SemplicePA

SEMantic instruments for PubLIc administrator and CitizEns

Martina Miliani - martina.miliani@semplicepa.it

Abstract. **La trasformazione digitale italiana sta procedendo ancora a rilento rispetto a quella Europea, con un digital divide che penalizza soprattutto i comuni più piccoli e gli open data che faticano ad essere pienamente valorizzati, con il risultato che il Foia è ancora ben lungi dall'essere applicato come dovrebbe. Eppure non mancano le iniziative civiche, alcune stimolate dalle stesse amministrazioni. Così come non mancano le tecnologie di eccellenza, sviluppate all'interno di start-up che collaborano con università statali e centri di ricerca. SemplicePA nasce in questo contesto con lo scopo di fornire uno strumento utile alla cittadinanza e alle amministrazioni a partire dalla trasformazione digitale di un archivio spesso sconosciuto, l'Albo Pretorio.**

***Index Terms*—Conservazione e condivisione dei dati, Natural Language Processing, Machine Learning, Big Data.**

I. Introduzione

Otto paesi su dieci, in Europa, hanno attivato una regolamentazione sugli open data. L'Italia, che si trova sotto la media europea, è tra i "follower" delle buone pratiche, con un "Mezzogiorno nettamente indietro"[1]. Tra le cause, il grande divario tra le piccole e le grandi amministrazioni. In base all'osservatorio dell'Istat, le province autonome e l'85,5% dei comuni sopra i 60.000 abitanti possiede un ufficio dedicato all'ICT[1], ovvero poco più dell'1% del totale dei comuni[2]. Anche i risultati del primo monitoraggio sull'applicazione del Freedom of Information Act (Foia) sono tutt'altro che positivi: il 73% degli utenti non ha ricevuto risposta e un diniego su tre era invece illegittimo[3]. Anche per questo, accanto all'Agenzia per l'Italia Digitale (AgID), sta lavorando il Team per la Trasformazione Digitale che ha una diversa concezione delle informazioni possedute dalla PA: "I dati sono nostri e li gestiamo insieme"[4]. Eppure in Italia si registrano già alcune iniziative sull'uso degli open data. Talvolta sono gli stessi enti pubblici a indire contest per premiare il migliore utilizzo degli open data: a vincere l'hackathon sulla disabilità indetto dal Comune di Lecce, è stato il censimento delle barriere architettoniche in città[5]. Ma sono tantissime anche le iniziative civiche, come il progetto di crowdfunding Ricostruzione Trasparente[6], il cui obiettivo è quello di tenere traccia di tutti gli atti pubblici che consentano di esercitare un controllo diffuso sugli attori della ricostruzione in seguito al terremoto del 2016 avvenuto nel Centro Italia. Si tratta di dati rilasciati dalle pubbliche amministrazioni, che sono stati poi rielaborati e resi fruibili in modi totalmente nuovi dai cittadini. Ma tante sono le risorse ancora non adeguatamente valorizzate, come ad esempio, l'Albo Pretorio, l'archivio degli atti di ciascun comune amministrativo.

La prima legge che sancisce la trasformazione digitale dell'Albo Pretorio risale al gennaio 2009 e giunge a pieno regime nel 2013[7]. Nonostante la trasformazione sia avvenuta, cercare un atto all'interno dell'Albo sembra possibile solo se si ha ben presente quale documento ci interessa: gli atti sono ricercabili solo in un arco di tempo ristretto, in genere 15 giorni, e nei siti di molti comuni per recuperare un certo provvedimento è necessario conoscerne la data, l'organo che lo ha emanato, l'oggetto o il suo numero identificativo. A mancare sono soprattutto le relazioni tra i singoli documenti, non solo tra quelli di uno stesso comune, ma anche e soprattutto tra comuni differenti.

II. Un motore di ricerca semantico

Nato nel 2015, SemplicePA[8] ha l'obiettivo di valorizzare i contenuti degli atti registrati negli albi pretori dei comuni di tutta Italia, di rendere navigabili queste informazioni e le relazioni che tra esse intercorrono. Un albo pretorio nazionale che sia possibile esplorare attraverso un motore di ricerca semantico, in grado di estrarre elementi testuali significativi, quali nomi di aziende e organizzazioni, e mostrare come sono collegati attraverso strumenti di visual analytics.

Un motore di ricerca semantico nasce con lo scopo di districarsi tra le ambiguità del lessico. La collocazione dei termini all'interno di un'ontologia permette di distinguere, ad esempio nei casi di omonimia, a cosa l'utente che lo interroga si stia riferendo. Se l'interrogazione è posta in linguaggio naturale, inoltre, è il contesto nel quale un termine è inserito ad aiutare il motore di ricerca ad associare il termine alla corretta categoria.

In Italia, Cogito[9] si basa sull'analisi semantica dei testi grazie a un'ampia banca dati che vede raggruppate più ontologie differenti, costruite anche in diverse lingue. È nato invece all'Università di Pavia, Facility Live, che mostra il suo valore nei domini più ristretti: l'ontologia dietro a motori di ricerca come questo è molto più "specializzata" e per questo anche precisa e puntuale nel recupero delle informazioni richieste dall'utente[10]. Légilocal è un motore di ricerca semantico che in Francia si occupa della gestione degli atti, dedicando anche un framework apposito per la loro stesura, in modo che siano facilmente leggibili ed elaborabili dal motore di ricerca[11]. Sugli enti locali in Italia si è specializzato Sophia Semantic Search. Questo motore di ricerca riconosce le entità elencate all'interno dei documenti e li classifica per similarità[12].

III. SemplicePA

Più similmente a Légilocal, SemplicePA è una piattaforma, un ambiente dotato di vari componenti, di seguito elencati.

A. Estrazione delle entità

[1] Information & Communication Technology.

All'interno di ogni documento sono individuate diverse entità: persone, dei luoghi, delle aziende, delle organizzazioni, importi, date e indirizzi email ma anche elementi più specifici dei provvedimenti amministrativi come riferimenti legislativi e ad altri atti, partite iva, codici identificativi di gara e codici fiscali. L'estrazione avviene attraverso un modello che integra due approcci diversi, uno basato su delle regole e un altro su calcoli statistici. L'approccio ruled-based è ordinato da algoritmi che contengono regole precise sull'estrazione. Ad esempio, la regola che estrae nomi e cognomi dovrà estrarre due parole una di seguito all'altra che inizino per lettera maiuscola, e così via. L'altro approccio vede la collaborazione del Dipartimento di Filologia, Letteratura e Linguistica dell'Università e del Centro di Linguistica Computazionale del Cnr di Pisa: un modulo di T2K[13], detto NLP Analyzer[2], ed Extra[14] estraggono in maniera automatica e non supervisionata le entità in base a calcoli statistici. La probabilità che i termini estratti siano delle entità è dedotta dalla distribuzione del lessico all'interno del testo, appresa inizialmente da un corpus, un insieme di atti, annotato manualmente. Un altro modulo di elaborazione è infine dedicato alla "normalizzazione" che sostituisce con una forma univoca e standard le diverse declinazioni in cui una stessa entità è presente nei documenti.

B. Ontologia

L'ontologia su cui si basa SemplicePA è costruita con un modello top-down e bottom-up, sia attraverso l'individuazione di termini e della loro classificazione da parte di esperti di dominio, sia estraendo i termini più rappresentativi di una classe in seguito all'analisi automatica degli atti. Anche di quest'ultima classificazione automatica, detta di topic modeling, si è occupata l'Università di Pisa. Si tratta di un metodo unsupervised, che non necessita cioè di una precedente annotazione degli atti. Il tool open source utilizzato per questo compito è chiamato Mallet[15] e si avvale dell'algoritmo di Latent Dirichlet Allocation[16]. Le classi corrispondenti ai topic quindi, da una parte vengono individuate in base alla presenza dei termini dell'ontologia, e dall'altra da termini automaticamente estratti dagli stessi testi: il risultato è un approccio integrato dei due metodi. Attraverso il LDA è inoltre possibile estrarre da uno stesso documento più di una categoria, per cogliere le diverse sfumature semantiche dell'atto amministrativo.

C. Network Analysis

Le relazioni tra le entità presenti nei documenti sono calcolate dalla piattaforma sulla base della compresenza all'interno degli atti. Una sezione è appositamente dedicata alla visualizzazione di reti in cui nodi sono le entità estratte, le relazioni sono gli archi che le collegano mentre il peso degli archi è dato dal numero di documenti in cui le entità collegate sono entrambe nominate. Si può partire da una persona, un'organizzazione o un'azienda e decidere di visualizzare in una rete diversi tipi entità ad essa "collegate" (ancora aziende, persone, organizzazioni). Oppure è possibile visualizzare le relazioni tra gli elementi a partire da un gruppo di documenti selezionati.

[2] Ovvero Natural Language Processing Analyzer.

D. Altri strumenti

Tra gli altri strumenti offerti dalla piattaforma, una mappa, caricata automaticamente da OpenStreetMap[17] all'invio di ciascuna query, segnala i comuni e i luoghi citati all'interno dei documenti; le entità che appaiono all'interno dei documenti restituiti all'utente sono ordinate per frequenza, in modo da fornire una panoramica generale dei contenuti; viene inoltre mostrato l'iter per ciascun testo, che segue i riferimenti agli atti amministrativi nei documenti ricostruendo, in forma gerarchica, l'ordine di pubblicazione dei testi che si riferiscono ad un unico provvedimento; in fondo alla pagina di consultazione dell'atto sono presentati i documenti simili; una sezione di visual analytics mostra infine i trend delle pubblicazioni degli atti nel tempo, in base ai vari argomenti; e infine, una chat mette in contatto quanti sono connessi alla piattaforma.

IV. Conclusioni

SemplicePA nasce per valorizzare gli atti amministrativi dei comuni di tutta Italia, grazie all'applicazione delle tecnologie del linguaggio più innovative che consentono anche all'utente inesperto del dominio della PA di consultare con facilità gli atti pubblicati.

V. Bibliografia

[1] Luca Tremolada, "L'Europa dei dati. Otto paesi su dieci hanno regole sugli open data, Il Sole 24 Ore, 5 Aprile 2017 (goo.gl/ZqVAfG).
[2] Istat, "Le tecnologie dell'informazione e della comunicazione nella pubblica amministrazione locale", 2015 (goo.gl/7q4Loq).
[3] Diritto di Sapere, "Ignoranza di Stato" (goo.gl/K4Cgx1).
[4] Raffaele Lillo, "Data & Analytics Framework", Medium, 13 Febbraio 2017 (goo.gl/CtQXNy).
[5] Lecce, Luoghi accessibili per disabilità varie e di interesse, Umap, 30 Aprile 2016 (goo.gl/NJq6g7).
[6] Ricostruzione Trasparente (http://partecipa.ricostruzionetrasparente.it/).
[7] Qualità PA, "Albo Pretorio Online" (goo.gl/LxNLV7).
[8] SemplicePA (http://www.semplicepa.it/).
[9] Cogito, Expert System (goo.gl/CWJE2o)
[10] Luca Piana, "Facility LIve, start-up italiana che sfida Google", L'Espresso, 14 Dicembre 2015 (goo.gl/sjxblF).
[11] Florence Amardeilh e altri, "The Légilocal project: the local low simply shared", 2013 (goo.gl/B4r10N).
[12] The Best of Innovazione ICT, Celi (goo.gl/C0cHT5).
[13] "Dal testo alla conoscenza: le tecnologie del linguaggio per il Knowledge Management", Consiglio Nazionale delle Ricerche (goo.gl/5tu0py).
[14] Lucia C. Passaro e Alessandro Lenci, "Extracting Terms with EXTra", 2015 (goo.gl/cq8uda).
[15] Blei, 2012. Probabilistic Topic Models. Magazine Communications of the ACM CACM Homepage archive Volume 55 Issue 4, April 2012. Pages 77-84. ACM New York, NY, USA.
[16] David M. Blei, Andrew Y. Ng and Michael I. Jordan, 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research 3 (2003) 993-1022.
[17] OpenStreetMap (https://goo.gl/V96Vsw).