Sessione ordinaria 2025
Seconda prova scritta

Ministero dell'istruzione e del merito

A038 - ESAME DI STATO CONCLUSIVO DEL SECONDO CICLO DI ISTRUZIONE

Indirizzo ITIA - INFORMATICA E TELECOMUNICAZIONI ARTICOLAZIONE "INFORMATICA"

(Testo valevole anche per gli indirizzi quadriennali IT32 e ITIT)

Disciplina: INFORMATICA

Il candidato svolga la prima parte della prova e due tra i quesiti proposti nella seconda parte.

PRIMA PARTE

Premessa

Un *dataset* (letteralmente "insieme di dati" in italiano) è una collezione strutturata di una grande quantità di dati, che, ai fini di questa prova, consideriamo organizzata in forma relazionale, all'interno della quale sono descritti elementi di interesse del mondo reale (es. eventi, notizie, oggetti, ecc.) con una serie di *caratteristiche*.
Un dataset può contenere numeri, parole, immagini, suoni, o qualsiasi altro tipo di informazione.
I dataset sono fondamentali nell'addestramento di alcune applicazioni di intelligenza artificiale.
Fare il *labeling* (o etichettatura) di un dataset significa aggiungere delle etichette ad ogni elemento per indicare cosa esso rappresenta, o meglio a quale categoria o classe appartiene. Ad esempio, immaginando di avere un dataset di immagini di fiori, il labeling assegnerebbe ad ogni immagine un'etichetta, cioè una stringa contenente il nome del tipo di fiore che quella immagine mostra. L'operazione di labeling è normalmente svolta da chi sa classificare gli elementi presenti nel dataset, in questo esempio un botanico. Tale operazione è necessaria quando si vogliono utilizzare tecniche di intelligenza artificiale e *machine learning* basate su algoritmi di apprendimento supervisionato.
In tal caso viene predisposto un dataset con un numero elevato di elementi già etichettati, detto *training dataset*, che costituisce un insieme di esempi per addestrare l'algoritmo (o più correttamente per addestrare il modello di intelligenza artificiale). Quando l'algoritmo avrà "imparato" dagli esempi forniti, sarà in grado di classificare autonomamente anche nuovi elementi. Ad esempio, fornendo in input una nuova immagine di un fiore, l'algoritmo addestrato sarà in grado di restituire in output l'etichetta che con grande probabilità lo classifica, ad esempio "margherita".

Caso professionale

Al fine di contrastare il fenomeno delle *fake news*, ad una società informatica è stato commissionato lo sviluppo di una piattaforma web per effettuare il labeling di un training dataset di grandi dimensioni, per poi addestrare un modello di intelligenza artificiale a classificare le news presenti sul web.

Ogni news è caratterizzata dalla fonte da cui proviene, di cui viene indicata la tipologia (blog, social media, giornale online, piattaforma di streaming, ecc.) e il nome (New York Times, Gazzetta del mezzogiorno, Focus, Facebook, Instagram, TikTok, Spotify, YouTube, ecc.) o comunque il dominio del sito di provenienza. Ogni news è inoltre caratterizzata da un URL che la localizza sul web, una data di pubblicazione, un eventuale titolo, l'autore se disponibile, il contenuto testuale (derivante da articolo di giornale, post su un social, transcript di un video o podcast, ecc.); ad essa possono essere eventualmente associati più contenuti multimediali (audio, video e immagini) ed anche più commenti che possono accompagnare la notizia.

Ministero dell'istruzione e del merito

A038 - ESAME DI STATO CONCLUSIVO DEL SECONDO CICLO DI ISTRUZIONE

L'obiettivo è classificare ogni news per assegnare alla stessa due etichette, denominate: Topic e Result. L'etichetta *Topic* serve ad indicare a quale argomento la news si riferisce, e potrà assumere un valore tra quelli contenuti in un elenco del tipo: Economia, Politica, Medicina e Salute, Cultura, Cronaca, Scienza e Tecnologia, Sport, ecc.

L'etichetta *Result* sarà quella che classificherà effettivamente la notizia, assegnando uno tra i seguenti possibili valori: "Fake" o "Vera" o "Dubbia". Nel caso in cui la notizia sia etichettata come fake, sarà inoltre necessario stabilire una tra le seguenti possibili motivazioni[1]: "contenuto fabbricato", "contenuto manipolato", "contenuto diffuso da impostori", "falso contesto", "contenuto ingannevole", "falsa connessione", "satira o parodia", ed inserire poi una nota a sostegno della motivazione.

L'operazione di labeling sarà affidata ad un gruppo di esperti *junior* e ad un gruppo di esperti *senior* che agiranno in due fasi successive: nella prima fase il gruppo di esperti junior effettuerà una etichettatura provvisoria delle news; successivamente, il gruppo di esperti senior effettuerà la validazione finale della classificazione svolta nella prima fase, lasciando invariate o correggendo le etichettature con le relative motivazioni compilate dagli esperti junior. Durante le operazioni di etichettatura si avranno quindi news non ancora etichettate, news che hanno ricevuto una etichettatura provvisoria, e news con etichettatura validata.

Alla fine dell'etichettatura del dataset, la piattaforma dovrà anche consentire alcune analisi sui dati etichettati.

Il candidato, effettuate le opportune ipotesi aggiuntive, sviluppi:

1. un'analisi della realtà di riferimento, giungendo alla definizione dello schema concettuale della base di dati relazionale che, a suo motivato giudizio, sia idoneo a gestire la piattaforma web;
2. il relativo schema logico della base di dati relazionale;
3. la definizione in linguaggio SQL di un sottoinsieme delle relazioni della base di dati in cui siano presenti alcune delle relazioni che contengono vincoli di integrità referenziale e/o vincoli di dominio, laddove presenti;
4. le seguenti interrogazioni espresse in linguaggio SQL che permettono di ottenere:
    a. il numero di news classificate come false (fake) per ciascun valore dell'etichetta Topic pubblicate nell'ultimo anno;
    b. Il numero di news analizzate da un certo esperto senior di cui siano stati forniti il nome ed il cognome;
    c. dato un argomento (Topic) e un periodo di pubblicazione, il numero delle news classificate come fake per ogni singola motivazione ("contenuto fabbricato", "contenuto manipolato", ..., "satira o parodia");
    d. Il nominativo dell'operatore senior che ha effettuato il maggior numero di validazioni nell'ultimo mese delle news di tipo "fake";

A038 - ESAME DI STATO CONCLUSIVO DEL SECONDO CICLO DI ISTRUZIONE

[1] All'interno di ciò che definiamo disinformazione, la guida di *First Draft* identifica sette modalità differenti, secondo una categorizzazione creata da Claire Wardle nel 2017 con l'obiettivo di sostituire il termine generico di 'fake news'

5. il progetto di massima della struttura dell'applicazione web per la gestione della realtà sopra presentata, indicando quali risorse tecnologiche sia di tipo hardware che di tipo software si ritengono idonee alla realizzazione;
6. una parte dell'applicazione, utilizzando appropriati linguaggi a scelta sia lato client che lato server, che sviluppi una delle seguenti funzionalità:
    a. presentazione all'esperto junior di tutti i dati di una specifica news non ancora etichettata, per consentirne l'etichettatura provvisoria;
    b. presentazione, eventualmente in forma grafica, del numero delle news etichettate come fake per ciascun argomento (topic);
    c. modulo che consenta di scegliere un argomento ed un periodo di pubblicazione e visualizzi il numero delle news classificate come fake per ogni singola motivazione.

SECONDA PARTE

I. In relazione al tema presentato nella prima parte, si ipotizzi che la società abbia la necessità di monitorare ciascuna delle due fasi dell'etichettatura. Si discutano le possibili modifiche da apportare alle funzionalità della piattaforma per tracciare il numero di news etichettate giornalmente da parte di ciascun esperto e il tempo complessivamente impiegato da ciascun gruppo di esperti (junior e senior) per completare l'etichettatura dell'intero dataset.

II. In relazione al tema presentato nella prima parte, si ipotizzi di voler tenere traccia dei possibili cambiamenti effettuati dagli esperti senior rispetto alle etichettature inizialmente inserite dagli esperti junior. Si descrivano le modifiche da apportare alla base di dati per tale esigenza.

III. Si ipotizzi che per accedere ad un sito web, oltre alle classiche credenziali username e password, si voglia aggiungere un ulteriore livello di verifica dell'utente (autenticazione a due fattori). Si descriva una possibile struttura della tabella degli utenti che permetta di supportare tale modalità di autenticazione. Si descrivano inoltre i passaggi fondamentali che il codice di autenticazione, scritto con linguaggi lato server, dovrà compiere per consentire l'accesso al sito web.

IV. Un database relazionale prevede l'accesso condiviso e contemporaneo da parte di più utenti attraverso la rete. Si illustrino le possibili problematiche che possono incidere sulla integrità e la consistenza dei dati e attraverso quali misure e soluzioni sia possibile prevenirle.

Durata massima della prova: 6 ore.
È consentito soltanto l'uso dei manuali di riferimento dei linguaggi di programmazione (language reference) e di calcolatrici scientifiche o grafiche purché non siano dotate della capacità di elaborazione simbolica algebrica e non abbiano la disponibilità di connessione a Internet.
È consentito l'uso del dizionario della lingua italiana.
È consentito l'uso del dizionario bilingue (italiano-lingua del paese di provenienza) per i candidati di madrelingua non italiana.
Non è consentito lasciare l'Istituto prima che siano trascorse 3 ore dalla consegna della traccia.