EXTRACTIVE SUMMARIZATION CON LSA E TEXTRANK

Text Mining and Search - A.A. 2020/2021

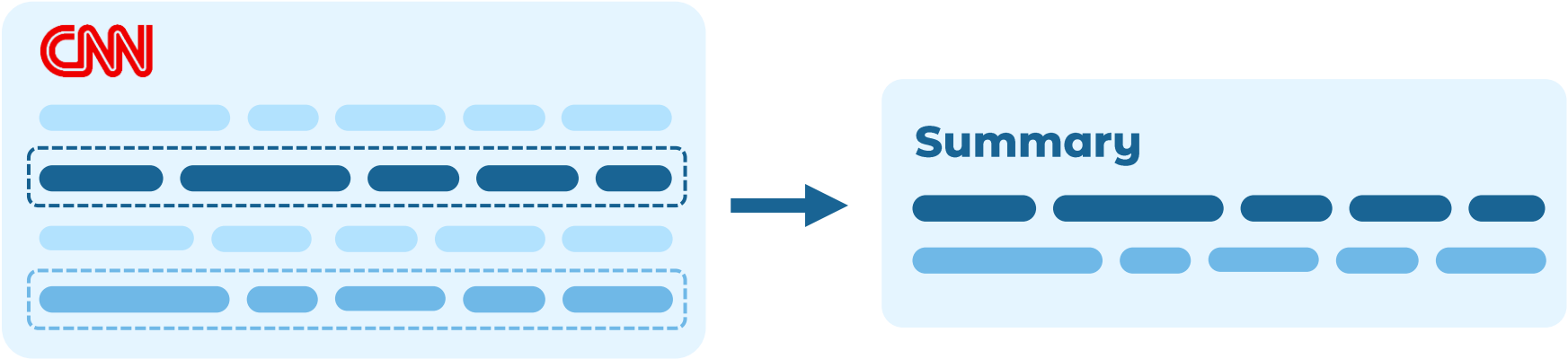
UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

MATTEO CAMPIRONI
matricola 801850

SERENA DI MAGGIO
matricola 821063

Obiettivo

Applicare tecniche di Text Summarization per generare riassunti di tipo estrattivo, utilizzando due diverse metodologie, LSA e TextRank.

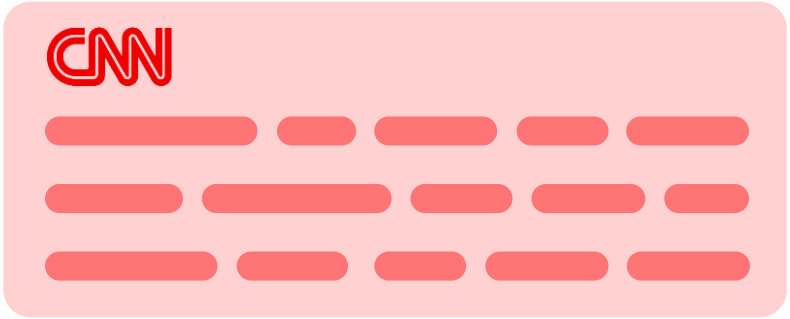


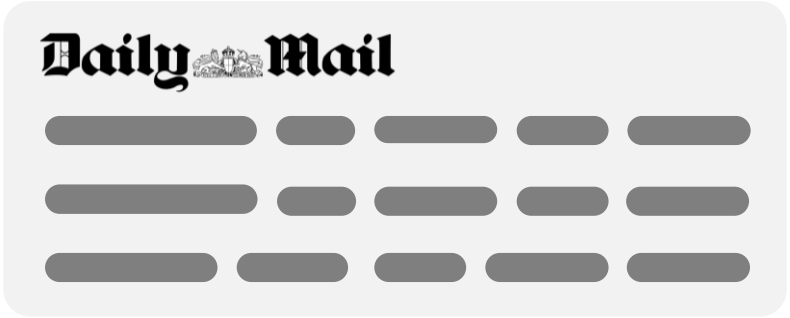
Dataset

Sono stati considerati articoli di giornale in lingua inglese estratti dal sito della CNN e del DailyMail riguardanti notizie degli ultimi anni.

Ad ogni articolo è assegnato un riassunto prodotto manualmente.

Dei 287.000 articoli del dataset originale, ne sono stati campionati casualmente solo 20.000 a causa di motivi computazionali.





Preprocessing

Per prima cosa sono state applicate diverse tecniche di preprocessing in modo da preparare i testi per gli step successivi:

- rimozione dell'autore e della data di pubblicazione
- conversione in lettere minuscole
- rimozione di numeri e punteggiatura
- sostituzione di contrazioni con la forma estesa
- eliminazione delle stop-words
- tokenizzazione
- POS tagging
- lemmatizzazione

~~By Shari Miller. PUBLISHED:. 05:45 EST, 22 November 2013. |. UPDATED:. 08:39 EST, 22 November 2013.~~ Some people can literally wait a lifetime before they find true love - and for one 83-year-old Canadian more than seven decades would pass before she married her childhood sweetheart.

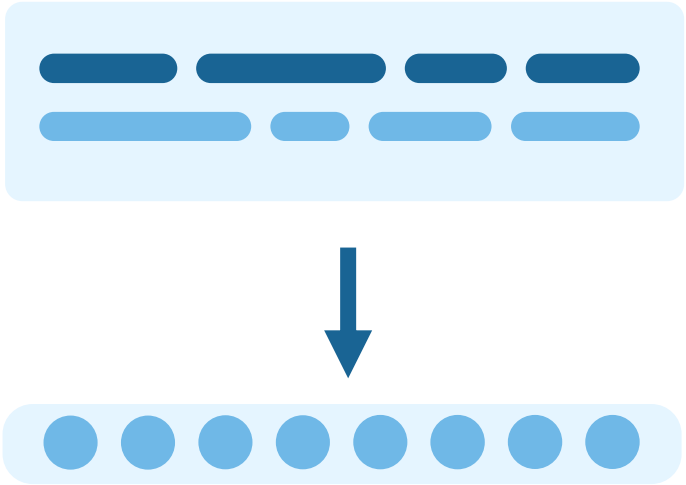
people literally wait lifetime find true love one year old canadian seven decade would pass marry childhood sweetheart.

Rappresentazione vettoriale

Sia LSA che TextRank necessitano che gli articoli vengano suddivisi in frasi e che queste siano rappresentate sotto forma di vettori.

Si è quindi deciso di usare la rappresentazione tramite TF-IDF, così definito:

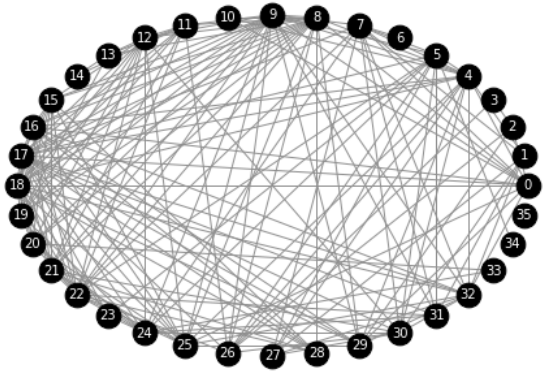
$$\text{TF-IDF} = \left(tf_{t,d}\right) \times \log_{10}\left(\frac{N}{df_t}\right)$$



TextRank

Appartenente agli Indicator Representation methods, questo modello prevede l'utilizzo di una matrice di similarità, costruita utilizzando la cosine similarity tra le frasi trasformate in vettori.

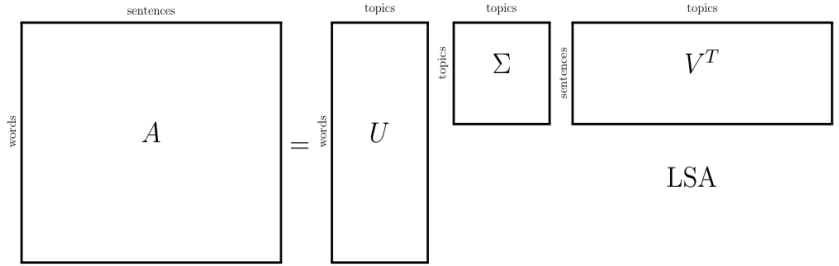
Questa matrice viene poi sfruttata per la realizzazione di uno grafo, i cui nodi rappresentano le frasi e gli archi la loro similitudine. Tramite l'algoritmo **PageRank** si assegna un punteggio ad ogni frase in base alla sua connessione con le altre.



$$\text{Similarity}(p,q) = \frac{p \cdot q}{\|p\|\|q\|} = \frac{\sum_{i=1}^{n} p_i q_i}{\sqrt{\sum_{i=1}^{n} p_i^{\,2}}\sqrt{\sum_{i=1}^{n} q_i^{\,2}}}$$

Latent Semantic Analysis

La Latent Semantic Analysis è una tecnica non supervisionata che permette di esprimere implicitamente il contenuto semantico degli articoli, basandosi sulla decomposizione ai valori singolari.



La matrice $D = \Sigma V^T$ unisce il peso dei topic e la rappresentazione della frase per indicare quanto questa descriva il topic, dove d_{ij} indica il peso del topic i nella frase j

Selezione delle frasi

Per la selezione delle frasi si è scelto di implementare una versione adattata a riassunti generici della Maximal Marginal Relevance:

$$\mathrm{MMR}(s_i) = \lambda\,\mathrm{sim}(s_i, D) - (1-\lambda)\max_{s_j \in D}(\mathrm{sim}(s_i, s_j))$$

In questo modo si cerca di massimizzare la rilevanza rispetto all'articolo e minimizzare la ridondanza tra le frasi da inserire nel riassunto. Fissando $\lambda = 0.5$, le 4 frasi con MMR maggiore sono state selezionate, costituendo dunque i riassunti finali.

Valutazione dei risultati

ROUGE-n sono un insieme di metriche di valutazione basata sulla presenza di elementi comuni nel riassunto ottenuto e nei modelli. In particolare, si è scelto di utilizzare il ROUGE-1 e il ROUGE-2, che considerano rispettivamente il numero di parole singole e di bigrams che sono presenti sia nel riassunto di riferimento che in quello prodotto.

$$\text{ROUGE-n} = \frac{p}{q}$$

p = numero di n-grams in comune tra riassunto ottenuto e quello di riferimento

q = numero di n-grams estratti solo dal riassunto di riferimento

Risultati

	Rouge-1	Rouge-2
LSA	**0.25**	**0.08**
TextRank	**0.26**	**0.09**

Entrambi i metodi producono risultati molto simili. I valori medi non sono molto elevati, ma questo dipende anche dalla natura stessa dei riassunti di riferimento che non sono stati ottenuti in maniera estrattiva, ma sono stati scritti dagli autori degli articoli.

Riassunto di riferimento

Prince was pictured holding the bear as he landed at Kensington Palace. Promised girl he would pass on bear to the Duke and Duchess of Cambridge. Was given the present during visit to brain injury charity in Nottingham.

Riassunto prodotto da LSA

A friend told the paper that Harry wants to get married and settle down with Cressida. True to his word: **Prince Harry pictured holding the blue bear he promised to give to his unborn niece or nephew**. Prince Harry proved he is an excellent uncle-to-be when he hand-delivered a teddy bear bought for his unborn niece or nephew. Baby's first present: The bear was a gift from a girl the Prince met during a visit to Nottingham.

Riassunto prodotto da TextRank

It comes as Prince Harry's romance with girlfriend Cressida Bonas is reportedly 'on the rocks' because she is not ready to marry. 'On the rocks': Prince Harry's romance with girlfriend Cressida Bonas is reportedly in trouble. A friend told the paper that Harry wants to get married and settle down with Cressida. True to his word: **Prince Harry pictured holding the blue bear he promised to give to his unborn niece or nephew.**

Conclusioni

- Nonostante a livello di risultati non ci siano grosse differenze, LSA ha richiesto uno sforzo computazionale maggiore, dovuto alla complessità dell'algoritmo di SVD.

- Sebbene i valori di ROUGE-1 e ROUGE-2 non siano particolarmente elevati, i riassunti risultano essere comunque apprezzabili.