MLx7: Week 2
Build a search engine

Team:
Perceptron Party

Main Actors:
Dimitris, Dimitar, Yuri, Pyry

What have we achieved?

- retrained CBOW word2vec with wiki (text8) + MS Marco (v1.1) corpus
- created triplets query / positive / negative with random sampling
- built GloVe-based vocabulary and embedding matrix
- indexed passage embeddings into ChromaDB for fast retrieval
- mined hard negatives using ChromaDB
- train Two-Tower model again with hard negatives
- re-encode passages using the improved model
- tested model’s performance with different datasets

Query: where is the eiffel tower?

Top Results:
1. [Similarity: 0.9558] The BuzzFeed app is best way to stay up to date with the latest viral videos, images, links, and buzz while on the go....

2. [Similarity: 0.9417] Gorillas in the Mist isn't a terrible film, but it is a frustrating one, and you can't help but feel betrayed by how the fil
y, and whenever t...

3. [Similarity: 0.9260] Yes. The centrosome, also called the microtubule organizing center, is an area in the cell where microtubles are produced. W
ganelles, the cen...

4. [Similarity: 0.9196] Phagocytosis (literally, cell-eating) is the process by which cells ingest large objects, such as cells which have undergone

What have we learned?

- start simple and build in complexity
- input data significance
- good practice to examine model outputs from each layer
- .parquet is better
- ChromaDB files can be very large
- cannot create Dockerfile from the Computa GPU’s

What we could have done better?

- add an RNN
- optimise model's architecture (number of layers, activation functions, dropout, margin, etc.)
- figure out and implement a better Triplet Loss
- start with the presentation earlier