Advanced Machine Learning - Exercise 3
Covid-19 Challenge

SPRING SEMESTER 2022

Lecturer: Dr. Shay Fine

by Aviad Ariel 066149428, Dana Perlstein302282538

1. Instructions

The used code and notebooks are available via our github repo.

2. Data Preparation

According to the instructions, the latest 20K articles were taken. The title, abstract and body are saved into a DataFrame, then saved into csv to be used later on.

```
[ ] covid_df
```

	title	cord_uid	abstract	body
0	Blockchain-based governance models for COVID-1...	ppfxi5id	This paper analyses the requirements of a bloc...	Within the existing literature, papers both ad...
1	On intelligent agent-based simulation of COVID...	uyf9ds7s	COVID-19 has impacted all areas of human activ...	Over the past decades, significant changes hav...
2	Concern with COVID-19 pandemic threat and atti...	d0s0f0t1	Tightening social norms is thought to be adapt...	Tightening social norms is thought to be adapt...
3	An antifragile strategy for Rome post-Covid mo...	ct7nc16b	We are aware that we will have to live with CO...	Since exactly one year, COVID has changed our ...
4	COVID-19 Time Series Forecasting – Twenty Days...	bs206r15	The new Coronavirus, responsible for the COVID...	One of the most issues addressed in 2020 and 2...
...	...	...	...	...
19995	Patient and Provider Experience With Cystic Fi...	l1y1ezfo	In response to the novel coronavirus (COVID-19...	On March 11, 2020 the novel coronavirus diseas...
19996	Association between voriconazole exposure and ...	lccgk110	Therapeutic drug monitoring (TDM) is essential...	Therapeutic drug monitoring (TDM) is essential...
19997	Network Pharmacology-Based Analysis of Pogoste...	8ehcnyp5	Nonalcoholic fatty liver disease (NAFLD) is th...	Nonalcoholic fatty liver disease (NAFLD) is a ...
19998	A Novel Approach to the Viability Determinatio...	htzqgwp6	Mycobacterium avium subsp. paratuberculosis (M...	Mycobacterium avium subsp. paratuberculosis (M...
19999	A Herbal Mixture Formula of OCD20015-V009 Prop...	p0ztyrb3	OCD20015-V009 is an herbal mix of water-extrac...	The genome of the influenza A virus (IAV) cont...

20000 rows × 4 columns

For more detailed information see data_preparation.ipynb

3. Data Exploration and Preparation

Exploring the data, displaying WordCloud of the titles, abstracts and bodies of the papers:

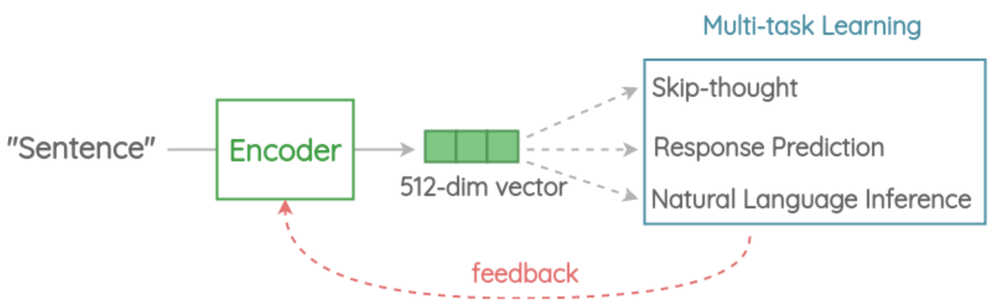
An additional data process of removing stopwords for the abstract and bodies is done. For more detailed information see data_exploration_and_processing.ipynb

4. Compression Method

Universal Sentence Encoder (USE) is our selected method of compression.

a. Main Idea

Design an encoder that summarizes a given sentence to a 512-dimensional sentence embedding. This embedding is used to solve multiple tasks and based on the mistakes it makes on those, the sentence embedding is updated.



b. Process

The process uses PTB tokenizer, the encoder uses Transformer or DAN architecture (depending on the loaded model).

c. Similarity & Clustering

Each of the sentences are encoded to 512 length vectors, then cosine similarity is calculated as a measure of distance. After the sentences are encoded, we use K-Means in order to cluster them in 512 dimensions, then for display we use PCA (2D) for dimensionality reduction. For more detailed information (as well as synthetic examples) see universal_sentence_encoder.ipynb

5. Paper Similarity

The processed data is loaded, then we use our utility class to calculate similarities (as explained above). The body and text are encoded via the Transformer architecture, and the title uses the DAN architecture. The first index was selected to be compared to, its title:

```
covid_df.iloc[0].title
'Blockchain-based governance models for COVID-19 digital health certificates: A legal, technical, ethical and security requirements analysis'
```

“Blockchain-based governance models for COVID-19 digital health certificates: A legal, technical, ethical and security requirements analysis”. We can see that it is relevant to Blockchain.

a. Body Similarities

After embedding the body and calculating similarities, we sort the similarities in descending order and pull the top K instances. The top K=4 similar instances:

```
indx, similarities = sentence_util_body.get_k_most_similar(compared_index=0, k=4)
indx, similarities

(array([11522,  3037,  2257,  9617]),
 array([0.78430235, 0.7703583 , 0.7508329 , 0.748494  ], dtype=float32))
```

By inspecting their titles we can see that they are also related to Blockchain

```
[23] pd.set_option('display.max_colwidth', None)
     covid_df.iloc[indx].title

11522                                        Blockchain-based Platform for Secure Sharing and Validation of Vaccination Certificates
3037                                                  A Systematic Literature Review of Blockchain Technology Adoption in Bangladesh
2257     Perceived Security Risk Based on Moderating Factors for Blockchain Technology Applications in Cloud Storage to Achieve Secure Healthcare Systems
9617                                                        BEAT: Blockchain-Enabled Accountable Infrastructure Sharing in 6G and Beyond
Name: title, dtype: object
```

b. Abstract Similarities

Repeating the same process for the abstracts, we get:

```
sentence_util_abstract = SentenceUtil(covid_df.processed_abstract, module_url="https://tfhub.dev/google/universal-sentence-encoder/4")
```

```
module https://tfhub.dev/google/universal-sentence-encoder/4 loaded
100%|██████████| 19999/19999 [03:35<00:00, 92.72it/s]
```

```
indx, similarities = sentence_util_abstract.get_k_most_similar(compared_index=0, k=4)
indx, similarities
```

```
(array([ 6378,  4301, 11838,  4472]),
 array([0.65570027, 0.6114664 , 0.61127365, 0.5858599 ], dtype=float32))
```

```
covid_df.iloc[indx].title
```

```
6378                Blockchain Matters—Lex Cryptographia and the Displacement of Legal Symbolics and Imaginaries
4301                                                      Cybersecurity, Data Privacy and Blockchain: A Review
11838                                 Know Your Customer: Balancing Innovation and Regulation for Financial Inclusion
4472     Research contributions and challenges in DLT-based cryptocurrency regulation: a systematic mapping study
Name: title, dtype: object
```

Index 11838 is the exception, not having blockchain mentioned in its title, but further investigation finds that it is related to Blockchain.

c. Combined Similarities

We decided to use body, abstract and title similarities, in a linear combination with different coefficients in order to get the final similarity.

```
C_title = 0.4
C_abstract = 0.5
C_body = 0.1
```

```
combined_similarities = C_title*sentence_util_title.similarity + C_abstract*sentence_util_abstract.similarity + C_body*sentence_util_body.similarity
```

```
def get_k_most_similar(similarity, compared_index, k):
    topk_ind = similarity[compared_index, :].argsort()[-(k + 1):][::-1][1:]
    return topk_ind, similarity[compared_index, :][topk_ind]
```

```
indx, similarities = get_k_most_similar(combined_similarities, compared_index=0, k=4)
indx, similarities
```

```
(array([11522,  4512,  2257, 16642]),
 array([0.54065406, 0.4891051 , 0.47174174, 0.4547489 ], dtype=float32))
```

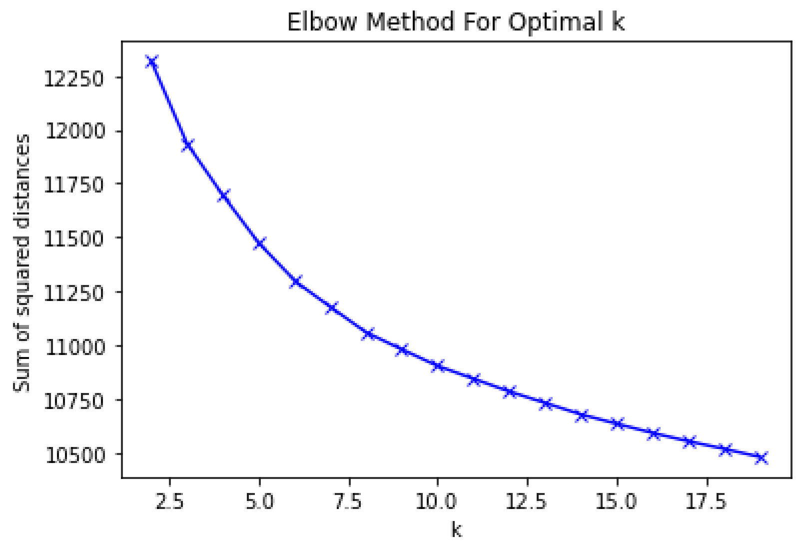
```
covid_df.iloc[indx].title
```

```
11522                                        Blockchain-based Platform for Secure Sharing and Validation of Vaccination Certificates
4512                                         Cyber governance studies in ensuring cybersecurity: an overview of cybersecurity governance
2257     Perceived Security Risk Based on Moderating Factors for Blockchain Technology Applications in Cloud Storage to Achieve Secure Healthcare Systems
16642                                                           A Privacy-Preserving Platform for Recording COVID-19 Vaccine Passports
Name: title, dtype: object
```

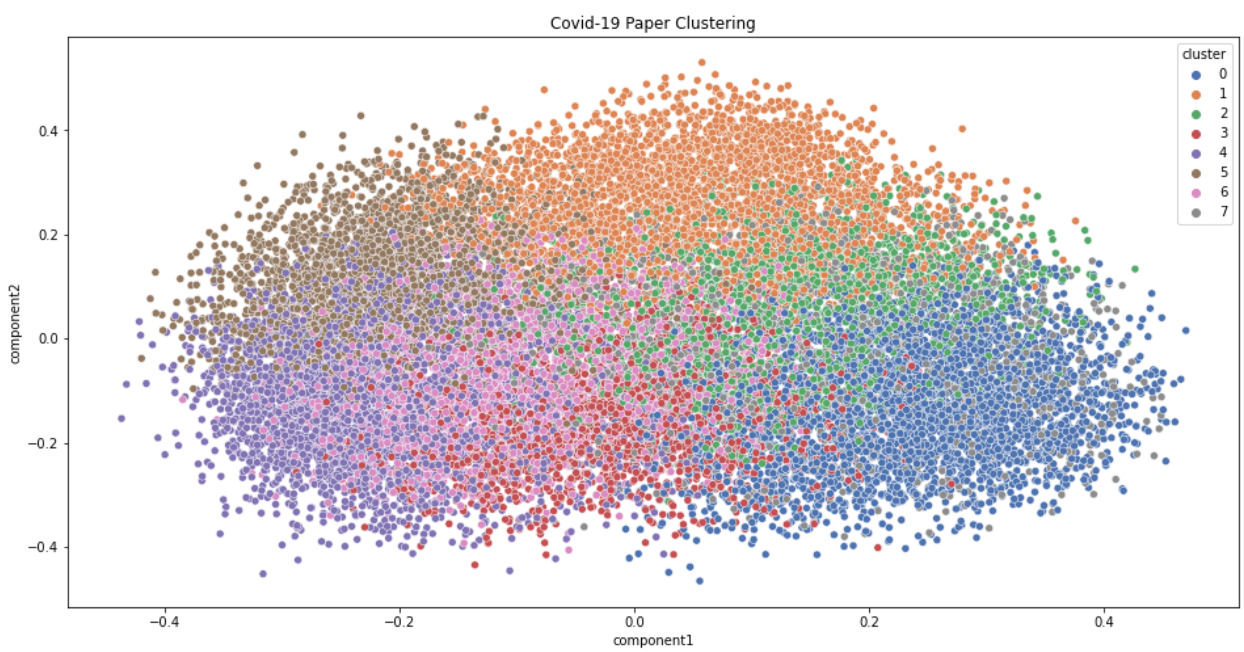
further investigations show that they seem to be related. For more detailed information see paper_similarity.ipynb

6. Paper Clustering

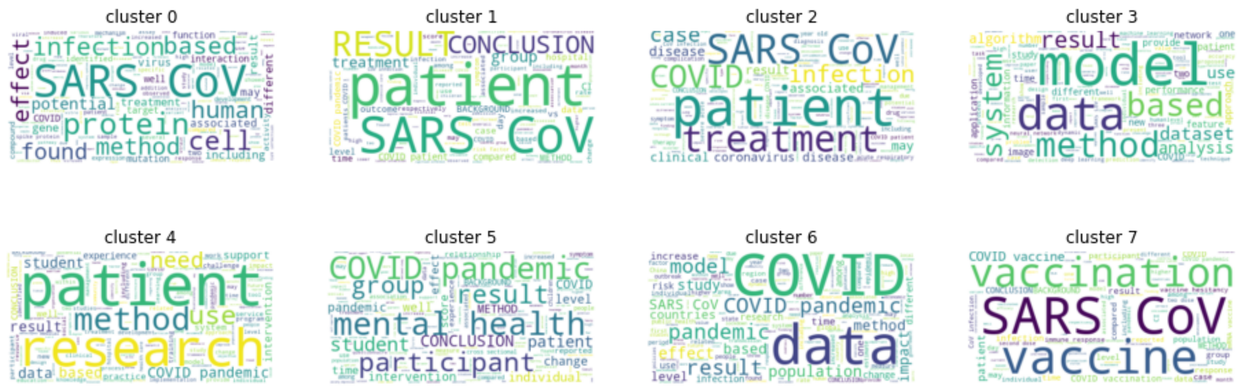
We use the processed abstracts in order to cluster the papers (in the method explained above). Using the elbow method to find the best K:



From the elbow we can see that the best K=8. using that K to cluster then performing PCA for display:



We can actually see that the clusters are grouped. Using WordCloud on the clusters:



Then using LDA on each cluster for topic modeling.

```python
for i in range(k_clusters):
    print('')
    print('Topics for Cluster ' + str(i) + ':')
    tm[i].print_topics()
```

```
Topics for Cluster 0:
topic 0 | resistance infections bacterial antimicrobial pathogens resistant antibiotic antibiotics bacteria pathogen
topic 1 | sars-cov-2 viral protein infection virus covid-19 proteins human host binding
topic 2 | samples study analysis results gene supplementary rna genes information available
topic 3 | enzymes orf6 glycocalyx consuming nad n-glycans proteins proteoglycans shield horseshoe
topic 4 | metabolic metabolism mitochondrial hiv liver brain lipid stress glucose hiv-1
topic 5 | dna autophagy modification methylation maternal quercetin proteins dengue allosteric myricetin
topic 6 | activity compounds properties potential study drug applications review different showed
topic 7 | sars-cov-2 variants spike mutations variant antibodies omicron detection antibody covid-19
topic 8 | fold curcumin nlrp3 inflammasome lungs activation infections surfactant increase reduced
topic 9 | cells cell expression immune patients mice levels cancer response inflammatory

Topics for Cluster 1:
topic 0 | oral dental efficacy treatment trial procedures bone aes lesions active
topic 1 | group time treatment clinical surgery care mean total significant patient
topic 2 | kidney aki antibiotic antibiotics renal ckd use resistance urinary bacterial
topic 3 | mortality risk group disease associated clinical age outcomes higher severe
topic 4 | respiratory influenza support children infections covid los hrv pneumonia viral
topic 5 | cases infection sars-cov-2 variant vaccination delta variants individuals infections infected
topic 6 | sars-cov-2 positive samples infection test viral testing negative sensitivity antibody
topic 7 | pandemic health care data years children age medical associated patient
topic 8 | group levels blood significantly lung parameters function higher disease ratio
topic 9 | women des les fracture eye hip maternal birth pregnant une
```

For more detailed information see [paper_clustering.ipynb]()