Programma di formazione

Titolo
Estrazione automatica di riferimenti bibliografici, relativi puntatori e metadati strutturati da documenti accademici in formato PDF

Responsabile scientifico
Professor Silvio Peroni <silvio.peroni@unibo.it>, Dipartimento di Filologia Classica e Italianistica, Università di Bologna / Direttore di OpenCitations, che può essere contattato per ulteriori informazioni.

Obiettivi
OpenCitations (http://opencitations.net/) [1] è una infrastruttura Open Science che mette a disposizione una grossa mole di metadati bibliografici e dati citazionali accademici, di qualità e copertura tali da competere con servizi proprietari, come Web of Science e Scopus. OpenCitations è no-profit e tutti i suoi servizi sono completamente gratuiti. Essa è gestita dal Research Centre for Open Scholarly Metadata dell'Università di Bologna (https://openscholarlymetadata.org).

Negli ultimi tre anni all'interno di OpenCitations è stato sviluppato e testato un nuovo software (lanciato nel dicembre 2022) per creare una nuova raccolta, chiamata OpenCitations Meta (https://opencitations.net/meta). I metadati esposti da OpenCitations Meta includono i metadati bibliografici utilizzati per descrivere informazioni di base di una risorsa bibliografica. In particolare, conserva gli identificatori delle risorse bibliografiche (ad esempio, DOI, PMID, ISSN e ISBN), il titolo, il tipo, la data di pubblicazione, le pagine e il luogo della risorsa (con i numeri di volume e di fascicolo se il luogo è una rivista). Inoltre, OpenCitations Meta contiene metadati riguardanti i principali attori coinvolti nella pubblicazione di una risorsa bibliografica, ovvero gli autori, gli editori e gli editori, ogni attore può essere caratterizzato con altri identificatori (ad esempio ORCID) se disponibili.

Tutte le risorse bibliografiche incluse in OpenCitations Meta [2] derivano da fonti/risorse esistenti (attualmente tali risorse includono Crossref, DataCite e PubMed). In particolare, esiste un record specifico per ogni entità (che cita o che viene citata) coinvolta in ogni citazione inclusa negli Indici di OpenCitations [3] - un indice di citazioni aperte contenente tutte le entità che citano e sono citate, identificate da identificatori persistenti usati dalle fonti (cioè DOI o PMID), che sono coinvolti nelle citazioni.

Uno dei principali scopi di OpenCitations, che è anche l'obiettivo principale di questo lavoro, è quello di processare, con sistemi automatici basati su tecnologie di machine learning e su altri approcci computazionali, documenti in formato PDF di articoli scientifici al fine di individuare determinate porzioni dell'articolo, tra cui i metadati descrittivi dello stesso (titolo, autori, affiliazioni), i riferimenti bibliografici inclusi i metadati che li contraddistinguono (titolo, autori, data di pubblicazione, sede editoriale, curatori, editore, identificatori), e i relativi puntatori a questi ultimi inseriti nel contenuto testuale dell'articolo. Una volta individuate queste porzioni, l'obiettivo è quello di estrarre ed arricchire le informazioni relative,

avvalendosi anche di servizi esterni disponibili su Web ed utilizzabili attraverso API REST, e metterle a disposizione in un formato strutturato che ne permetta il facile processamento da parte di software.

Piano di attività

L'Assegno di Ricerca avrà durata di 21 mesi a partire da Aprile 2024. L'Assegnista di Ricerca lavorerà direttamente con il Professor Silvio Peroni nel contesto del Research Centre for Open Scholarly Metadata, presso il Dipartimento di Filologia Classica e Italianistica dell'Università di Bologna (Italia). Il Centro di Ricerca è un ambiente vivo e stimolante, ed è atteso che l'Assegnista di Ricerca fornisca contributi personali centrali alle attività di OpenCitations. Il lavoro a distanza può essere possibile se strettamente necessario, ma altrimenti la presenza di persona nel Centro di Ricerca è preferibile.

Durante il periodo di lavoro è prevista una fase iniziale introduttiva e conoscitiva del contesto applicativo. Dopodichè il lavoro dell'Assegnista di Ricerca può essere organizzato e riassunto in questi punti:

1) Raccogliere ed organizzare un corpus adeguato di documenti in formato PDF di articoli scientifici da utilizzare come training data e gold standard.
2) Sviluppare uno o più sistemi diversi, basati su tecnologie di machine learning, natural language processing, o altre metodologie computazionali, per identificazione e l'estrazione delle varie parti di interesse che caratterizzano gli articoli accademici, ovvero i metadati descrittivi dell'articolo in considerazione, i riferimenti bibliografici inclusi i metadati che li contraddistinguono, e i relativi puntatori a questi ultimi inseriti nel contenuto testuale dell'articolo.
3) Testare i vari sistemi per identificare i più promettenti – eventualemente anche combinandoli tra loro.
4) Sviluppare un software web-based e delle API che permettano utilizzo di questo sistema di estrazione in modo programmatico.
5) Scrivere e pubblicare una documentazione appropriata per descrivere i risultati ottenuti da questo lavoro.

Mentre il professor Peroni dirigerà e supervisionerà il lavoro, il Borsista di Ricerca avrà la responsabilità di gestire in modo autonomo e sistematico queste attività.

Requisiti

Tutti/e i/le candidati/e devono avere eccellenti abilità come programmatori/trici e, come valore aggiunto, devono essere in grado di parlare, scrivere, e presentare verbalmente a conferenze in un buon inglese. Esperienze dimostrabili di programmazione in Python e utilizzo dei più comuni librerie Python, e sistemi di versionamento basati su Git (in particolare GitHub) sono fortemente desiderabili. In più, è altresì fortemente desiderabile che il/la candidato/a abbia una forte e dimostrabile attitudine verso la Scienza Aperta e la capacità di lavorare in gruppo. Conoscenze dimostrabili nelle tecnologie del Web Semantico, Linked Data e tecnologie Web in generale sono elementi favorevoli per la candidatura.

I requisiti minimi formali per la posizione sono il possesso di una Laurea Magistrale LM43 o equivalente. Il candidato deve avere un'esperienza adeguata e dimostrabile come

programmatore, comprovata dai documenti da allegare in fase di domanda. La candidatura (in Italiano o in Inglese) deve almeno includere un Curriculum Vitae completo di informazioni riguardanti attività scientifico-professionali e relative alla produttività scientifica. Eventuali lettere di raccomandazione sono opzionali, ma fortemente consigliate.

L'Università di Bologna è un'istituzione che da pari opportunità di impiego, e la selezione per questa posizione verrà fatta esclusivamente sul merito.

Riferimenti

1. Peroni, S., & Shotton, D. (2020). OpenCitations, an infrastructure organization for open scholarship. Quantitative Science Studies, 1(1), 428–444. https://doi.org/10.1162/qss_a_00023
2. Massari, A., Mariani, F., Heibi, I., Peroni, S., & Shotton, D. (2023). OpenCitations Meta. arXiv. https://doi.org/10.48550/arXiv.2306.16191
3. Heibi, I., Peroni, S., & Shotton, D. (2019). Software review: COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations. Scientometrics, 121(2), 1213–1228. https://doi.org/10.1007/s11192-019-03217-6