AREA SCIENCE PARK

PROFILO A

Traccia n. 3

Il candidato discuta in maniera sintetica due dei seguenti argomenti, indicando esplicitamente la scelta all'inizio del proprio elaborato:

A) Il candidato descriva il problema del clustering, delineandone gli obiettivi generali. Successivamente, il candidato fornisca una descrizione generale dell'algoritmo di clustering K-Means e ne discuta le limitazioni, descrivendo, a parole, un caso specifico in cui l'algoritmo non è adatto. Infine, il candidato illustri un metodo per determinare il numero ottimale di cluster (K) dai dati.

B) Il candidato descriva brevemente cosa sia un Transformer, evidenziandone gli elementi architetturali caratteristici. Successivamente, il candidato risponda alla seguente domanda:

- Qual è la differenza tra il tipo di pre-training utilizzato nelle architetture "BERT-like" e quello utilizzato nelle architetture "GPT-like"?

C) Si consideri un laboratorio dotato di strumenti scientifici in grado di generare grandi volumi di dati. I dati grezzi devono essere trasferiti e salvati automaticamente in un centro di calcolo. Una volta salvati con successo, si deve automaticamente avviare una pipeline di analisi dei dati. Inoltre, i dati e gli strumenti computazionali correlati devono essere resi accessibili ai ricercatori tramite una piattaforma cloud che supporti l'uso di strumenti open-source per l'analisi, il machine learning e il deep learning. L'obiettivo è garantire che la ricerca sia il più possibile riproducibile. Il candidato delinei gli elementi essenziali di un'infrastruttura di ricerca di questo tipo. Inoltre, il candidato affronti brevemente almeno due dei seguenti argomenti:

- Gestione dei dati secondo i principi FAIR
- Interoperabilità
- "Infrastructure as Code" per lo sviluppo e la manutenzione della pipeline

The candidate should discuss, in a concise and effective manner, two of the following topics, explicitly indicating their choice at the beginning of their work:

A) The Candidate should describe the clustering problem, outlining its general objectives. Next, the Candidate should provide a broad description of the K-Means clustering algorithm and discuss its limitations, describing, in words, a specific case where the algorithm is unsuitable. Finally, the Candidate should illustrate a method for determining the optimal number of clusters (K) from the data.

B) The Candidate should briefly describe what a Transformer is, outlining its characteristic architectural elements.
The Candidate should then answer the following question:

- What is the difference between the type of pre-training used in BERT-like architectures and in GPT-like architectures?

C) Consider a laboratory equipped with high-throughput scientific instruments capable of generating large volumes of data. The raw data must be automatically transferred and stored in a computing center. Once the data has been successfully saved, a data analysis pipeline should automatically start. Additionally, the data and related computational tools must be made accessible to researchers via a cloud platform that supports the use of open-source tools for analysis, machine learning, and deep learning. The goal is to ensure that the research is as reproducible as possible. The Candidate should outline the essential elements of such a research infrastructure. Additionally, the Candidate should briefly address at least two of the following topics:

- FAIR Data Management
- Interoperability
- "Infrastructure as Code" for the development and maintenance of the pipeline