Come tutti sappiamo, il Web è la più grande, eterogenea e dimanica sorgente di informazioni liberamente fruibile da chiunque:

* **Grande** perché il numero di produttori di informazioni è uguale al numero di consumatori, facendo crescere costantemente le informazioni sul Web.
* **Eterogenea** perché dipende sia dal formato delle pagine sia dal contenuto testuale. Nel primo caso, esistono 3 tipologie:
  + **Pagine non strutturate**: pagine scritte in linguaggio naturale e non sono caratterizzate da nessuna struttura;
  + **Pagine strutturate**: pagine generate da una sorgente dati di tipo strutturato (database): i dati vengono pubblicati una volta che vengono inseriti in una qualche struttura (forma tabellare);
  + **Pagine semistrutturate**: sono una via di mezzo delle tipologie precedentemente descritte.

Nel secondo caso, l’eterogeneità del Web è definita dal fatto che le pagine vengono create da milioni di persone aventi differente cultura, abilità e linguaggio. Da questo si deduce che le pagine Web possono avere informazioni simili o uguali, ma presentate e organizzate in maniera completamente differente.

* **Dinamica** perché ogni secondo vengono create, distrutte e modificate migliaia di pagine. Il contenuto e la struttura delle pagine Web cambiano con frequenza.

Estrarre l’informazione dal Web, quindi, richiede l’utilizzo di metodologie di differenti campi: [sulle slide]

Ma l’informazione non è codificata solo nel contenuto delle pagine o nella struttura del sito Web, ma anche nei log di navigazione, ovvero un elenco di eventi generati dagli utenti durante la visita o l’interazione delle pagine del sito Web.

[cambia slide]

La pagina Web, elemento di un determinato sito, è formata da differenti rappresentazioni:

* **Rappresentazione Testuale**: le pagine Web sono composte dal testo, componente fondamentale della pagina stessa perché permette il trasferimento dell’informazione. Durante la navigazione, infatti, gli utenti estrapolano conoscenza semplicemente leggendo il contenuto testuale delle pagine che visita.
* **Rappresentazione Visuale**: quando una pagina Web viene renderizzata da un browser, viene applicato uno stile di formattazione visuale, chiamato CSS, che definisce gli elementi della pagina Web come contenitori rettangolari innestato o concatenati, formando un albero chiamato Rendered Box Tree. Quest’albero non deve essere confuso con gli elementi definiti dai tag Html, poiché ci sono dei tag invisibili <head> oppure elementi resi invisibili nel css. Inoltre, la generazione di tale albero richiede l’esecuzione di codice JS e CSS.
* **Rappresentazione Strutturale**: elementi della pagina Web inscritti nei tag Html, i quali possono essere testo, hyperlink o dati multimediali.

Tali rappresentazioni vengono sfruttate dagli attuali algoritmi di Clustering per poter organizzare l’enorme quantità di dati archiviati nel Web in maniera veloce e mirata.

[cambia slide]

Con il termine Clustering ci si riferisce al processo di selezione e raggruppamento, a partire da un set di dati iniziali, di elementi omogenei, avendo come base la somiglianza tra gli stessi. Questa attività non viene utilizzata solo nel contesto del Web, e più nello specifico nell’information Retrieval per suddividere le pagine in gruppi, ognuna delle quali cattura un aspetto particolare dell’elemento stesso. Infatti è stata usata in Biologia per creare una tassionomia di tutte le specie viventi; è stata usata per analizzare e classificare i climi della Terra; è stata usata in psicologia e medicina per identificare variazioni di malattie e depressioni; ed è stata usata nel campo del Business per dividere i consumatori in gruppi per successive attività di analisi e di marketing.

[cambia slide…]

L’obiettivo di questa tesi è quello di raggruppare le pagine appartenenti alla stessa tipologia semantica (per esempio pagine di professori, corsi e prodotti) di un sito Web. In particolare si è cercato di capire se, combinando le diverse rappresentazioni della pagina Web precedentemente viste, è possibile migliorare la qualità dei Cluster (dei raggruppamenti) prodotti. Non solo, si è cercato di capire se è possibile aumentare la bontà dei Cluster prodotti dando più importanza alle pagine vicine alla homepage.

[cambia slide…]

Gli algoritmi di Clustering esistenti in letteratura raggruppano pagine Web di differenti siti, trascurando quelle di uno specifico sito. Sono classificati in 4 categorie:

* **Contenuto testuale**: sfruttano il testo per estrarre informazioni utili al processo di raggruppamento delle pagine;
* **Web log**: sfruttano gli schemi di comportamento degli utenti durante la navigazione nel Web per raggruppare le pagine;
* **HTML**: le pagine Web, a differenza dei documenti testuali, sono caratterizzate dalle proprietà strutturali come i tag HTML, che permettono di definire la loro rappresentazione strutturale. Il codice HTML fornisce una differente e complementare rappresentazione rispetto a quella testuale.
* **Hyperlink**: anche la struttura ad hyperlink, che interconnettono le pagine Web, caratterizzano le loro proprietà strutturali. Gli hyperlink interconnettono due pagine Web, sfruttando l’informazione codificata negli stessi per Clusterizzare le pagine.

Ma tali algoritmi presentano delle limitazioni:

* **Contenuto testuale**: le pagine Web pur avendo lo stesso contenuto informativo, potrebbero essere contestualmente differenti, ovvero il contenuto informativo simile viene inserito in elementi Web aventi diverse regole semantiche (esempio: tabelle o menu di navigazione) e differenti funzionalità (link, pulsanti).
* **Web log**: la difficoltà sta nel fatto che ogni utente ha un comportamento differente nel Web, avendo come conseguenza una diversa profilazione per ogni utente, portando quindi alla creazione di differenti Cluster di pagine Web, in base ad ogni profilo.
* **HTML**: La qualità dei Cluster potrebbe abbassarsi se i tag sono differenti ma offrono una visualizzazione della pagina Web simile: i dati strutturati, per esempio, inseriti in tabelle o in liste (rispettivamente <table> e <ul>) hanno una visualizzazione simile, pur avendo una rappresentazione strutturale diversa.
* **Hyperlink**: se le pagine contengono un numero insufficiente di link, allora la qualità dei Cluster prodotti sarà bassa. Inoltre analizzare tutte le relazioni tra i nodi è computazionalmente costoso.

[cambia slide…]

Da queste motivazioni, si è definita come soluzione quella di combinare le informazioni estratte dal contenuto testuale e quelle della struttura del sito Web (grafo) e di rappresentare queste informazioni in uno spazio vettoriale.

[cambia slide…]

Per arrivare a questa soluzione, si è definita una metodologia di 3 fasi: Crawling del sito Web, Costruzione del Dataset e Clustering delle pagine Web.

[cambia slide…]

Il primo passo è quello di utilizzare un Crawler, ovvero un software automatizzato che permette l’indicizzazione delle pagine Web. L’indicizzazione è utile per aumentare l’efficienza delle interrogazioni che gli utenti effettuano durante la navigazione (ricerca su google).

L’algoritmo del Crawler è relativamente semplice: dato un insieme di URL di pagine Web, vengono scaricate tutte le pagine associate all’URL, estratti gli hyperlinks e, iterativamente, scaricate le pagine associate a questi link. Il loro contenuto verrà analizzato e salvato per essere successivamente indicizzato.

[cambia slide…]

Un sito Web può essere formalmente descritto come un grafo orientato, dove V è l’insieme dei nodi (url di pagine Web) ed E è l’insieme degli archi, ovvero una coppia di elementi (url) in cui il secondo url è contenuto nella pagina del primo url.

Nella sperimentazione di questa tesi sono state utilizzate due tecniche di Crawling: il tradizionale, che considera e indicizza tutte le pagine del sito Web e quello con il vincolo sulle Liste Web, dove una lista web è una collezione di due o più elementi Web che hanno una struttura HTML simile, sono visualmente adiacenti ed allineati. L’allineamento può essere visto lungo l’asse x, y, griglia. Questo crawler filtra le pagine da indicizzare usando proprio le liste web. Questa soluzione è stata utilizzata perché il grafo del sito può essere pieno di link rumore, come hyperlink scorciatoia, che non sono rilevanti nel processo di Clustering. In più la struttura del sito Web è codificata in sistemi di navigazione che offrono una visuale della sua organizzazione. Questi sistemi vengono implementati come collezioni di link che hanno lo stesso dominio. Nel Box A in figura sono stati esclusi i link all’interno della lista perché non appartengono al dominio del sito.

Il Grafo prodotto da entrambi i Crawler è un sottografo G’, i cui insiemi sono sottoinsiemi rispettivamente di V ed E.

Una volta Crawlato il sito Web, gli url vengono normalizzati per evitare che il processo di Crawling li analizzi più di una volta. Normalizzare significa modificare e standardizzare gli url, in modo tale da poter confrontare due url per definire se sono equivalenti, anche se sono sintatticamente differenti. Nell’esempio, anche se i due link di facebook sono sintatticamente differenti, sono equivalenti, poiché puntano allo stesso sito. Nella sperimentazione si è scelto di normalizzare gli url eliminando

* Protocollo (http:// e https://)
* www
* slash finale

[cambia slide…]

Una volta concluso il processo di Crawling, viene prodotto il grafo del sito Web e il contenuto testuale di ogni pagina del sito.

Dal grafo viene prodotto un dataset basato sulle rappresentazioni vettoriali della struttura, del testo e della struttura con il testo. Nel primo caso sono stati utilizzati Word2Vec con i RandomWalk standard, Word2Vec modificato con i Random Walk a partenza fissa (dalla homepage) e LINE. Nel secondo caso sono stati utilizzati tf-idf e doc2vec. Nell’ultimo caso la concatenazione dei vettori che contengono le informazioni della struttura con quelli testuali.

[cambia slide…]

I random walk sono delle passeggiate aleatorie, usate nella sperimentazione di questa tesi per l’esplorazione del grafo del sito Web in analisi. Vengono generate sequenze di URL, partendo da un nodo del grafo e seguendo ricorsivamente un hyperlink casuale fino ad un limite prefissato. Esistono due tipologie: quella standard, in cui ogni nuova sequenza parte da un nodo casuale, e quella a partenza fissa, in cui ogni nuova sequenza parte da un nodo specificato. Per la sperimentazione si è scelto di usare come nodo specificato la homepage del sito.

La scelta dei random walk è data dal fatto che, per esplorare un grafo, bisognerebbe esplorare tutte le possibili opzioni. Questo tipo di esplorazione è altamente complesso in termini di computazione. Le passeggiate aleatorie permettono di avere buone approssimazioni nell’esplorazione del grafo del sito. Analizzando la figura, però, ci si rende conto che potrebbe arrivare ad un nodo che non presenta ulteriori collegamenti (nel caso di F). Nel caso del Web, i nodi sono pagine che non hanno link al loro interno. La soluzione generalmente adottata è quella di effettuare un “salto” verso una qualsiasi altra pagina (nodo). In questa tesi, se si arriva ad una pagina senza link, l’attraversatore semplicemente si blocca, anche se non arriva al limite prefissato.

I random walk sono stati utilizzati per far apprendere a Word2Vec la struttura del grafo del sito Web.

[cambia slide…]

Word2Vec è un algoritmo di Word Embedding, ovvero una funzione parametrizzata che trasforma le parole in un vettore ad n dimensioni, come si evince dall’esempio. Questa rete neurale costruisce un vocabolario delle parole, in cui ogni termine ha un vettore. Word2Vec ha due modelli di apprendimento: CBOW e Skipgram. CBOW consiste nel predire una determinata parola a partire dal suo contesto, ovvero le parole che vengono prese in considerazione durante l’apprendimento, data una determinata parola. In questa tesi è stato utilizzato come modello di apprendimento Skip-Gram[, poiché permette di ottenere risultati più accurati]. skipgram consiste nel trovare rappresentazioni vettoriali delle parole per predire quelle circostanti in una frase. Formalmente, data una sequenza di parole w1…wt viene costruito un vocabolario, i cui termini hanno un vettore con n dimensione generato casualmente. L’obiettivo di skipgram è massimizzare la probabilità media logaritmica della parola del contesto, data la parola target. Supponiamo di avere la frase in esempio con dimensione della finestra di contesto 1, verrà costruito un dataset come in esempio e si cercherà di massimizzare la probabilità media logaritmica per ogni coppia di parole. Ad ogni passo vengono aggiornati i vettori delle parole precedentemente apprese. L’intero processo di apprendimento viene ripetuto sull’intera collezione di dati.

[cambia slide…]

Per raggiungere uno degli obiettivi di questa sperimentazione, si è deciso di modificare skipgram della libreria deeplearning4j, in maniera tale da considerare solo il contesto sinistro, ovvero le parole che si trovano alla sinistra di quella target. Dall’esempio si evince il funzionamento. Durante il lavoro di tesi, si è analizzato il valore b, che permette di aumentare o diminuire la finestra di contesto, dando più importanza alle parole più vicine a quella in analisi.

[cambia slide…]

LINE è un nuovo modello di apprendimento, capace di produrre rappresentazioni vettoriali dei vertici in una rete o grafo. Un grafo è una coppia G = (V, E) di insiemi, dove V insieme dei nodi ed E degli archi. L’obiettivo di LINE è quello di rappresentare ogni vertice v appartenente a v in un vettore Rd, dove d è molto più piccolo della dimensione dell’insieme V. Una particolarità di LINE è che permette di preservare la prossimità di primo e di secondo ordine. In entrambi si cerca di ottimizzare le funzioni di perdita riportate sulla slide. Per spiegare la differenza tra prossimità di primo e secondo ordine, analizziamo la figura: con prossimità di primo ordine si intende la somiglianza di due nodi, ovvero due nodi sono simili se sono vicini e collegati da un arco (altrimenti la prossimità è 0); con prossimità di secondo ordine, invece, si intende che due nodi sono simili se condividono lo stesso vicinato. La prossimità di primo ordine non è sufficiente per preservare la struttura del network, poiché può capitare che due nodi simili non siano direttamente collegati. Per questo è stata implementata la tipologia di secondo ordine.

La differenza con Word2Vec sta nel fatto che LINE permette di analizzare al più i vicini dei vicini di un dato nodo (relazioni aventi profondità 2); Word2Vec permette di analizzare relazioni più ampie.

[cambia slide…]

Tf-idf è una funzione che viene usata per misurare l’importanza di un termine rispetto ad un documento o ad una collezione di documenti. È composto da due fattori: tf e idf.

Tf misura quante volte un termine appare in un documento. Dato che ogni documento ha un differente numero di parole, è possibile che un termine possa apparire più volte in documenti più lunghi che in quelli più corti. Questo problema viene risolto dividendo la frequenza dei termini per la lunghezza del documento. La frequenza del termine è calcolata dalla formula in slide, dove nij è il numero di occorrenze del termine ti che si trova nel documenti dj, il denominatore è la dimensione del documento dj.

Idf misura i termini che si presentano più volte in un documento, ma con meno frequenza in tutta la collezione di documenti. Questo perché potrebbero esserci termini più significativi che appaiono raramente in un documento, ma frequentemente nella collezione dei documenti. La formula calcola proprio questo, dove |D| è il numero di documenti nella collezione, il denominatore è il numero di documenti che contengono il termine ti.

[cambia slide…]

Doc2Vec è una estensione di word2vec che aggrega tutte le parole di un paragrafo in un vettore. Il processo di apprendimento avviene in maniera simile a word2vec, ma limitando l’apprendimento delle parole ad ogni paragrafo. Vengono create due matrici: una dei paragrafi e una delle parole. Vengono usati gli stessi modelli di apprendimento di Word2Vec. Il vettore del paragrafo è condiviso per tutte le parole che si trovano nello stesso, ma non per altri paragrafi. I vettori dei paragrafi e delle parole vengono usati durante l’apprendimento e aggiornati ad ogni passo.

[cambia slide…]

L’ultimo passo della metodologia è il Clustering delle pagine apprese. Sono stati utilizzati algoritmi di clustering che sfruttano le rappresentazioni vettoriali, ovvero Kmeans e hdbscan. Sono stati scelti questi due algoritmi perché sono ben conosciuti ed hanno molte proprietà complementari. Infatti kmeans basa il processo di raggruppamento in base alla distanza dei centroidi del cluster (media di tutti i campioni); hdbscan si basa sulla densità del cluster. Kmeans richiede il numero di cluster da creare in input (etichette distinte in groundtruth); hdbscan richiede il numero minimo di elementi che un cluster deve avere per essere considerato come tale (5).

I vettori, prima di essere usati nel processo di Clustering, sono stati normalizzati con L2. Normalizzare significa ridurre l’effetto degli outlier, ovvero dei valori anomali e aberranti che potrebbero portare a risultati fuorvianti (esempio della temperatura di 10 oggetti in una stanza). Esistono vari metodi di normalizzazione, ma secondo diversi studi, la tecnica che ha permesso di ottenere risultati di normalizzazione superiori è stata la L2 per ogni vettore. L2 si basa sulla distanza euclidea [non so magari spiega la formula.. e di che una volta ottenuto il valore, si applica una divisione tra ogni valore del vettore e il valore ottenuto].

[cambia slide…]

Per la sperimentazione sono stati usati 4 siti del dipartimento di informatica, riportati in tabella in tabella. Sempre in tabella è possibile notare le dimensioni di ogni sito al termine del Crawling. La colonna numero pagine sono le pagine estratte sia dal crawler a liste di costrizione che quello tradizionale. Sono stati inoltre riportati sia il numero di archi crawlati con il processo tradizionale (seconda colonna), sia il numero degli archi crawlati con il processo a liste di costrizione (terza colonna). Nell’ultima colonna sono stati inseriti il numero dei cluster identificati manualmente dall’esperto durante la generazione della ground truth. La ground truth è una tabella di verità, dove sono state riportate le pagine che sono state clusterizzate manualmente dall’esperto.

[cambia slide…]

Per la sperimentazione sono state scelte le seguenti metriche per misurare la qualità dei cluster prodotti dagli algoritmi precedentemente citati.

* Con omogeneità si calcola se ogni cluster contiene elementi appartenenti alla stessa classe, ovvero della stessa tipologia.
* La completezza è simmetrica all’omogeneità, e calcola che tutti gli elementi di una stessa classe siano in uno stesso cluster.
* V-Measure è la media armonica tra omogeneità e completezza, ovvero 2 \* (h \* c) / (h + c).
* Adjusted Mutual Information calcola la corrispondenza delle due informazioni, ignorando le permutazioni.
* Adjusted Random Index calcola la percentuale di coppie di elementi per i quali la ground truth e l’algoritmo di Clustering concordano sull’assegnazione. In altre parole, misura l’accuratezza del processo di Clustering.
* Silhouette misura la forma di ogni Cluster, ovvero quanto un elemento di un Cluster è coeso con gli altri dello stesso raggruppamento e se è ben separato dagli altri insiemi.

La silhouette è l’unica metrica che non ha bisogno della groundtruth per restituire un valore, ed è l’unica metrica ad avere come range di valori possibili [-1; 1], a differenza delle altre metriche che hanno [0;1]. Più alto è il valore delle metriche, maggiore sarà la qualità dei cluster prodotti.

[cambia slide…]

Spiega lo schema… specificando che i random walk sono stati di 100k 500k e 1m con profondità massima 10 15 20 e w2v con finestra di 2 3 5 7. Di che una volta effettuato il clustering sulle informazioni della struttura e del testo, sono state scelte le configurazioni migliori per la concatenazione.

[cambia slide…]

Per motivi di spazio, sono stati riportati solo i risultati finali dei 4 siti web della sperimentazione. Le tabelle sono formate da due parti: nella prima parte si raccoglie le metriche migliori analizzando la struttura del sito (prima riga), quelle migliori analizzando il contenuto testuale (seconda riga) e quelle migliori della concatenazione delle informazioni. Le sigle nc ed lc rispettivamente sono senza costrizioni e a liste di costrizioni, ovvero il grafo utilizzato per ottenere le metriche riportate. Nella seconda parte, invece, si raccolgono le migliori configurazioni dei random walk analizzando la struttura (prima riga) e la concatenazione (seconda riga).

In generale, si può notare come il miglior algoritmo di clustering che riesce a sfruttare le informazioni estratte è **KMeans** . non solo, la versione tradizionale di skipgram permette di estrarre informazioni più utili per migliorare la bontà del processo di clustering: escludere il contesto destro causa un filtraggio di informazioni importanti. L’uso delle liste non migliora in maniera significativa le performance degli algoritmi di clustering. Evidentemente i siti web sono altamente strutturati, anche perché, se non fosse stato così, l’esperto non sarebbe riuscito ad assegnare manualmente le pagine ai vari cluster, creando la groundtruth.

La sperimentazione, in conclusione, ci mostra come i risultati migliori sono stati ottenuti andando a combinare l’informazione della struttura del sito con quella testuale delle pagine Web. In priceton, si evince come l’informazione testuale è soddisfacente per realizzare cluster di buona qualità: si suppone che le pagine del sito abbiano più testo che hyperlink ad altre pagine, oppure che la struttura del sito non fornisce unformazioni utili. In stanford, invece, l’informazione della struttura è sufficiente: si può supporre che il sito contenga pochi termini significativi oppure poco testo contenente informazioni utili.

[cambia slide…]

Concludendo, il testo e la struttura del sito forniscono informazioni diverse e complementari che aumentano effettivamente la qualità dei cluster. Le liste Web non hanno prodotto miglioramenti significativi. Neanche considerare più importanti le pagine più vicine alla homepage, utilizzando la versione modificata di skipgram, non ha prodotto miglioramenti significativi.

Uno sviluppo futuro può essere quello di applicare la metodologia descritta in questa tesi su più sito e meno strutturati, per capire se effettivamente le liste web sono utili a filtrare i link rumore oppure no.

[cambia slide…]

Grazie per l’attenzione…