Ονοματεπώνυμο: Παντελεήμων Μαλέκας
A.M: 1115201600268

1. For the first question, three question answering models were developed, using the pre-trained models and cross-encoders from UKPLab (https://github.com/UKPLab/sentence-transformers)

Important note: This assignment makes use of the CUDA functionality. Make sure to activate the GPU before executing any cells of the notebook.

For the pre-processing stage, each file is taken from the CORD-19 data and after removing the unnecessary sections, each article is written out to one file (final.json).

Then, the passages are created using each section from the final.json file.

Finally after initializing the transformers/cross-encoders, the questions and defining a search function (using util.search from UKPLab), we move on to the predictions.

The first model uses the ‘stsb-roberta-large’ transformer and ‘cross-encoder/stsb-TinyBERT-L-4’ cross-encoder.

The second model uses the ‘msmarco-distilroberta-base-v2’ transformer and ‘cross-encoder/ms-marco-TinyBERT-L-4’ cross-encoder.

The third model uses the ‘msmarco-roberta-base-v2’ transformer and ‘cross-encoder/ms-marco-electra-base’ cross-encoder.

Here are the questions used:

1. 'What are the coronoviruses?'
2. 'What was discovered in Wuhan in December 2019?'
3. 'What is Coronovirus Disease 2019?'
4. 'What is COVID-19?'
5. 'What is caused by SARS-COV2?'
6. 'How is COVID-19 spread?'
7. 'Where was COVID-19 discovered?'
8. 'How does coronavirus spread?'
9. 'How can the spread of COVID-19 be prevented?'
10. 'How many coronaviruses are there?'
11. 'Is COVID-19 related to SARS?'
12. 'What are the symptoms of COVID-19?'
13. 'How can COVID-19 be cured?'
14. 'How many cases of COVID-19?'
15. 'Is COVID-19 a pandemic?'

16. 'What is the financial impact of COVID-19?',
17. ‘'Where does the name coronavirus come from?'

The questions 9-17 are the ones I came up with.

Here are the results of each model:

Model 1:
Question: What are the coronoviruses?
Found in Title: Structure and Inhibition of the SARS Coronavirus Envelope Protein Ion Channel Score: 0.27000197768211365

Question: What was discovered in Wuhan in December 2019?
Found in Title: Note from the editors: Don't stop thinking about tomorrow Eurosurveillance editorial team Score: 0.24593238532543182

Question: What is Coronovirus Disease 2019?
Found in Title: pathogens Emergence of Novel Coronavirus 2019-nCoV: Need for Rapid Vaccine and Biologics Development Score: 0.4932737946510315

Question: What is COVID-19?
Found in Title: Systematic Comparison of Two Animal-to-Human Transmitted Human Coronaviruses: SARS-CoV-2 and SARS-CoV Score: 0.40332743525505066

Question: What is caused by SARS-COV2?
Found in Title: Systematic Comparison of Two Animal-to-Human Transmitted Human Coronaviruses: SARS-CoV-2 and SARS-CoV Score: 0.4778613746166229

Question: How is COVID-19 spread?
Found in Title: Systematic Comparison of Two Animal-to-Human Transmitted Human Coronaviruses: SARS-CoV-2 and SARS-CoV Score: 0.3468919098377228

Question: Where was COVID-19 discovered?
Found in Title: Systematic Comparison of Two Animal-to-Human Transmitted Human Coronaviruses: SARS-CoV-2 and SARS-CoV Score: 0.26053303480148315

Question: How does coronavirus spread?
Found in Title: BMC Infectious Diseases Mutational dynamics of the SARS coronavirus in cell culture and human populations isolated in 2003 Score: 0.3180394768714905

Question: How can the spread of COVID-19 be prevented?
Found in Title: Identification of COVID-19 Can be Quicker through Artificial Intelligence framework using a Mobile Phone-Based Survey in the Populations when Cities/Towns Are Under Quarantine Score: 0.4440874755382538

Question: How many coronaviruses are there?
Found in Title: Revisiting the dangers of the coronavirus in the ophthalmology practice Score: 0.5892760157585144

Question: Is COVID-19 related to SARS?
Found in Title: Systematic Comparison of Two Animal-to-Human Transmitted Human Coronaviruses: SARS-CoV-2 and SARS-CoV Score: 0.487301230430603

Question: What are the symptoms of COVID-19?
Found in Title: Systematic Comparison of Two Animal-to-Human Transmitted Human Coronaviruses: SARS-CoV-2 and SARS-CoV Score: 0.4963934123516083

Question: How can COVID-19 be cured?
Found in Title: Systematic Comparison of Two Animal-to-Human Transmitted Human Coronaviruses: SARS-CoV-2 and SARS-CoV Score: 0.4134822487831116

Question: How many cases of COVID-19?
Found in Title: Systematic Comparison of Two Animal-to-Human Transmitted Human Coronaviruses: SARS-CoV-2 and SARS-CoV Score: 0.3717901110649109

Question: Is COVID-19 a pandemic?
Found in Title: Systematic Comparison of Two Animal-to-Human Transmitted Human Coronaviruses: SARS-CoV-2 and SARS-CoV Score: 0.4221062958240509

Question: What is the financial impact of COVID-19?
Found in Title: Clinical Medicine Real-Time Estimation of the Risk of Death from Novel Coronavirus (COVID-19) Infection: Inference Using Exported Cases Score: 0.2919902503490448

Question: Where does the name coronavirus come from?
Found in Title: Mucosal Immune Response to Feline Enteric Coronavirus Infection Score: 0.36045417189598083

Model 2:
Question: What are the coronoviruses?
Found in Title: Detection and Characterization of Distinct Alphacoronaviruses in Five Different Bat Species in Denmark Score: 0.06262648850679398

Question: What was discovered in Wuhan in December 2019?
Found in Title: Epidemiological Identification of A Novel Pathogen in Real Time: Analysis of the Atypical Pneumonia Outbreak in Wuhan Score: 0.9783265590667725

Question: What is Coronovirus Disease 2019?
Found in Title: pathogens Emergence of Novel Coronavirus 2019-nCoV: Need for Rapid Vaccine and Biologics Development Score: 0.6576673984527588

Question: What is COVID-19?
Found in Title: Early epidemiological analysis of the coronavirus disease 2019 outbreak based on crowdsourced data: a population- level observational study Score: 0.781908392906189

Question: What is caused by SARS-COV2?
Found in Title: Reverse Genetics of SARS-Related Coronavirus Using Vaccinia Virus-Based Recombination Score: 0.9848011136054993

Question: How is COVID-19 spread?
Found in Title: Cell Discovery Phase-adjusted estimation of the number of Coronavirus Disease 2019 cases in Wuhan, China Score: 0.9270247220993042

Question: Where was COVID-19 discovered?
Found in Paper ID: 37e17a2bab698f8850dc89f7689eda93502821fb Score: 0.9182679057121277

Question: How does coronavirus spread?
Found in Title: The novel coronavirus outbreak in Wuhan, China Score: 0.6929607391357422

Question: How can the spread of COVID-19 be prevented?
Found in Title: Identification of COVID-19 Can be Quicker through Artificial Intelligence framework using a Mobile Phone-Based Survey in the Populations when Cities/Towns Are Under Quarantine Score: 0.9718709588050842

Question: How many coronaviruses are there?
Found in Title: Genome-wide analysis of codon usage bias in Bovine Coronavirus Score: 0.8849629163742065

Question: Is COVID-19 related to SARS?
Found in Title: A new coronavirus associated with human respiratory disease in China Score: 0.9755364656448364

Question: What are the symptoms of COVID-19?
Found in Title: Q&A: The novel coronavirus outbreak causing COVID-19 Score: 0.9922263622283936

Question: How can COVID-19 be cured?
Found in Title: Outbreak of Novel Coronavirus (SARS-Cov-2): First Evidences From International Scientific Literature and Pending Questions Score: 0.9107555747032166

Question: How many cases of COVID-19?
Found in Title: Clinical Medicine Characteristics of and Public Health Responses to the Coronavirus Disease 2019 Outbreak in China Score: 0.9378999471664429

Question: Is COVID-19 a pandemic?
Found in Title: Preliminary Identification of Potential Vaccine Targets for the COVID-19 Coronavirus (SARS-CoV-2) Based on SARS-CoV Immunological Studies Score: 0.7251870632171631

Question: What is the financial impact of COVID-19?
Found in Title: First two months of the 2019 Coronavirus Disease (COVID-19) epidemic in China: real- time surveillance and evaluation with a second derivative model Score: 0.8832363486289978

Question: Where does the name coronavirus come from?
Found in Title: Emergence of the Middle East Respiratory Syndrome Coronavirus Score: 0.9285920262336731

Model 3:
Question: What are the coronoviruses?
Found in Title: Case Report Neurological Complications of Middle East Respiratory Syndrome Coronavirus: A Report of Two Cases and Review of the Literature Score: 0.003970596939325333

Question: What was discovered in Wuhan in December 2019?
Found in Paper ID: fd28e6d03eef27b0454f13ca539dc1498242a4c2 Score: 0.9935118556022644

Question: What is Coronovirus Disease 2019?
Found in Title: Consensus statement The species Severe acute respiratory syndrome- related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2 Coronaviridae Study Group of the International Committee on Taxonomy of Viruses* Score: 0.8691843748092651

Question: What is COVID-19?
Found in Paper ID: af000c5a8e181550fd16291e5d4f0f70ca9161a1 Score: 0.9742367267608643

Question: What is caused by SARS-COV2?
Found in Title: Human Coronaviruses: Insights into Environmental Resistance and Its Influence on the Development of New Antiseptic Strategies Score: 0.9557797312736511

Question: How is COVID-19 spread?
Found in Title: Identification of COVID-19 Can be Quicker through Artificial Intelligence framework using a Mobile Phone-Based Survey in the Populations when Cities/Towns Are Under Quarantine Score: 0.9765667915344238

Question: Where was COVID-19 discovered?
Found in Paper ID: 37e17a2bab698f8850dc89f7689eda93502821fb Score: 0.9933891296386719

Question: How does coronavirus spread?
Found in Title: Use of Toll-Like Receptor 3 Agonists Against Respiratory Viral Infections Score: 0.8378279805183411

Question: How can the spread of COVID-19 be prevented?
Found in Paper ID: 37e17a2bab698f8850dc89f7689eda93502821fb Score: 0.990584135055542

Question: How many coronaviruses are there?
Found in Title: Early events during human coronavirus OC43 entry to the cell OPEN Score: 0.9926537275314331

Question: Is COVID-19 related to SARS?
Found in Title: Q&A: The novel coronavirus outbreak causing COVID-19 Score: 0.9924227595329285

Question: What are the symptoms of COVID-19?
Found in Title: Q&A: The novel coronavirus outbreak causing COVID-19 Score: 0.9935359954833984

Question: How can COVID-19 be cured?
Found in Title: Outbreak of Novel Coronavirus (SARS-Cov-2): First Evidences From International Scientific Literature and Pending Questions Score: 0.9821274876594543

Question: How many cases of COVID-19?
Found in Paper ID: 37e17a2bab698f8850dc89f7689eda93502821fb Score: 0.9923303723335266

Question: Is COVID-19 a pandemic?
Found in Title: Comment Score: 0.9768864512443542

```
Question: What is the financial impact of COVID-19?
Found in Title: Clinical Medicine Communicating the Risk of Death from Novel Coronavirus
Disease (COVID-19) Score: 0.13666510581970215

Question: Where does the name coronavirus come from?
Found in Title: Clinical Medicine Optimization Method for Forecasting Confirmed Cases of
COVID-19 in China Score: 0.9542181491851807
```

As we can see, the first model gives the worst results. This is expected since the pre-trained models are not trained for question answering tasks. I thought I should utilize them anyway, to see how well they will perform.

Now the second and third models (which are trained for question answering tasks) give far better scores. The third model gives somewhat better cross encoder scores than the second one (13/17 questions over 90% compared to 10/17). However, this doesn't mean it's necessarily better. We can see in the second question that a lot of >90% answers are not that accurate. On the other hand, some <90% answers are actually quite accurate.

Nevertheless, for this assignment I chose to use this metric in order to make a comparison. Time isn't really important for these models as all the comparisons happen almost instantly. I should note though, the time to make the embeddings for each model is different (Specifically: 65.75 minutes for the first, 24.95 minutes for the second and 37.61 minutes for the third). Although the third model takes up a bit more time, I deemed the accuracy metric more important so the model used in the second question is the third model.

2. For the second question, a slightly different search method is defined, in order to print the relevant passage of the article that the model found. Also, two additional passages are given as alternate answers. These passages have a smaller cross encoder score than the first passage.

The results of the model can be seen in the notebook file.

3. I wasn't able to finish the third question. The .ipynb file simply contains some ideas of what I was going to do, if I had more time. :-)