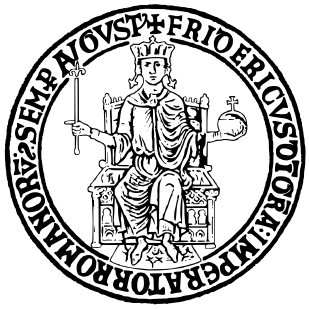
Università degli Studi di Napoli Federico II



Scuola Politecnica e delle Scienze di Base

Dipartimento di Ingegneria Elettrica e Tecnologie dell'Informazione

Corso di Laurea Magistrale in Ingegneria Informatica

Sistema Pipelined di big data per l'analisi di articoli di geopolitica in lingua italiana.

Studenti
Raffaele del Gaudio M63001389
Davide Vitale M63001469

Docente
Prof. Vincenzo Moscato

Anno Accademico 2023/2024

Indice

Capitolo 1

Introduzione

L'obiettivo del progetto è realizzare una pipeline di big data analytics per l'estrazione di **entità** e **relazioni** da articoli di giornale per costruirne un **Knowledge Graph**. Tale grafo deve essere usato per produrre delle analytics che sono usufruibili da un utente mediante una dashboard web.

Il sistema realizzato si specializza sulla costruzione di un grafo per analisi geopolitiche ed economiche a partire da articoli in lingua italiana del sito *notiziegeopolitiche.net*.

1.1 Modello delle Entità e Relazioni

L'estrazione di entità e relazioni dagli articoli di giornale viene guidata attraverso una precisa definizione ex ante del *modello delle entità e relazioni*.

Questo modello enumera:

- Tutte e sole le tipologie di entità che verranno ricercate negli articoli. Ad ogni tipologia di entità verrà associata una diversa label nel grafo della conoscenza.
- Tutte e sole le relazioni che verranno ricercate negli articoli. Queste sono le uniche relazioni che sarà possibile trovare nel grafo della conoscenza.

Le relazioni previste sono di due tipologie: *neutre* e *sentimentali*. Una relazione neutra è tale quando essa esprime una relazione che non ha senso quantificare secondo la sentiment analysis (e.g.: Nella frase "Joe Biden incontra Angela Merkel a Berlino." la relazione "incontrare" non può essere soggetta ad una sentiment analysis).

Una relazione sentimentale è tale quando essa può essere soggetta a sentiment analysis e quindi può esistere nell'accezione positiva o negativa (e.g.: Nelle due frasi "Il pubblico apprezza i discorsi di Joe Biden." e "Il pubblico disprezza le politiche di austerità." è presente la relazione "apprezzare/disprezzare" - rispettivamente - in forma positiva e negativa. Si noti che nella frase "Marco Bianchi non apprezza il lavoro di Joe Biden" la relazione "apprezzare/disprezzare" è in forma negativa).

1.1.1 Tipologie di entità

- **Persona**: Nome di individuo rilevante nel contesto dell'articolo. Es. politici, leader economici, amministratori delegati, analisti economici, ecc. Es. "Joe Biden", "Angela Merkel", "Elon Musk".

- **Organizzazione**: Nome di ente, azienda, istituzione, ONG, partito politico, e organismo internazionale. Es. "Nazioni Unite", "Banca Mondiale", "Google", "Partito Democratico".

- **Località**: Nome di luogo geografico, come paese, città, regione, e continente. Es. "Italia", "New York", "Asia", "Europa".

- **Data o Periodo Temporale**: Riferimento a data specifica, anno, mese, giorno della settimana, o periodo di tempo. Es. "12 marzo 2024", "nel 2023", "secondo trimestre", "decennio del 1990".

- **Misura o Quantità**: Misura o quantità numerica che si riferisce a misure specifiche, come l'inflazione, il PIL, e altre metriche economiche. Es. "100 milioni di dollari", "50.000 euro", "3 milioni di barili", "200 miliardi di metri cubi", "5%", "2,5%".

- **Prodotto o Servizo**: Nome di bene o servizo, spesso associato a specifici settori economici. Es. "iPhone", "energia solare", "servizi bancari".

- **Evento**: Nome di evento significativo che ha impatto economico o geopolitico. Es. "Brexit", "G20", "Forum Economico Mondiale".

- **Legge, Regolamento o Programma Politico**: Nome di legge, regolamento, trattato o accordo internazionale. Es. "Accordo di Parigi", "Regolamento GDPR", "Trattato di Maastricht", "politica monetaria espansiva", "riforma fiscale", "politica di austerità", "Belt and Road Initiative", "Piano Marshall", "Progetto Manhattan".

- **Settore Economico**: Riferimento a specifico settore dell'economia. Es. "settore tecnologico", "industria manifatturiera", "settore dei servizi".

- **Risorsa Naturale**: Nome di risorsa naturale che ha rilevanza economica e geopolitica. Es. "petrolio", "gas naturale", "minerali rari".

- **Entità Governativa**: Nome di dipartimento, ministero o altra entità governativa. Es. "Ministero delle Finanze", "Dipartimento di Stato", "Banca Centrale Europea".

- **Infrastruttura**: Nome di infrastruttura strategica che può influire su geopolitica ed economia. Es. "Canale di Suez", "Pipeline Nord Stream", "Autostrada del Sole".

- **Tecnologia**: Nome di tecnologia emergente o consolidata che influenza l'economia e la geopolitica. Es. "intelligenza artificiale", "blockchain", "5G".
- **Incidente o Crisi**: Riferimento a incidente significativo o crisi economica o politica. Es. "crisi del debito sovrano", "attacco informatico SolarWinds", "pandemia COVID-19".

1.1.2 Relazioni

Relazioni neutre

- **Incontrare**: Es."Joe Biden incontra Angela Merkel a Berlino."
- **Discutere**: Es."I leader mondiali discutono delle questioni climatiche al G20."
- **Partecipare**: Es."Elon Musk partecipa a una conferenza sull'innovazione tecnologica."
- **Annunciare**: Es."La Banca Mondiale annuncia nuovi finanziamenti per l'Africa."
- **Riferire**: Es."Le Nazioni Unite riferiscono sulle condizioni umanitarie in Siria."
- **Prevedere**: Es."Gli economisti prevedono una crescita del PIL per il 2025."
- **Collaborare**: Es."Google e Microsoft collaborano su progetti di intelligenza artificiale."
- **Pubblicare**: Es."L'ONU pubblica un rapporto sul cambiamento climatico."
- **Registrare**: Es."L'Italia registra un aumento delle esportazioni nel primo trimestre."
- **Lanciare**: Es."Tesla lancia un nuovo modello di auto elettrica."
- **Monitorare**: Es."La BCE monitora l'inflazione nell'Eurozona."
- **Analizzare**: Es."Gli analisti economici analizzano le tendenze di mercato."
- **Accettare**: Es."La Grecia accetta gli aiuti finanziari internazionali."
- **Sviluppare**: Es."L'azienda sviluppa nuove tecnologie per l'energia solare."
- **Rivedere**: Es."Il governo rivede le normative sul lavoro."
- **Osservare**: Es."I media osservano le reazioni dei mercati finanziari."

- **Intervistare**: Es."I giornalisti intervistano il CEO di Tesla."
- **Firmare**: Es."I leader europei firmano un accordo commerciale."
- **Trasmettere**: Es."Le notizie trasmettono aggiornamenti sulla situazione economica."
- **Raccogliere**: Es."Gli esperti raccolgono dati sulle nuove tecnologie."
- **Valutare**: Es."Le agenzie di rating valutano il rischio di credito delle aziende."
- **Pianificare**: Es."Il governo pianifica nuove politiche fiscali."
- **Coordinare**: Es."Le organizzazioni internazionali coordinano gli aiuti umanitari."
- **Proporre**: Es."Il ministro delle Finanze propone una nuova riforma fiscale."
- **Progettare**: Es."La società progetta un nuovo impianto di produzione."
- **Consultare**: Es."Il governo consulta gli esperti prima di adottare nuove leggi."
- **Implementare**: Es."Le aziende implementano nuove strategie di marketing."
- **Condurre**: Es."I ricercatori conducono uno studio sulle energie rinnovabili."
- **Indagare**: Es."Il governo indaga sulle cause dell'inflazione."
- **Verificare**: Es."Gli ispettori verificano la conformità delle normative ambientali."
- **Esplorare**: Es."Gli scienziati esplorano nuove fonti di energia rinnovabile."
- **Confermare**: Es."Il ministro conferma la data delle elezioni."
- **Presentare**: Es."Il consiglio direttivo presenta il bilancio annuale."
- **Promuovere**: Es."L'organizzazione promuove l'educazione finanziaria."
- **Stabilire**: Es."Il trattato stabilisce nuove regole commerciali."

Relazioni sentimentali

NOTA BENE: si considerano relazioni in senso negativo anche quelle che utilizzano la negazione della forma positiva, ad esempio "Luca non apprezza il lavoro di Maria" presenta la relazione "Apprezzare/Disprezzare" in senso negativo.

- **Apprezzare/Disprezzare**: Es. positivo "Il pubblico apprezza i discorsi di Joe Biden.". Es. negativo "Il pubblico disprezza le politiche di austerità."
- **Supportare/Ostacolare**: Es. positivo "La Germania supporta le iniziative dell'Unione Europea.". Es. negativo "La Germania ostacola le nuove regolamentazioni fiscali."
- **Promuovere/Boicottare**: Es. positivo "Google promuove l'adozione di tecnologie sostenibili.". Es. negativo "Google boicotta le nuove leggi sulla privacy."
- **Investire/Tagliare**: Es. positivo "La Banca Mondiale investe nei progetti di sviluppo in Africa.". Es. negativo "La Banca Mondiale taglia i fondi per i progetti in Europa."
- **Celebrare/Ignorare**: Es. positivo "L'Italia celebra la Festa della Repubblica il 2 giugno.". Es. negativo "L'Italia ignora le raccomandazioni dell'Unione Europea."
- **Commemorare/Dimenticare**: Es. positivo "Le Nazioni Unite commemorano il Giorno della Terra.". Es. negativo "Le Nazioni Unite dimenticano gli anniversari storici importanti."
- **Accogliere/Respingere**: Es. positivo "L'Europa accoglie i rifugiati.". Es. negativo "L'Europa respinge le richieste di asilo."
- **Valorizzare/Trascurare**: Es. positivo "L'Egitto valorizza il Canale di Suez come infrastruttura chiave.". Es. negativo "L'Egitto trascura la manutenzione delle strade locali."
- **Adottare/Rifiutare**: Es. positivo "Le aziende adottano l'intelligenza artificiale.". Es. negativo "Le aziende rifiutano di implementare nuove tecnologie."
- **Incoraggiare/Dissuadere**: Es. positivo "Il governo incoraggia l'innovazione tecnologica.". Es. negativo "Il governo dissuade gli investimenti esteri."
- **Esaltare/Denigrare**: Es. positivo "La stampa esalta le capacità diplomatiche di Angela Merkel.". Es. negativo "La stampa denigra le scelte politiche di Boris Johnson."

- **Collaborare/Confliggere**: Es. positivo "I paesi collaborano per affrontare il cambiamento climatico.". Es. negativo "I paesi confliggono per le risorse naturali."
- **Proteggere/Minacciare**: Es. positivo "La nuova legge protegge i diritti dei lavoratori.". Es. negativo "La nuova politica minaccia la libertà di stampa."
- **Favorire/Impedire**: Es. positivo "Le politiche favoriscono la crescita economica.". Es. negativo "Le regolamentazioni impediscono l'espansione del mercato."
- **Riconoscere/Negare**: Es. positivo "L'ONU riconosce l'indipendenza del nuovo stato.". Es. negativo "L'ONU nega la legittimità delle elezioni."
- **Innovare/Restare indietro**: Es. positivo "Apple innova con nuovi prodotti ogni anno.". Es. negativo "Apple resta indietro rispetto alla concorrenza nel settore laptop."
- **Sostenere/Ostacolare**: Es. positivo "Il governo sostiene le start-up tecnologiche.". Es. negativo "Il governo ostacola le iniziative imprenditoriali locali."
- **Consolidare/Indebolire**: Es. positivo "Le nuove leggi consolidano la stabilità finanziaria.". Es. negativo "Le nuove regolamentazioni indeboliscono il settore bancario."
- **Motivare/Demoralizzare**: Es. positivo "Il presidente motiva i cittadini con discorsi ispiratori.". Es. negativo "Le politiche di austerità demoralizzano la popolazione."
- **Risolvere/Complicare**: Es. positivo "L'accordo internazionale risolve le dispute commerciali.". Es. negativo "Le nuove tariffe complicano le relazioni economiche."

1.2 Glossario dei termini

- **entità**: parte di un discorso che rappresenta un nome. E.g.: Nella frase "Luca corre verso il mare", "Luca" e "mare" sono due entità. In particolare "Luca" è un'entità di tipo *persona* e "mare" è di tipo *luogo*.
- **relazione**: parte di un discorso che rappresenta un legame tra due entità. Dall'esempio precedente, "corre verso" è la relazione che lega "Luca" e "mare".
- **Knowledge Graph**: un grafo dove i nodi sono le entità e gli archi sono le relazioni.

Capitolo 2

Requisiti del sistema

2.1 Requisiti funzionali [RF]

1. Il sistema deve prevedere una **pagina principale** dove vengono mostrati il numero totale di entità, tipologie di entità e di relazioni conservate nel sistema.

2. Il sistema deve prevedere un **costruttore dinamico di grafo**. Questo vuol dire che l'utente può ricercare per nome un'entità *radice* dalla lista di quelle disponibili e posizionarla nel grafo. Posizionata la radice, egli deve poter selezionare, dalla lista delle tipologie di entità del modello, quelle di interesse e il grafo deve aggiornarsi per visualizzare tutte le entità di quella/e tipologia che hanno una relazione con la radice. Cambiando il periodo temporale il grafo deve aggiornarsi automaticamente.

3. Il sistema deve prevedere un **analizzatore di trend**. Questo vuol dire che l'utente può visualizzare un grafico dell'andamento nel tempo del numero di volte che un'entità è stata rilevata negli articoli analizzati. Questo grafico deve essere costruito con una granularità mensile dell'asse temporale.

4. Il sistema deve prevedere un **analizzatore di sentimento**. Questo vuol dire che l'utente può ricercare per nome un'entità radice ed una relazione sentimentale e visualizzare un grafico in cui vengono poste tutte le entità che hanno avuto relazioni sentimentali con l'entità radice. Questo grafico deve quantificare il sentimento positivo o negativo della relazione tra le entità visualizzate e quella radice.

5. Il sistema deve prevedere una **mappa delle connessioni**. Questo vuol dire che l'utente può visualizzare una mappa in cui è possibile distinguere quanto ogni paese è stato menzionato negli articoli analizzati di mese in mese. L'utente dovrà poter vedere i dati della mappa anche in forma tabellare ed in altri modi che gli possano permettere di apprezzare questa proprietà.

2.2 Requisiti non funzionali [RNF]

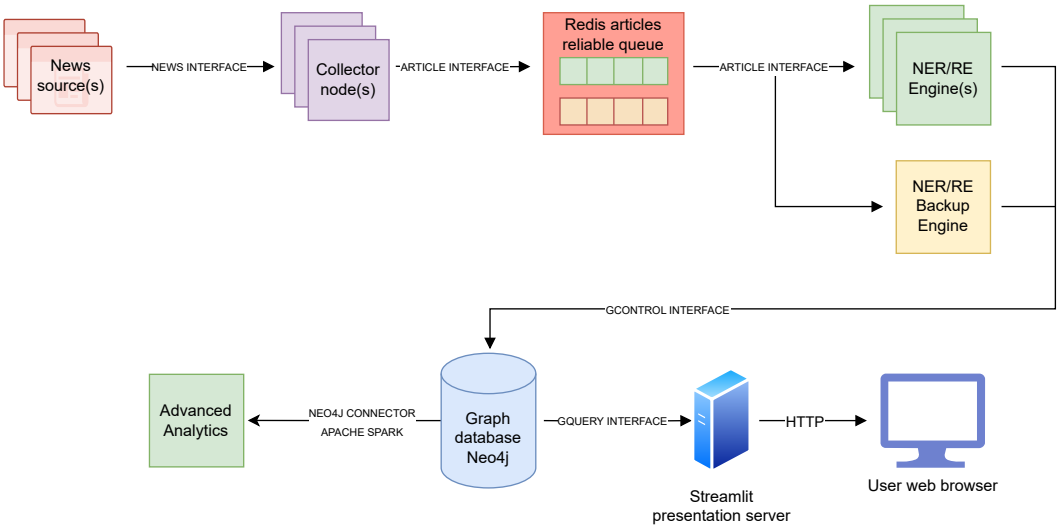
1. Il sistema deve essere modulare e scalabile orizzontalmente per supportare grandi moli di articoli analizzati.

2. Il sistema deve essere tenere in considerazione i fallimenti dei componenti e non deve mai essere possibile perdere un articolo se questo è stato inserito nella pipeline di analisi.

3. Il sistema deve prevedere un meccanismo di asincronicità tra l'estrazione degli articoli e l'analisi di questi ultimi per effettuare NER ed RE. Questo vuol dire che il componente che estrae gli articoli può funzionare ad un rate molto diverso da quello che li analizza e produce il grafo della conoscenza.

Capitolo 3

Architettura del sistema



3.1 Componenti

3.1.1 Collector node

Il *Collector node* è il componente che offre al sistema complessivo le due seguenti funzionalità:

- Si interfaccia con i siti di notizie tramite la *NEWS INTERFACE* per estrarre gli articoli.

- Trasforma gli articoli in forma grezza recuperati dai diversi siti di notizie in una forma standard definita dalla *ARTICLE INTERFACE*.

- Usa le funzionalità della *ARTICLE INTERFACE* per inserire nella *Redis articles reliable queue* gli articoli in forma standard.

L'architettura prevede la possibilità di istanziare molteplici collector in parallelo.

3.1.2 Redis articles reliable queue

Il componente *Redis Articles Reliable Queue* consente di evitare la perdita di articoli attraverso l'implementazione di due code su Redis: una principale e una di backup.

Dalla coda principale è possibile effettuare un'operazione di pop che rimuove un batch di articoli e lo inserisce nella coda di backup. Dalla coda di backup, i batch di articoli possono essere prelevati tramite un'operazione di peek o eliminati tramite la chiave.

La presenza di *Redis Articles Reliable Queue* consente ai componenti *NER/-RE Engine* e *NER/RE Backup Engine* di elaborare gli articoli ad un rate diverso rispetto a quello di estrazione dai siti di notizie.

3.1.3 NER/RE Engine

Il componente *NER/RE Engine* si occupa dell'elaborazione dei batch di articoli, estraendone entità e relazioni rilevanti e adattandole ad un formato standard. In particolare esso offre le seguenti funzionalità:

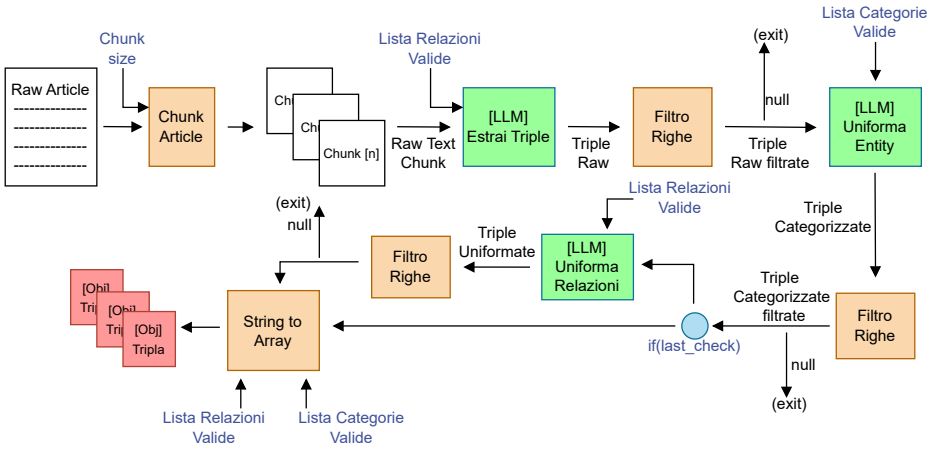
- Si interfaccia alla Redis reliable queue tramite *ARTICLE INTERFACE* per prelevare i batch di articoli dalla coda principale.

- Si occupa di analizzare e manipolare il testo di ciascun articolo restituendo entità e relazioni sotto forma di triple standardizzate e categorizzate secondo il *modello delle relazioni*.

- Effettua l'inserimento delle triple nel Graph Database tramite la *GCONTROL INTERFACE*

- Rimuove il batch dalla coda di backup di Redis utilizzando le funzioni della *ARTICLE INTERFACE*.

Il processo di estrazione delle triple *(entità1[categoria1],relazione,entità2[categoria2])* è modellato attraverso una pipeline che alterna l'uso di un LLM (LLama3 8B) a controlli e filtraggi deterministici.



I filtraggi deterministici consentono di uscire anticipatamente dalla pipeline se l'output intermedio non è conforme allo standard.
I prompt forniti al LLM sono composti da: le direttive sull'elaborazione, il chunk corrente e le liste delle Relazioni e Categorie di entità valide. Lo stadio finale effettua un controllo deterministico sui termini utilizzati e converte la stringa delle triple in una serie di oggetti.
NER/RE Engine consente di configurare i seguenti parametri che influenzano la velocità di elaborazione degli articoli:

- Numero massimo di tentativi di estrazione di triple da uno stesso chunk
- Dimensione del chunk
- Flag di ultimo controllo delle relazioni tramite LLM

L'architettura prevede la possibilità di istanziare il componente *NER/RE Engine* più volte per aumentare il rate di processamento degli articoli.

3.1.4 NER/RE Backup Engine

NER/RE Backup Engine fornisce le stesse funzionalità del componente *NER/RE Engine*, ad eccezione del fatto di prelevare articoli dalla coda di backup. Questo componente, difatti, serve a garantire che gli articoli che erano sotto analisi durante il fallimento di un componente *NER/RE Engine* vengano rianalizzati e le triple estratte inserite nel knoledge graph.

A differenza di *NER/RE Engine*, questo componente non può essere istanziato più volte in parallelo.

3.1.5 Streamlit Presentation Server

Streamlit Presentation Server implementa una dashboard interattiva che prevede le seguenti pagine:

1. *Homepage*: mostra statistiche generali del grafo e un menu delle top 10 entità più citate per il mese selezionato.

2. *Map explorer*: permette di visualizzare quanto ogni paese è stato menzionato negli articoli analizzati mese per mese, con la possibilità di visualizzare i dati anche in forma tabellare e con un grafico a barre.

3. *Trend Viewer*: permette di visualizzare un grafico dell'andamento mensile del numero di volte che un'entità è stata rilevata negli articoli analizzati.

4. *Political Compass*: permette di ricercare un'entità radice e una relazione sentimentale, visualizzando un grafico che quantifica il sentimento positivo o negativo delle relazioni con l'entità radice.

5. *Graph Explorer*: permette all'utente di ricercare un'entità radice per nome e selezionare le tipologie di entità di interesse per aggiornare il grafo automaticamente in base al periodo temporale selezionato.

3.1.6 Advanced Analytics

Il componente *Advanced Analytics* si interfaccia al Graph Database tramite il *Neo4j Connector Apache Spark* ed effettua analisi sull'intero grafo utilizzando Spark, GraphX e SparkML. In particolare, vengono applicati i seguenti algoritmi: PageRank, TriangleCounting e LabelPropagation. I coefficienti estratti dal grafo vengono poi utilizzati per addestrare un modello di predizione delle relazioni attraverso l'algoritmo RandomForest.

3.2 Interfacce

3.2.1 News Interface

La *News Interface*:

- Definisce le funzioni usate dal Collector node per acquisire gli articoli da tutti i provider supportati

- Definisce per ogni provider supportato la funzione di standardizzazione degli articoli

3.2.2 Article Interface

La *Article Interface* definisce la struttura articolo standard in questo modo:

```
{
    "title": "Titolo dell'articolo",
    "link": "https://www.article-link.com",
    "date": "24-10-2023",
```

```
    "text": "Recenti scoperte dimostrerebbero che..."
}
```

e definisce le funzioni per interagire con la Redis article reliable queue sia in push che in pop.

3.2.3 Gcontrol Interface

La *Gcontrol Interface* definisci le funzioni per effettuare l'inserimento delle triple nel Graph Database Neo4j.

3.2.4 Gquery Interface

La *Gquery Interface* definisce le funzioni utilizzate dallo *Streamlit Presentation Server* per estrarre informazioni dal Graph Database tramite query cypher.