STAATLICH
ANERKANNTE
HOCHSCHULE

A Semantic QA Framework for Knowledge Discovery from CORD-19

Presented by :

Manmeet Kumar Chaudhuri, 11011780

HOCHSCHULE
SRH HEIDELBERG
Intelligence in Learning

DAS CORE PRINZIP

Content

STAATLICH
ANERKANNTE
HOCHSCHULE

STAATLICH
ANERKANNTE
HOCHSCHULE

Simplifying the Thesis Title

Thesis Title

STAATLICH ANERKANNTE HOCHSCHULE

4

- Framework takes advantage of semantics at word level as well as sentence level

1

- The process of extracting useful information from data
- Knowledge discovery from the perspective of Information retrieval from Big Data

A Semantic QA Framework for Knowledge Discovery from CORD-19

- A methodology for Question Answering (QA) task in the domain of NLP
- QA is a specific type of knowledge retrieval

2

- Big Data that comprises of textual data

4

STAATLICH
ANERKANNTE
HOCHSCHULE

Motivation and Research Questions

Motivation

STAATLICH
ANERKANNTE
HOCHSCHULE

- **Identify** a real world problem statement/use case
- **Read** scientific resources/publications to build knowledge in the domain of NLP
- **Understand** the state-of-the-art AI and Deep Learning methods in NLP
- **Learn** to use these methods in the context of QA task
- **Apply** the learning in solving the identified problem statement/use case
- **Disseminate** the knowledge to others

Problem Statement: Background

STAATLICH
ANERKANNTE
HOCHSCHULE

- COVID-19 pandemic
- Allen Institute of AI and some leading research groups have released the COVID-19 Open Research Dataset (CORD-19)[1]
- CORD-19 comprises over 100,000 scholarly articles, including around 60,000 papers with full text, about COVID and the coronavirus family of viruses
- Growing need for AI-based approaches to help scientific community in Knowledge Discovery from ever growing CORD-19

Problem Statement: Questions

STAATLICH ANERKANNTE HOCHSCHULE

1. **Which Deep Learning NLP architecture should be selected for the KD task from CORD-19?**
   KD, in the context of this Thesis, is limited to the QA task

2. **Can an AI Framework be developed incorporating the selected DL NLP architecture to retrieve or discover knowledge from CORD-19?**
   Retrieving the knowledge is limited to identifying relevant publications, sentences from the corpus and creating knowledge graphs

Use the Framework developed to answer the following user-questions (indicative examples):

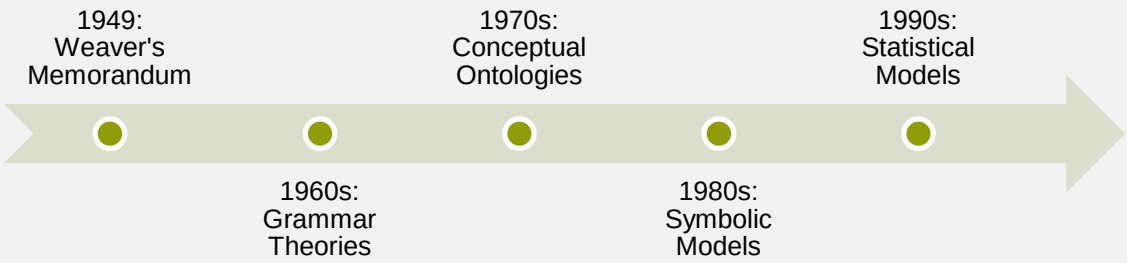
- What is known about the origin of virus and management measures at the human-animal interface?
- What is known about transmission, incubation, and environmental stability?
- What has been published about medical care?
- What is known about COVID-19 risk factors?
- What has been published regarding research and development and evaluation efforts for developing vaccines and therapeutics?

Which Deep Learning NLP architecture should be selected for the KD task from CORD-19?

Theoretical Background of NLP

NLP: Before Deep Learning Era

STAATLICH ANERKANNTE HOCHSCHULE

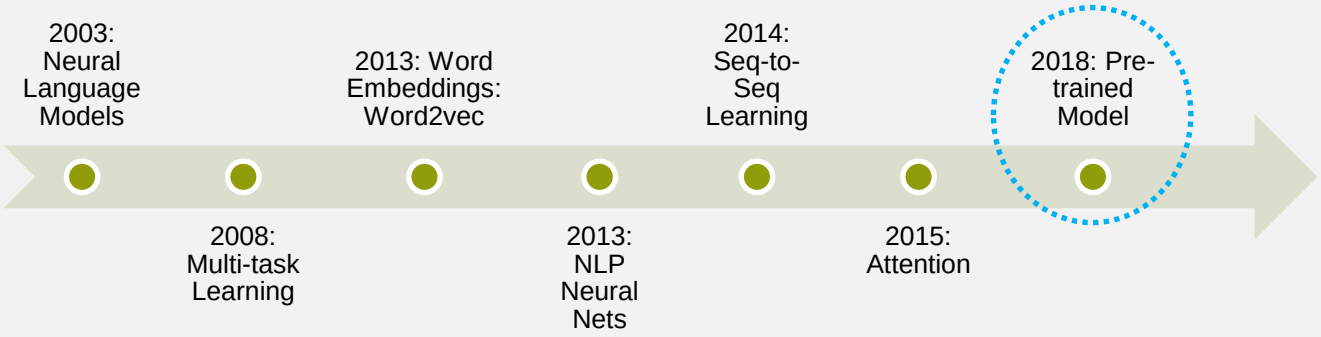


Constraint 1: Rule based complex statistical methods with handwritten rules

Constraint 2: Not effective in capturing semantics and context

NLP: After Deep Learning Era

STAATLICH ANERKANNTE HOCHSCHULE



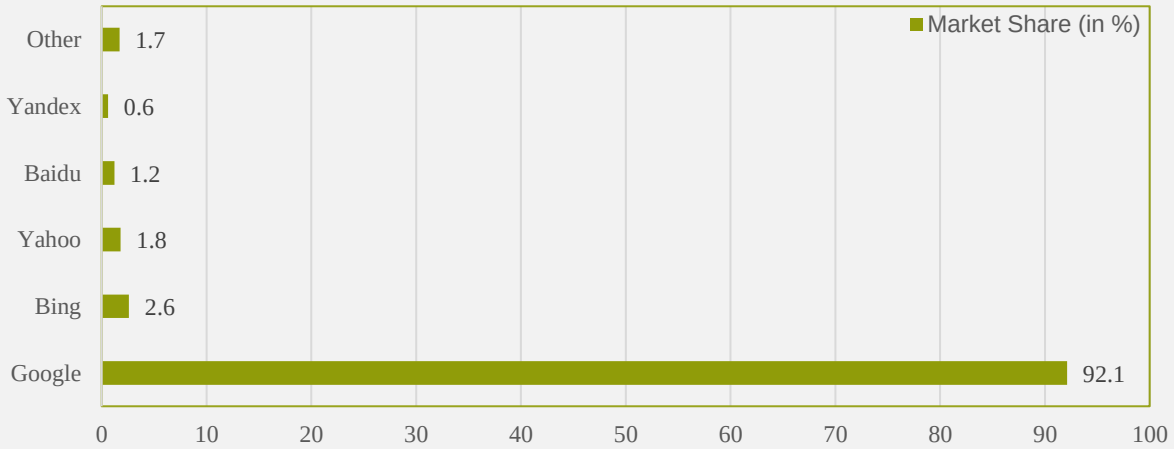
Examples of prominent Pre-trained language representation models: Embeddings from Language Model (ELMO), Generative Pre-trained Transformers (GPT), and Bidirectional Encoder Representations from Transformers (BERT)

Choice of Pre-trained Deep Learning Model

Choice of BERT Model

STAATLICH ANERKANNTE HOCHSCHULE

1. BERT: Driver of Google Search Engine



Market Share of Search Engines

https://gs.statcounter.com/search-engine-market-share

Choice of BERT Model

STAATLICH ANERKANNTE HOCHSCHULE

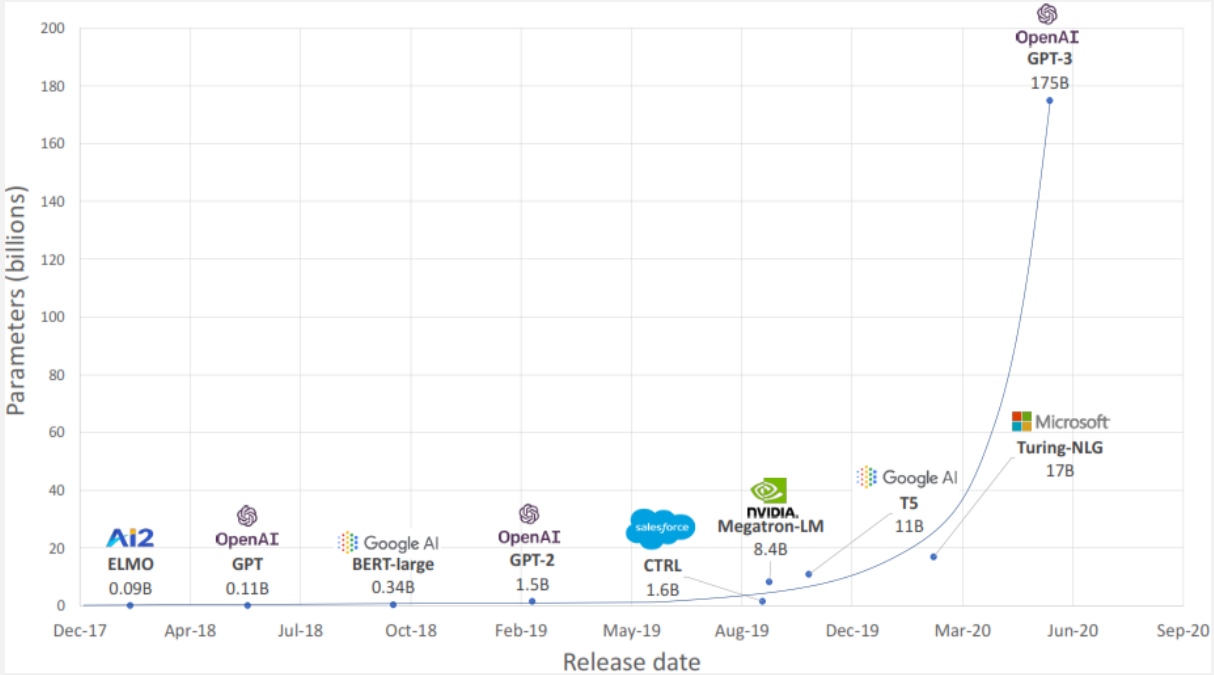
2. BERT as Knowledge Base

- Liu et al., in 2019, developed a knowledge base question-answering model by using pre-trained BERT architecture on QA datasets [2]
- In 2019, Petroni et al. established that a pre-trained BERT model already contains relational oracle knowledge comparable to traditional NLP models, thus demonstrating a huge potential of using BERT as a Knowledge base in an unsupervised QA system [3]
- In 2019 Poerner et al. supplemented this evidence by proving BERT's good performance on QA datasets is due to BERT’s inherent understanding of entity names instead of factual knowledge [4]

Choice of BERT Model

STAATLICH ANERKANNTE HOCHSCHULE

3. Optimum Model size of BERT



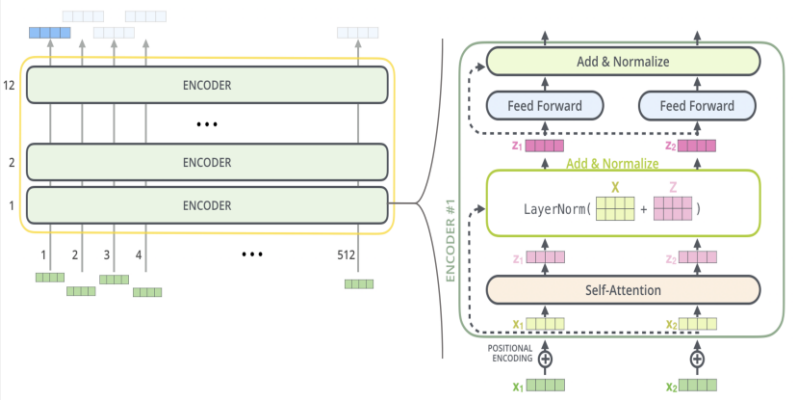
Popular Pre-trained Language Models and their parameter count

https://huggingface.co/transformers/pretrained_models.html

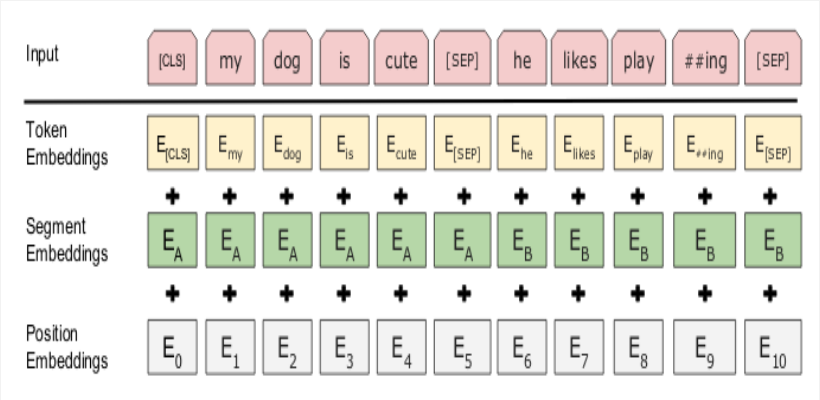
Architecture of BERT

STAATLICH ANERKANNTE HOCHSCHULE

- BERT (Devlin et al., 2018) [5] is the most advanced deep contextual language representation model introduced by Google AI researchers
- BERT is pre-trained on a large corpus of unlabeled text including the entire Wikipedia (that's 2,500 million words!) and Book Corpus (800 million words)
- BERT has been pre-trained simultaneously on two tasks, i.e., masked language modeling (MLM) and the next sentence prediction (NSP)



Architecture of BERT



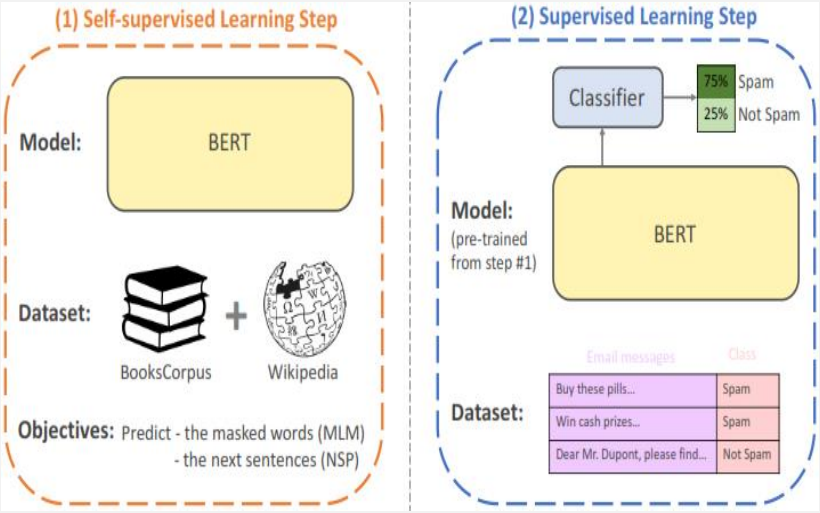
BERT Input Representation

http://jalammar.github.io/illustrated-bert/

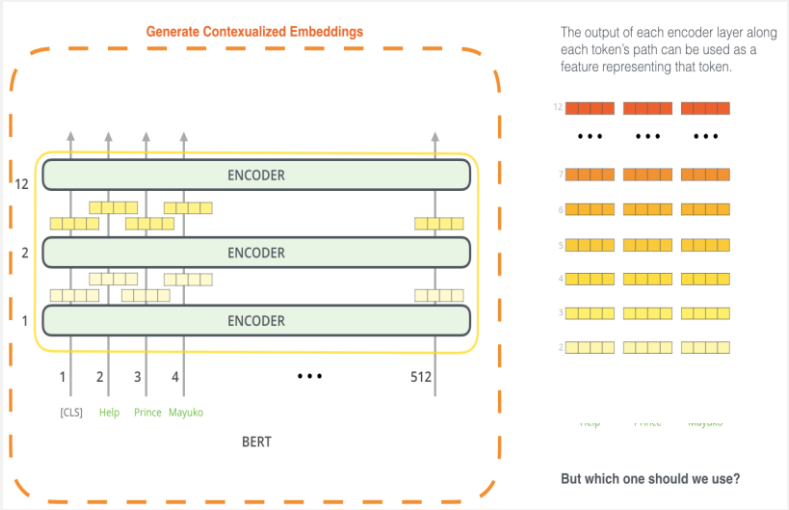
Using BERT in downstream NLP tasks

STAATLICH ANERKANNTE HOCHSCHULE

- Fine-tuning approach: Sentence classification, Next Sentence prediction



- Feature based approach: Name Entity Recognition, Similarity search



http://jalammar.github.io/illustrated-bert/

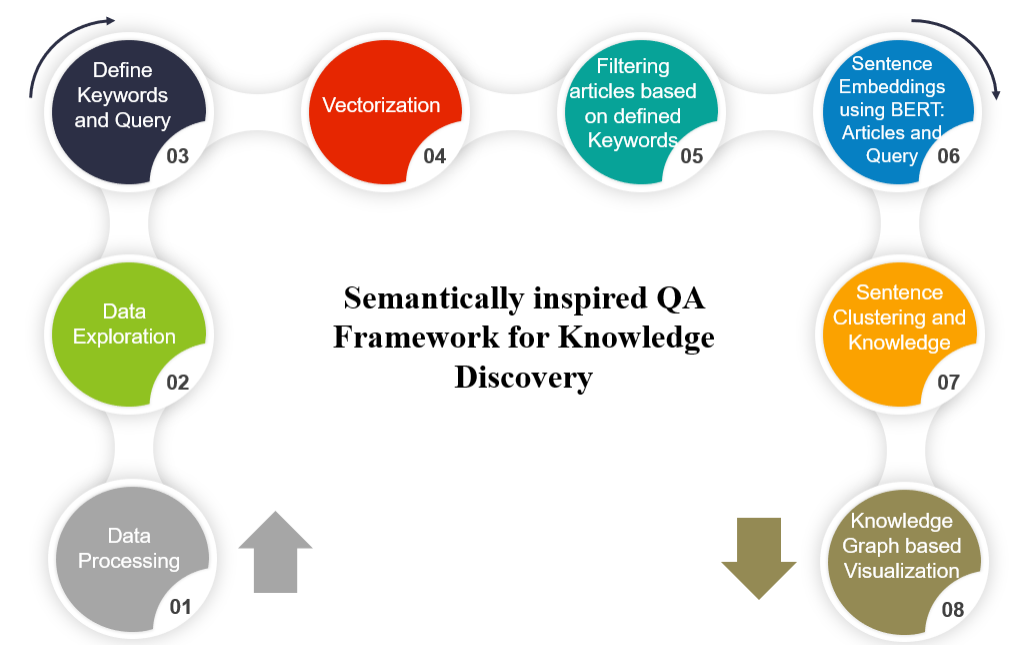
STAATLICH
ANERKANNTE
HOCHSCHULE

Can an AI Framework be developed incorporating the selected DL NLP architecture to retrieve or discover knowledge from CORD-19?

The Framework

Framework

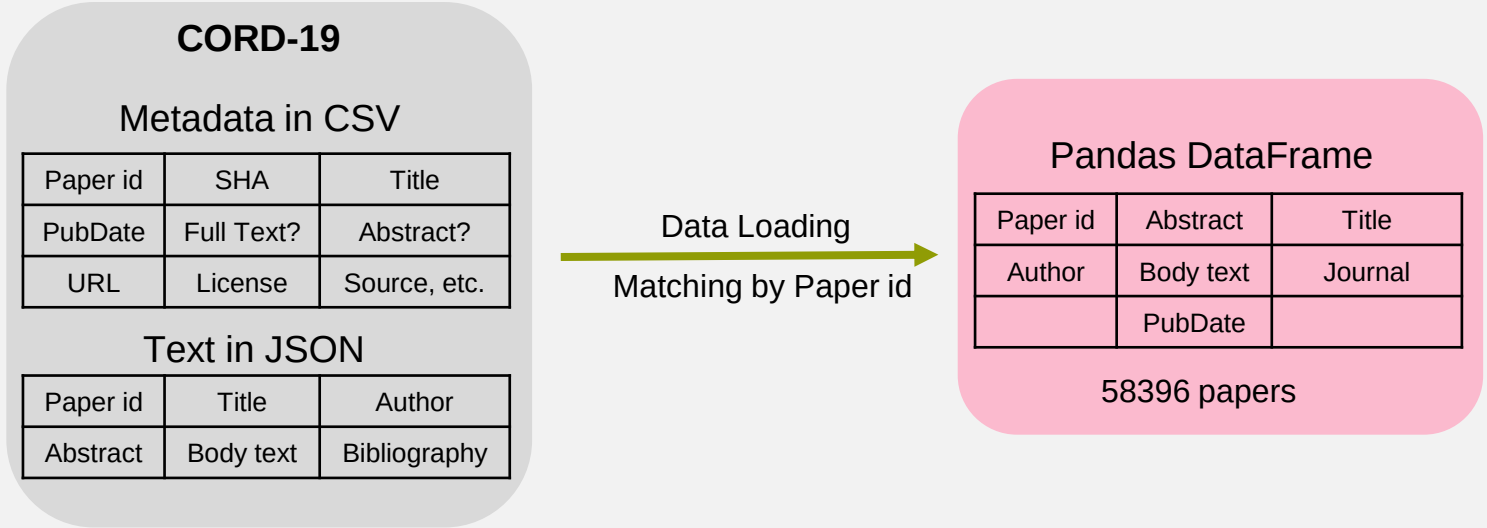
The Framework comprises of eight steps



Implementation

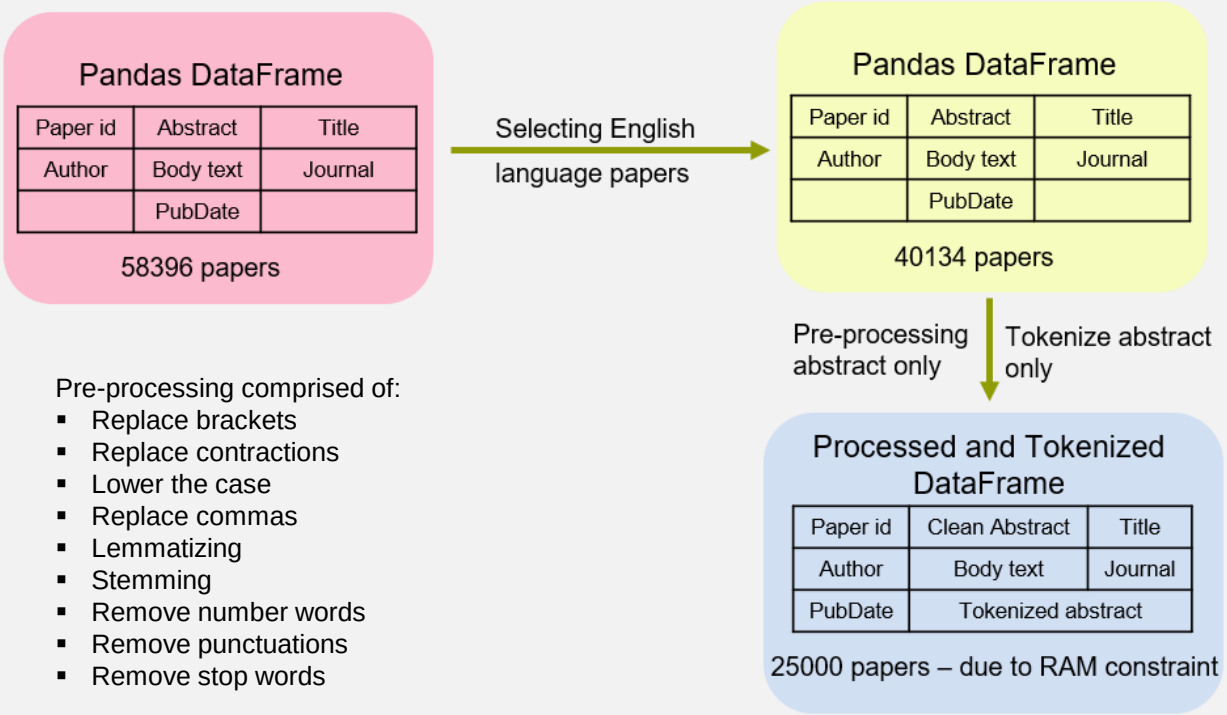
Data Loading

STAATLICH ANERKANNTE HOCHSCHULE



Data Processing

STAATLICH ANERKANNTE HOCHSCHULE



Pre-processing comprised of:

- Replace brackets
- Replace contractions
- Lower the case
- Replace commas
- Lemmatizing
- Stemming
- Remove number words
- Remove punctuations
- Remove stop words

Data Exploration

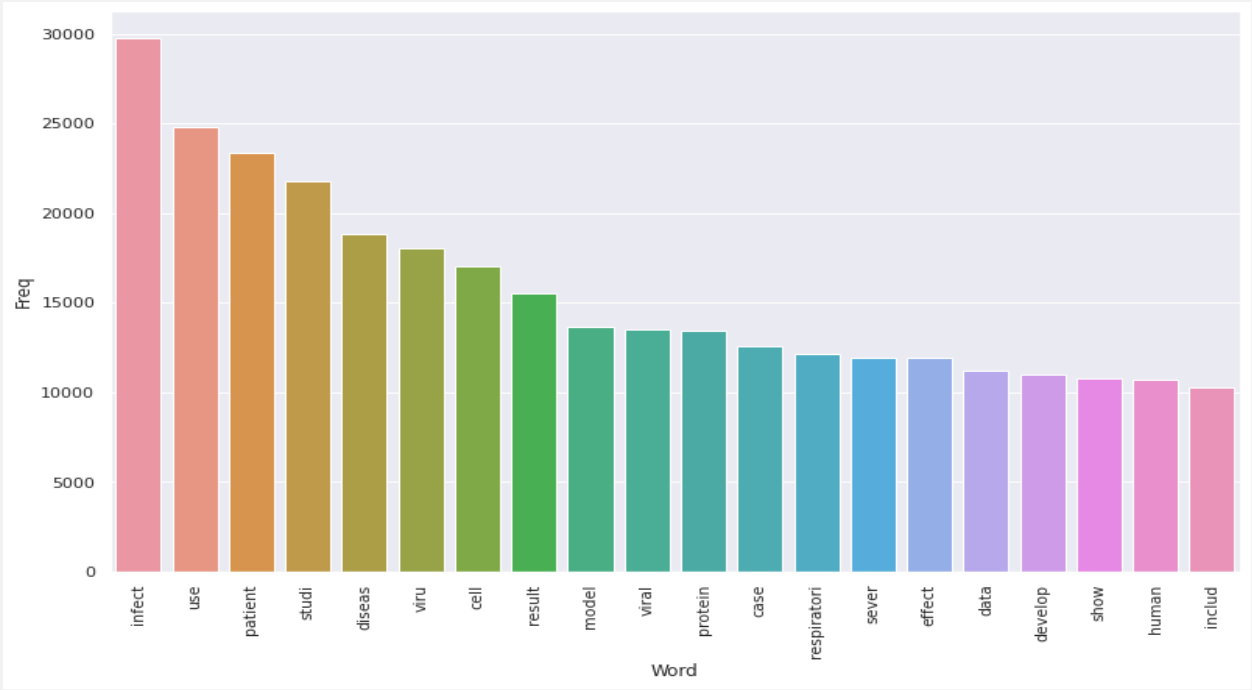
STAATLICH
ANERKANNTE
HOCHSCHULE

Word Cloud of Abstract

Data Exploration

STAATLICH
ANERKANNTE
HOCHSCHULE

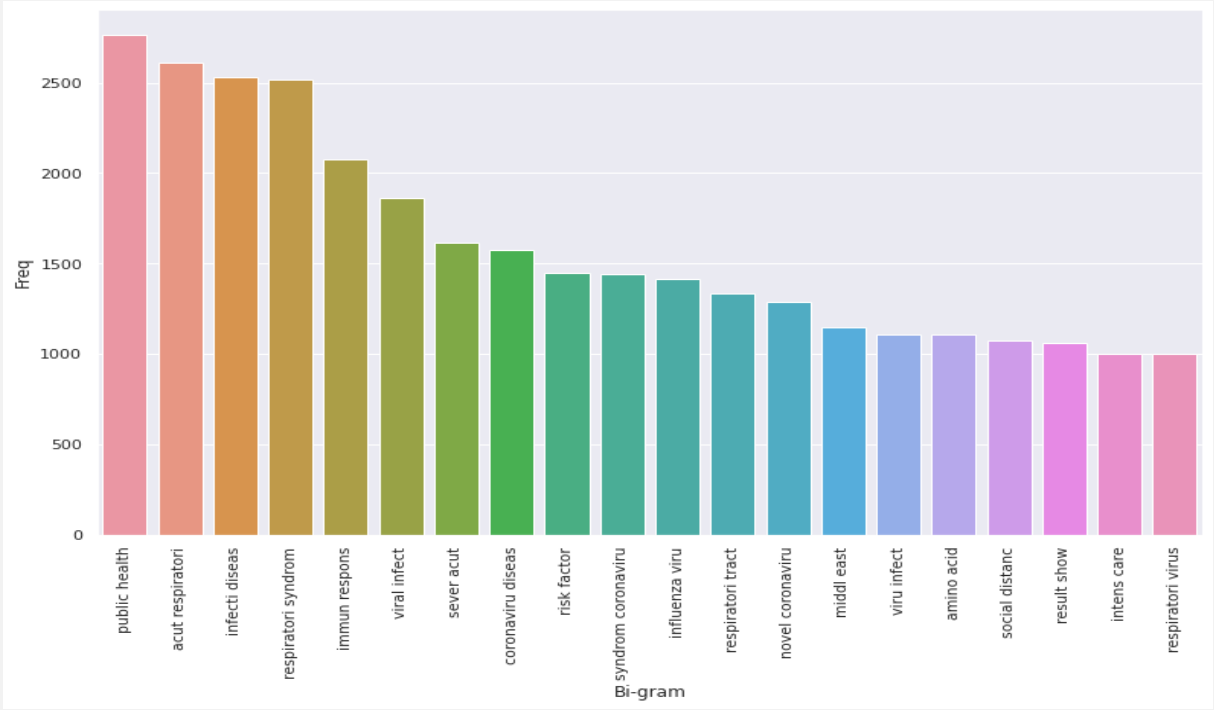
Top 20 unigrams



Data Exploration

STAATLICH
ANERKANNTE
HOCHSCHULE

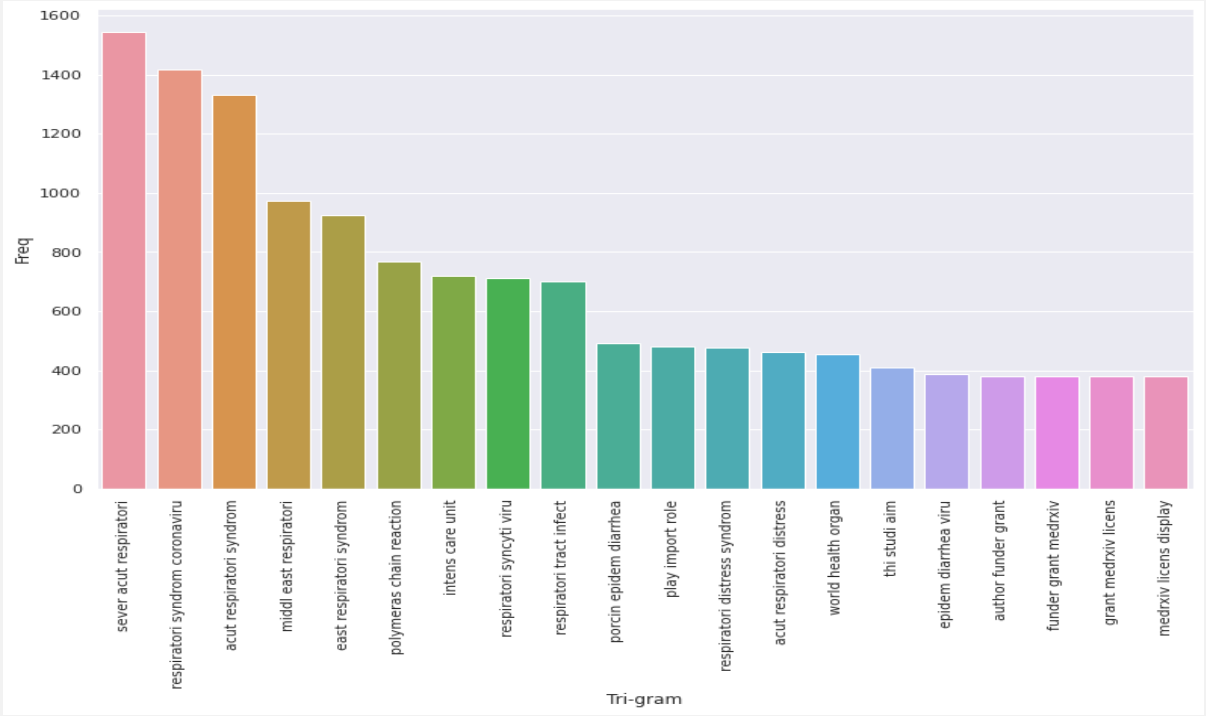
Top 20 bigrams



Data Exploration

STAATLICH
ANERKANNTE
HOCHSCHULE

Top 20 trigrams



Define Keywords and Query

STAATLICH ANERKANNTE HOCHSCHULE

Keywords: “SARS-CoV-2, Covid-19, HCoV-19, Covid corona, 2019-nCoV, sars, cov2, ncov wuhan, coronavirus, pneumonia”

Query:

Question	Query
Question1: What is known about the origin of virus and management measures at the human-animal interface?	Evidence of SARS-CoV-2 infection in animals and its transmission or spill-over to other hosts, including humans
Question 2: What has been published about medical care?	Knowledge of the frequency, manifestations, and course of extrapulmonary manifestations of SARS-CoV-2, including, but not limited to, possible cardiomyopathy and cardiac arrest
Question 3: What is known about the transmission, incubation, and environmental stability of the coronavirus?	Incubation periods for the corona disease in humans, how this varies across age and health status, how long individuals are contagious, even after recovery
Question 4: What is known about COVID-19 risk factors?	Data on potential risk factors: Transmission dynamics of the virus, including the basic reproductive number, incubation period, serial interval, modes of transmission and environmental factors
Question 5: What has been published regarding research and development and evaluation efforts for developing vaccines and therapeutics?	Capabilities to discover a therapeutic for the disease, and clinical effectiveness studies to discover therapeutics, including antiviral agents

Vectorization and Filtering papers based on Keywords

STAATLICH ANERKANNTE HOCHSCHULE

Processed and Tokenized DataFrame

Paper id	Clean Abstract	Title
Author	Body text	Journal
PubDate	Tokenized abstract	

25000 papers

Vectorize Keywords

Vectorize abstract,
TF-IDF search engine (Keywords and abstract, cosine similarity score)

Filtered Papers

7062 papers,
Cosine similarity score > 0

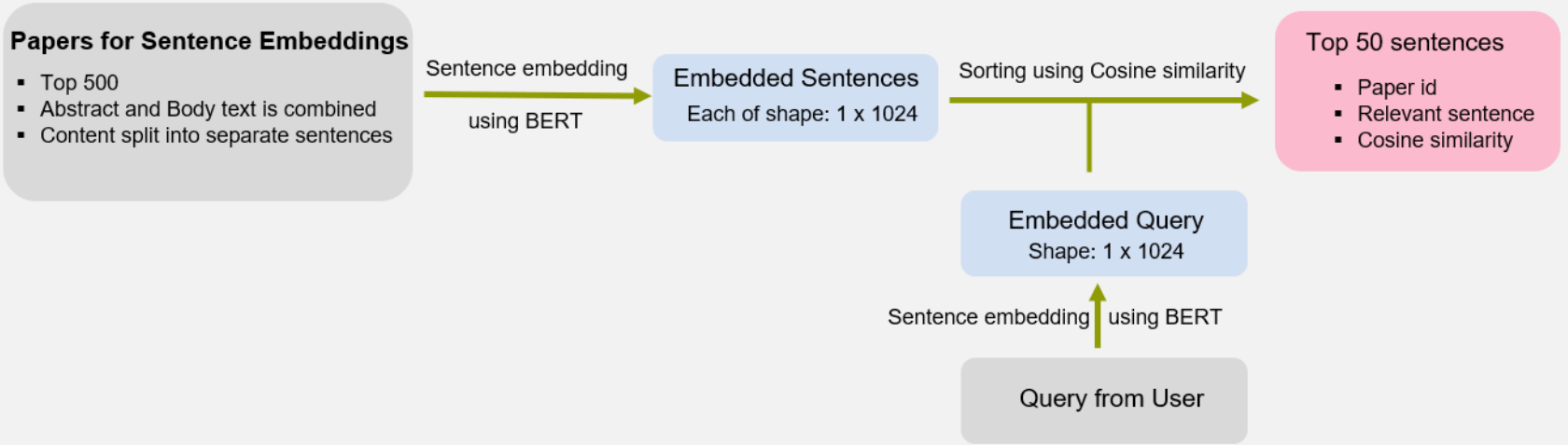
Sort and select

Papers for Sentence Embeddings

Top 500 papers – due to RAM constraint

Sentence Embeddings using BERT on filtered Articles and Query

STAATLICH ANERKANNTE HOCHSCHULE



Sentence Clustering and Knowledge

Papers for Sentence Embeddings

- Top 500 papers
- Abstract and Body text is combined
- Content split into separate sentences

Sentence embedding

using BERT

Embedded Sentences

K-mean
Clustering and
Knowledge

5-clusters of
sentences creating a
new form of
Knowledge

Knowledge Graph Visualization

STAATLICH ANERKANNTE HOCHSCHULE

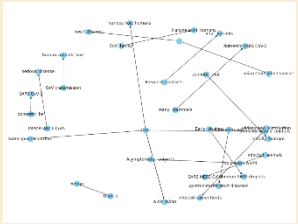
Top 50 sentences

- Paper id
- Relevant sentence
- Cosine similarity

Select top 20 sentences

Extract subject, object and relationship

Knowledge Graph



Result and Discussion

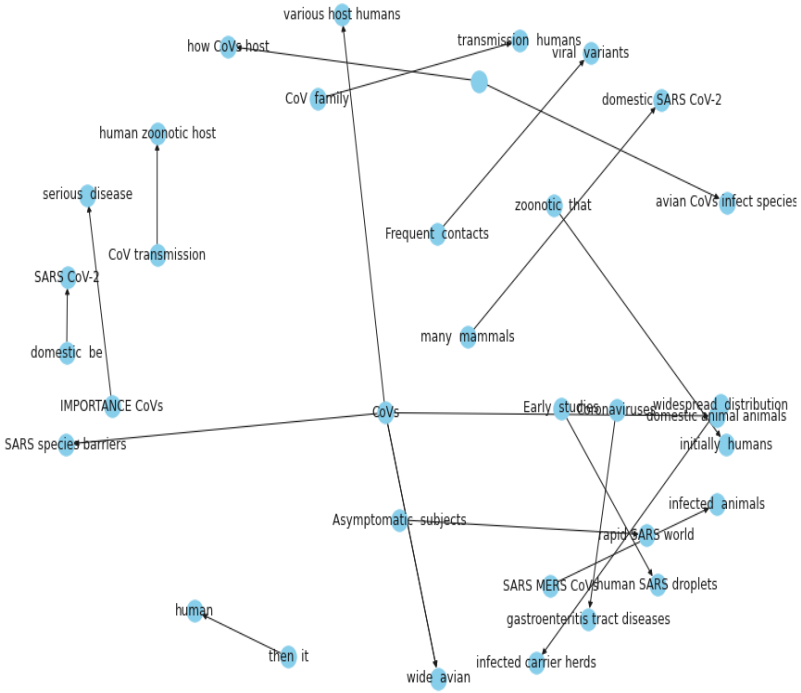
User Query 1

STAATLICH ANERKANNTE HOCHSCHULE

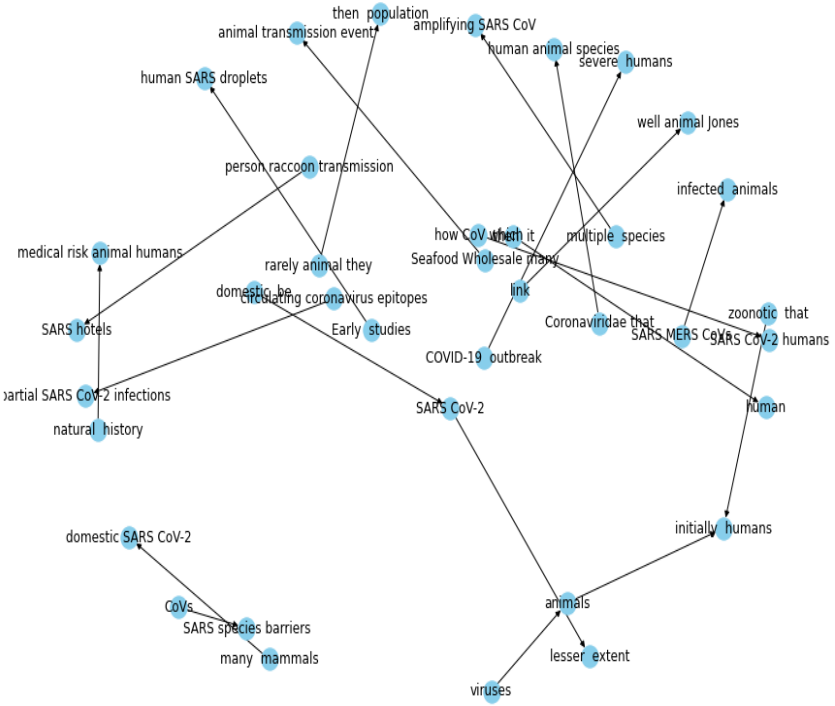
Query: Evidence of SARS-CoV-2 infection in animals and its transmission or spill-over to other hosts, including humans	
The top 3 answers from the combined abstract and body corpus are as follows:	
BERT	**RoBERTa**
Paper ID: 14eaf287845975191ed38c75bc22d181806ef167 Important sentence: Thus, SARS-CoV-2 disease can be considered a zoonotic disease (like SARS) that has initially spread from animals to humans. Cosine Similarity metric: 0.781	Paper ID: 32b941d4406406a70dbced2b906b6795b775f5d4 Important sentence: In the last 15 years, two outbreaks of previously unknown highly pathogenic coronaviruses, SARS-CoV and MERS-CoV, have demonstrated that CoVs will continue to spill over into human populations, likely facilitated by interaction between infected animals and humans. Cosine Similarity metric: 0.754
Paper ID: 32b941d4406406a70dbced2b906b6795b775f5d4 Important sentence: In the last 15 years, two outbreaks of previously unknown highly pathogenic coronaviruses, SARS-CoV and MERS-CoV, have demonstrated that CoVs will continue to spill over into human populations, likely facilitated by interaction between infected animals and humans. Cosine Similarity metric: 0.752	Paper ID: bd1973693386dff5704ca4460874257cb1f037dd Important sentence: Finally, since many mammals including domestic animals might be susceptible to SARS-CoV-2, both surveillance and experimental infection should be conducted. Cosine Similarity metric: 0.740
Paper ID: bd1973693386dff5704ca4460874257cb1f037dd Important sentence: Finally, since many mammals, including domestic animals might be susceptible to SARS-CoV-2, both surveillance and experimental infection should be conducted. Cosine Similarity metric: 0.747	Paper ID: 4c222bd3e1c78c48931f4e69164fc6ee7bffb5ff Important sentence: Currently, it is thought that SARS-CoV-2 has been introduced to human by an unidentified intermediary animal and then it has spread from human-to-human. Cosine Similarity metric: 0.736

User Query 1

STAATLICH ANERKANNTE HOCHSCHULE



Knowledge Graph using BERT



Knowledge Graph using RoBERTa

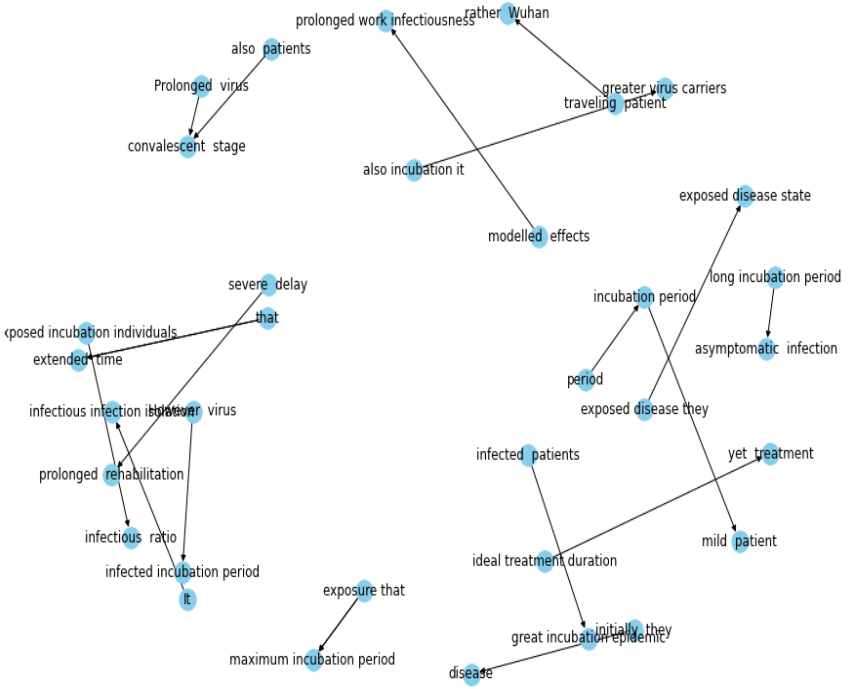
User Query 2

STAATLICH ANERKANNTE HOCHSCHULE

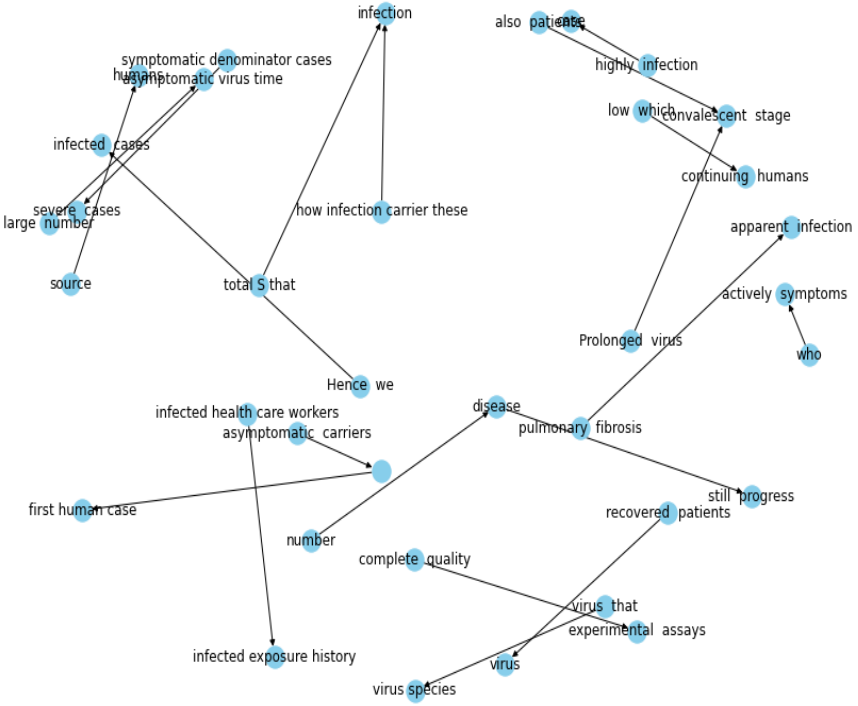
Query: Incubation periods for the corona disease in humans, how this varies across age and health status, how long individuals are contagious, even after recovery	
The top 3 answers from the combined abstract and body corpus are as follows:	
BERT	**RoBERTa**
Paper ID: 29d15895b6f3c6472f3bab820d6a99feefae60df Important sentence: Based on the current data, we do not know whether these patients are only asymptomatic initially after contracting the disease or if they are asymptomatic throughout the course of the disease. Cosine Similarity metric: 0.735	Paper ID: 86314a273c2fbd3015e2646773a98d8eb789a171 Important sentence: Deaths, sick persons who recover, and asymptomatic carriers continue to be found. Cosine Similarity metric: 0.734
Paper ID: 4bc41aaa2b754b3ec526eb40c384a18ab38b8d51 Important sentence: A long incubation period may lead to a high rate of asymptomatic and subclinical infection. Cosine Similarity metric: 0.721	Paper ID: b94ed97de06abc45de3ecc4e688c9ed357920162 Important sentence: The source and persistence of the infection in humans remains unknown. Cosine Similarity metric: 0.690
Paper ID: df97f804b68dcf16ffcc3c2c0894528d6d0c7ff2Important sentence: Susceptible individuals might acquire the infection at a given rate when they come in contact with an infectious person and enter the exposed disease state before they become infectious and later either recover or die. Cosine Similarity metric: 0.697	Paper ID: b94ed97de06abc45de3ecc4e688c9ed357920162 Important sentence: , showing the number of total cases and a timeline of persistence in human populations since onset of the first case. Cosine Similarity metric: 0.675

User Query 2

STAATLICH ANERKANNTE HOCHSCHULE



Knowledge Graph using BERT



Knowledge Graph using RoBERTa

User Query 3

STAATLICH ANERKANNTE HOCHSCHULE

Query: Knowledge of the frequency, manifestations, and course of extrapulmonary manifestations of SARS-CoV-2, including, but not limited to, possible cardiomyopathy and cardiac arrest	
The top 3 answers from the combined abstract and body corpus are as follows:	
BERT	**RoBERTa**
Paper ID: 867ebb2be2d466bcb35e5a7319708eaff39c4799 Important sentence: The complications of sedative procedures, of deep sedation and of endotracheal intubation is bradycardia and cardiac arrest, 2 case, because of hypoxia and vagal stimulation. Cosine Similarity metric: 0.715	Paper ID: 53595a15c4befbdcef60543bccbc76ad9567b640 Important sentence: In addition, peripheral pulmonary opacities were noted which were confirmed to be secondary to SARS-2-CoV pneumonia. Cosine Similarity metric: 0.751
Paper ID: 85a2db1bb25d3c369b7422b2b5f5a007175bec15 Important sentence: Cardiac complications such as electrocardiography abnormities, diastolic dysfunction, and acute myocardial injury were reported in patients with COVID-19 [124] [125] [126] [127]. Cosine Similarity metric: 0.682	Paper ID: 85a2db1bb25d3c369b7422b2b5f5a007175bec15 Important sentence: Current case reports show that SARS-CoV-2 infection may have cardiovascular symptoms in addition to the typical respiratory symptoms. Cosine Similarity metric: 0.751
Paper ID: 85a2db1bb25d3c369b7422b2b5f5a007175bec15 Important sentence: The activated SNS alters cardiac wall contractility and increases apoptotic pathways in cardiomyocytes, contributing to CVD development. Cosine Similarity metric: 0.659	Paper ID: 734779fad93249a2f6fede6afd10eeff7b37919b Important sentence: The transmission dynamics of the SARS-CoV-2 epidemic will depend on factors including the degree of seasonal variation in transmission strength, the duration of immunity, and the degree of cross-immunity between SARS-CoV-2 and other coronaviruses. Cosine Similarity metric: 0.727

Conclusion

Conclusion

STAATLICH
ANERKANNTE
HOCHSCHULE

- With the help of extensive literature review and surveys, it is established that pre-trained language models esp. BERT and it variants have been able to capture semantic and syntactic meanings of words: local context. They form the backbone of most of the advanced NLP tasks
- State of the art advanced NLP and machine learning techniques were deployed to develop a semantically intelligent QA framework to study queries related to COVID
- Framework's utilization and effectiveness have been demonstrated by way of utilizing high-level keywords and queries
- The Framework has been useful in providing the answers (in the form of the most relevant papers, content and knowledge graph)
- Querying and answering is an iterative process with refinement in query after every knowledge retrieval cycle

Future Scope

Future Scope

STAATLICH
ANERKANNTE
HOCHSCHULE

Three area have been identified for value addition in future:

1. Experimentation with other pre-trained language models esp. recently developed BioBERT

2. Extraction of more complex relationships for preparing the Knowledge Graph

3. Evaluation methodology for QA tasks where no training data is available

References

STAATLICH ANERKANNTE HOCHSCHULE

[1]. L. Wang, K. Lo, Y. Chandrasekhar and R. Reas, "CORD-19: The COVID-19 Open Research Dataset," *arXiv,* vol. 2004.10706v4, 2020. – accessed on 1st June 2020

[2]. A. Liu, Z. Huang, X. Wang and C. Yuan, "Bb-kbqa: Bert-based knowledge base question answering," *China National Conference on Chinese Computational Linguistics,* pp. 81-92, 2019.

[3]. F. Petroni, T. Rocktaschel, P. Lewis, Y. Wu and A. Miller, “Language Model as Knowledge Bases," *arXiv,* vol. 1909.01066v2, 2019.

[4]. N. Poerner, U. Waltinger and H. Schutze, "Factual knowledge vs. name-based reasoning in unsupervised qa," *arXiv,* vol. 1911.03681, 2019

[5]. J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv,* vol. abs/1810.04805, 2018.

STAATLICH
ANERKANNTE
HOCHSCHULE

Thank You!