into brain via activated microglial cells	PLoS Negl Trop Dis	0.7237
interspecies transmission of pathogens may	The Application of Genomics to Emerging Zoonotic Viral Diseases	PLoS Pathog	0.7086
endemic and seasonally recurring respiratory viruses are a	Surveillance of respiratory viruses in the outpatient setting in rural coastal Kenya: baseline epidemiological observations	Wellcome Open Res	0.6995
rubella virus rv has been reported to	The rubella virus E2 and E1 spike glycoproteins are targeted to the Golgi complex	J Cell Biol	0.6776

And what about related risk factors?

Top 5 most similar articles in corpus:

Abstract	Title	Journal	Score
in addition to protective	Sepsis and septic shock: endothelial molecular pathogenesis associated with vascular microthrombotic disease	Thromb J	0.7072
to examine the impacts of a multi	Effectiveness of Integrated HIV Prevention Interventions among Chinese Men Who Have Sex with Men: Evaluation of a 16-City Public Health Program	PLoS One	0.7042
offi ce,	European Hedgehogs as Hosts for Borrelia spp., Germany	Emerg Infect Dis	0.7039
interspecies transmission of pathogens may	The Application of Genomics to Emerging Zoonotic Viral Diseases	PLoS Pathog	0.7012
background surveillance and intervention are resource	Conceptualising the technical relationship of animal disease surveillance to intervention and mitigation as a basis for economic analysis	BMC Health Serv Res	0.7002

RECAP

1. Train SBERT on SciBERT and NLI dataset with Softmax
2. Use this fine-tuned model to get sentence-embeddings
3. Use these embeddings for:
    a. Clustering (after UMAP)
        i. HDBSCAN
        ii. K-means
        iii. LDA
    b. Semantic search

Maybe.. machines can truly help humans face nowadays challenges and succeed!

Let’s hope this Call to Action will bring tangible impact to research

Thank you for your attention!