# Progetti IA2 a.a. 2018/2019

**Traccia 1:**

Nel contesto dei Big Data dobbiamo confrontarci con grandi volumi di dati non strutturati, spesso testuali, provenienti da domini diversi e di tipologia diversa (es. articoli di giornale, blog post, post sui social network, pagine di siti specializzati, etc.). Nel cercare di dare un senso a questi dati ed estrarre valore da essi, il paradigma dell’*Open IE* ci permette di identificare in maniera veloce e robusta (rispetto all’eterogeneità dell’input) le relazioni asserite in documenti testuali.

Lo scopo di questa traccia è di applicare tecniche di *Open IE* per la costruzione di un *knowledge graph*. In particolare, si richiede di affrontare le problematiche che si presentano quando si vuole andare oltre la semplice estrazione di tuple testuali e costruire un vero e proprio knowledge graph.

I principali requisiti sono:

* Identificare il corpus documentale di input (es. pagine di Wikipedia, definizioni da Wikipedia, pagine web su uno stesso argomento o su argomenti diversi, articoli di giornale, etc.). Si può decidere di avere più corpus con caratteristiche diverse (es. stile espressivo, uso di anafora, affermazioni ridondanti), soprattutto nell’ottica di analizzare l’influenza di queste caratteristiche sulla progettazione e le prestazioni del sistema
* Generalità delle assunzioni linguistiche (es. non limitarsi a relazioni espresse da verbi, o presupporre che la relazione debba comparire necessariamente tra i propri argomenti)
* A differenza dell’Open IE “in senso stretto” che elabora frase per frase, si chiede di affrontare problemi di co-referenza all’interno di uno stesso documento. Come esempio si consideri un ipotetico documento formato da queste due frasi che esprimono 2 triple: “Albert Einstein was a physicist. He developed the theory of relativity.” Il pronome “he” è co-referente con la named entity “Albert Einstein”, pertanto le triple estratte dalle due frasi hanno lo stesso soggetto a livello semantico.
* Affrontare, almeno a livello di problematiche, la questione dell’identità all’interno del corpus. Per esempio, si può considerare il problema di distinguere Micheal Jordan giocatore da Micheal Jordan professore, etc. oppure quello di capire che Roma e Roma Capitale sono (in certi contesti) la stessa cosa
* Affrontare problemi analoghi per le relazioni. In questo caso, può essere utile servirsi di risorse lessico-semantiche come Princeton WordNet.
* Discutere le problematiche relative a relazioni non binarie
* Affrontare il problema della minimalità delle tuple estratte. Gli argomenti testuali non sono sempre delle named entity (es. Albert Einstein), ma spesso ci troviamo di fronte a noun phrase più complesse o all’uso di vari tipi di congiunzioni. Occorrerà domandarsi, per esempio, se dalla frase “oranges and lemons are sources of vitamin C” si debba estrarre una sola tupla con la relazione “be source of”.
* Considerare che una frase potrebbe non affermare semplicemente una proposizione (es. “John brings the umbrella”), ma esprimere un fatto condizionale (es. “If it rains, John brings the umbrella”) oppure l’attitudine di un soggetto nei confronti di una proposizione (es. “Mark believes that John brings the umbrella”).
* Stimare la confidenza di una tupla (che è affetta dal processo di estrazione e dalla fonte, che potrebbe fornire informazioni false)
* Aspetti diacronici (es. validità temporale di una tupla)
* Rappresentazione dei risultati in RDF per la creazione del knowledge graph.

Discutere se sono stati usati KG esterni, come fonte di conoscenza di background o per fare entity linking. Discutere se sono state usate altre risorse esterne (es. WordNet, VerbNet, etc.)

Analizzare i risultati prodotti dal sistema e valutarne le prestazioni. Per questo aspetto ed in modo particolare per quanto riguarda l’estrazione delle tuple, è possibile confrontarsi con altri sistemi di Open IE disponibili gratuitamente.

**Traccia 2**

I Big Data sono associati a grandi quantità (volume) di dati eterogenei (varietà) di cui occorre servirsi per trarne un qualche vantaggio (value), nonostante l’incertezza relativa alla qualità dei dati e del processo di estrazione (veracity). In aggiunta, si pretende che il processo di analisi venga attivato prontamente appena i dati sono disponibili e che i risultati siano prodotti in tempi ragionevolmente brevi (velocity).

Lo scopo di questa traccia è confrontarsi col problema del volume (ed in misura minore della velocity) nel contesto della gestione di dataset RDF. Nello specifico si chiede di discutere diversi approcci basati su architetture di calcolo distribuito (es. Hadoop o Spark) o sulla federazione di triplestore, di confrontarle ed implementare un paio di soluzioni.

Alcuni temi di interesse da affrontare nella discussione e nella realizzazione dei sistemi:

* Varietà degli obiettivi (es. affidabilità, disponibilità, scalabilità rispetto alla dimensione dei dati, etc.)
* Storage dei dati RDF: diverse strategie (spesso legate alla scelta dell’infrastruttura sottostante) per rappresentare le triple RDF e immagazzinarle
* Interrogazione (es. valutazione di singoli triple pattern, valutazione di basic graph pattern, etc.). Per esempio, si può pensare di implementare un frammento di SPARQL.
* Reasoning (rispetto a diversi tipi di semantica, per RDFS)

Il predicato *owl:sameAs* può essere usato per integrare dataset diversi, ma può determinare un’inflazione degli statement contenuti all’interno della base di conoscenza. Discutere eventuali ottimizzazioni nell’implementazione di questo predicato.

Discutere come il sistema gestisce l’aggiornamento dei dati (es. caricamento da zero vs aggiornamento puntuale).

Valutare le prestazioni del sistema, al variare delle risorse hardware disponibili, della taglia/tipologia dei dati e del carico di lavoro sul sistema.