Descrizione tecnica del servizio

Oggetto della manifestazione di interesse è la fornitura di un servizio di tipo *big data* che consiste nella messa a disposizione di un vasto database documentale di notizie (prodotte da varie fonti di informazione), che il Fornitore dovrà costruire, aggiornare e manutenere per il periodo del contratto, come infrastruttura per l'erogazione delle ulteriori componenti di servizio di seguito descritte.
Lo scopo generale del servizio è consentire all'Autorità di effettuare analisi dei dati sul sistema dell'informazione in Italia attraverso l'elaborazione di serie storiche di occorrenze di parole chiavi e attraverso l'uso *in-house* di tecniche di *NLP* (*Natural Language Processing*) applicate al testo di milioni di notizie afferenti ad un arco temporale esteso, che viene tenuto sempre *online*.

DATABASE DOCUMENTALE

Il Fornitore dovrà creare, aggiornare e gestire puntualmente un database documentale (*document store*) di notizie che costituisce l'infrastruttura abilitante all'erogazione del servizio. Da un punto di vista dimensionale il *database* documentale dovrà supportare la gestione di un numero di notizie avente, come ordine di grandezza, quello dei 100 milioni di documenti.
Occorre che nel *database* documentale confluisca il contenuto testuale di tutte le notizie pubblicate giornalmente da un insieme quanto più ampio possibile di fonti italiane (almeno 2.000), altamente rappresentativo (in termini di parametri quali, ad esempio, l'audience raggiunta, la diffusione sul territorio, l'effettiva presenza di contenuto informativo all'interno dell'offerta proposta) di ciascuna delle seguenti categorie:

Fonti:
- Quotidiani nazionali e locali (testate cartacee)
- Telegiornali (incluse le edizioni dei singoli TGR) e altri programmi televisivi di informazione
- Giornali radio e altri programmi radiofonici di informazione
- Agenzie di stampa (siti web)
- Siti web di quotidiani
- Siti web di Tv
- Siti web di radio
- Testate esclusivamente online
- Fonti scientifiche (siti web)
- Pagine e account social (riconducibili almeno alle piattaforme Facebook e Twitter) relative alle fonti di informazione di cui ai punti precedenti, nonché influencer legati al mondo dell'informazione
- Altri siti web, pagine e account social di fonti di informazione non tradizionale, che saranno indicati dall'Autorità.

L'insieme di fonti che alimenterà il *database*, selezionato in collaborazione con l'Autorità, dovrà essere flessibile alle periodiche esigenze di revisione e aggiornamento che dovessero occorrere in base ai mutamenti del contesto di riferimento. In tal senso, durante l'intero corso della fornitura, dovrà essere previsto il nuovo inserimento (e la rimozione) di fonti, in particolare per quelle categorie tipicamente soggette a variazioni più frequenti.
A ciascuna notizia presente nel *database* documentale dovrà essere associata una serie di metadati: identificativo del documento, titolo della notizia, sottotitolo, data, autore, fonte, gruppo di fonti, categoria (cronaca, cultura, economia, esteri, politica, scienza, spettacolo, sport). È necessario che i metadati che corrispondono a predizioni prodotte da algoritmi di *machine learning* supervisionato abbiano una precisione non inferiore al 90%. Più in generale, a servizio in corso, il Fornitore dovrà fornire i dettagli tecnici di tutti gli algoritmi di *machine learning* utilizzati e l'Autorità si riserverà la possibilità di richiedere al Fornitore i *training sets* utilizzati. Ciò, ai fini di disporre di strumenti di rilevazione di eventuali *bias* negli algoritmi di *machine learning* utilizzati nella piattaforma erogante il servizio.

ALIMENTAZIONE DEL DATABASE DOCUMENTALE

Le modalità di acquisizione automatica dei contenuti e conseguente alimentazione del *database* documentale con la trascrizione del testo di ciascuna notizia saranno a cura del Fornitore, tenendo conto delle caratteristiche specifiche di ciascuna categoria di fonte e dei diversi formati delle notizie. L'alimentazione dei dati avverrà con tempistiche di tipo *near real time*. Il *database* documentale dovrà contenere le notizie a partire almeno dal 1° gennaio 2020 e offrirle e mantenerle *online* per l'intera durata della fornitura.

MODALITÀ DI ACCESSO AL DATABASE DOCUMENTALE

a) Interfaccia *web* di ricerca

Il Fornitore dovrà rendere disponibile una *web application* ad accesso controllato di tipo interfaccia utente di motore di ricerca *full-text* con sintassi della query di tipo *Lucene* (con operatori logici AND OR, NOT, parentesi e operatori di prossimità di termini, etc.) e la possibilità, per l'utente, di poter specificare una serie di filtri: data di inizio, data di fine, fonti e gruppi di fonti. La query e i parametri devono poter essere esportati ed importati in formato *JSON* compatibile con le API descritte nel seguito.
Devono inoltre essere possibili ricerche sui metadati delle notizie e per identificativo del documento contenuto nel *database* documentale. La lista dei risultati degli "item di documento" sarà di tipo paginato, dovrà evidenziare il numero totale dei documenti trovati nel *database* e dovrà essere possibile ordinarla secondo un criterio di *ranking* di tipo *full text* oppure in semplice modalità cronologica diretta ed inversa.

b) *Application Programmatic Interface* (API)

Il *database* documentale dovrà essere accessibile attraverso una API *online* di tipo REST ad accesso controllato con formato di interscambio dei dati di tipo *JSON*. Tale API dovrà fornire le seguenti funzionalità: richieste di esecuzione di una query sul *database* documentale con relativi

filtri, fornendo, in risposta, serie storiche di occorrenze oppure, a richiesta, il testo completo delle notizie e tutti i relativi metadati. Dovrà inoltre poter essere possibile la paginazione dei risultati. Il *payload JSON* della *call* di query dovrà coincidere con la query esportata dall'interfaccia *web* di ricerca. Presumendo connessioni di rete efficienti, la velocità effettiva media di download del testo delle notizie, attraverso le API, non dovrà risultare inferiore a 100 notizie testuali al secondo.

ESPORTAZIONE MASSIVA DI DOCUMENTI

Il servizio erogato dovrà altresì consentire di procedere all'estrazione dalla base dati documentale di un numero elevato di notizie (dell'ordine dei milioni) per cui il *download online* via API potrebbe non essere praticabile. Nei casi di richieste di esportazioni massive, dunque, l'estrazione avverrà in modalità *offline* e il risultato sarà reso disponibile, in formato *JSON* compresso, su un server *ftp/sftp* protetto e appositamente messo a disposizione dal Fornitore. La richiesta di una esportazione massiva avverrà invece attraverso apposite API REST il cui formato del *payload JSON* coinciderà ancora con quello esportabile dall'interfaccia *web* di ricerca. Sempre attraverso le API dovrà essere possibile ottenere una stima dei tempi necessari all'espletamento della richiesta.