UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

TEXT MINING & SEARCH

FINAL PROJECT

Text Summarization

il dono della sintesi ai tempi odierni

Teresa Cigna – 813925 - t.cigna@campus.unimib.it

Davide Garavaldi – 818308 – d.garavaldi@campus.unimib.it

Febbraio 2021

Abstract

Il seguente progetto si pone come obiettivo quello di creare riassunti di tipo estrattivo in maniera automatica, sia generici sia in relazione ad una query scelta da un utente. E' stato utilizzato un metodo di tipo *graph based*, impostando pesi *TF-IDF* per il calcolo della similarità e si è poi applicato l'algoritmo *TextRank* per l'estrazione delle frasi più importanti all'interno del testo e l'approccio *Maximal Marginal Relevance* per la scelta di quelle da inserire. Inoltre, è stata data particolare importanza alla leggibilità dei riassunti ottenuti.

1. Introduzione

Quante persone, dagli adolescenti agli adulti, dovendo leggere un testo lungo come un articolo o addirittura un libro, si sono chieste se ci fosse un modo rapido per capirne i contenuti senza leggerlo integralmente? Una volta, ottenere il riassunto di un testo presupponeva per forza la lettura di quest'ultimo, perché era impossibile avere il "*dono della sintesi*" prendendo come riferimento solamente alcune frasi. Ai giorni nostri, tuttavia, con la nascita del Text Mining e l'avvento delle moderne tecniche di Machine Learning, il "*dono della sintesi*" è diventato possibile, in particolare grazie all'utilizzo della tecnica della *text summarization* automatica: quest'ultima è uno dei task di NLP più affascinanti in quanto consente di estrapolare le informazioni più importanti da una fonte per produrne una versione ridotta a seconda del diverso utilizzo e utilizzatore. La *text summarization* può essere divisa in due tipologie:

- *Extractive*, dove il riassunto viene creato estraendo frasi già presenti nel testo *originale*;
- Abstractive, dove il riassunto creato esprime il contenuto del documento originale, ma attraverso parole differenti (più simile a come agirebbe un essere umano).

In questo scenario si colloca questo progetto, il cui obiettivo è quello di effettuare una *text summarization* di tipo estrattivo, utilizzando un metodo basato sui grafi. Inoltre, si è deciso di creare sia riassunti generici sia *query-based*, ossia generati sulla base di una specifica richiesta da parte dell'utente.

2. Dataset

I dati utilizzati per questo progetto sono stati presi da un dataset della CNN (scaricabile qui) contenente circa 92’500 articoli con relativi riassunti non ottenuti in maniera estrattiva. I testi riguardano news e trascrizioni di show televisivi.
Visto che la numerosità del dataset risultava essere molto alta e visto che l’approccio scelto non necessitava di una fase di training (e quindi sarebbe stato richiesto un elevato numero di dati), si è scelto di utilizzare i primi 12'000 articoli. La selezione è stata fatta in questo modo e non tramite un campionamento casuale, in quanto non è stato riscontrato un criterio specifico con cui erano ordinati i dati forniti.

3. Approccio metodologico

Si è scelto di applicare una *text summarization* di tipo estrattivo, ossia creare riassunti formati da frasi appartenenti al testo originale. Si tratta, quindi, di identificare un numero di frasi ritenute importanti per catturare l’essenza del testo in oggetto e, successivamente, selezionare tra queste le frasi da inserire all’interno del riassunto finale.

3.1 TextRank

A questo scopo si è deciso di utilizzare un approccio di tipo *graph-based*, in cui ogni nodo rappresenta una frase e gli archi che connettono due nodi indicano quanto le due frasi siano “simili” fra loro.
Nello specifico si è scelto di utilizzare *TextRank*, un algoritmo basato su *PageRank* utilizzato per l’estrazione di frasi importanti all’interno di un testo.

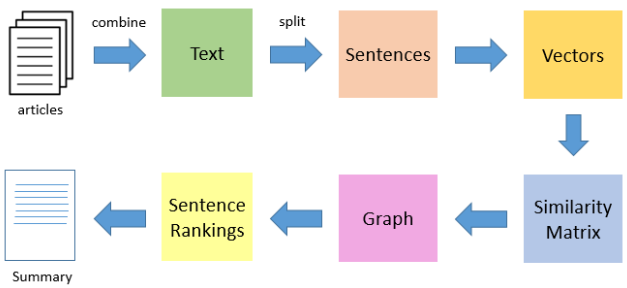

Figura 1. Step dell'algoritmo TextRank [1]

La procedura consiste nel dividere il testo in frasi, trasformare queste ultime in vettori (*e.g . word embedding, tf-idf, ecc*.), calcolarne la matrice di similarità e utilizzarla per costruire un grafo; successivamente, si assegna un punteggio ad ogni frase e vengono selezionati gli enunciati da inserire nel riassunto finale.

3.1.1 Preprocessing

Uno dei problemi della *text summarization* automatica riguarda la leggibilità dei riassunti. Si è scelto, quindi, di applicare due preprocessing distinti: in questo modo, da un lato sono stati processati più approfonditamente i testi utilizzati per il calcolo delle similarità e del ranking e, dall'altro, nei riassunti sono state inserite le frasi nella loro forma originale, quindi con un preprocessing minimo, in modo tale che il riassunto finale risultasse di facile comprensione.

- Preprocessing semplice: il preprocessing in questione ha riguardato la rimozione di link, la *tokenizzazione* e la sostituzione di numeri decimali con *'xxxx'*, in modo che il separatore decimale non venisse interpretato come un segno di interpunzione; inoltre, è stata effettuata la rimozione del punto all'interno degli acronimi (*e.g.* da *U.S.A.* a *USA*), la rimozione di simboli e la sostituzione di spazi bianchi consecutivi con un unico spazio bianco.

- Preprocessing approfondito: per la preparazione dei testi da utilizzare per il calcolo della matrice di similarità, oltre al preprocessing precedente, è stata effettuata la rimozione dei numeri, la modifica delle lettere in formato minuscolo, la lemmatizzazione (ossia la trasformazione delle parole nella loro forma base), la rimozione di *stopwords* inglesi (ovvero parole molto frequenti e con poco potere informativo) e la rimozione della punteggiatura.

Testo originale	Preprocessing semplice	Preprocessing approfondito
Burkhart, a 24-year-old German national, has been charged with 37 counts of arson following a string of 52 fires in Los Angeles.\n The charges are in connection with arson fires at 12 locations scattered through Hollywood, West Hollywood and Sherman Oaks, according to authorities.\n'	Burkhart , a 24 year old German national , has been charged with 37 counts of arson following a string of 52 fires in Los Angeles . The charges are in connection with arson fires at 12 locations scattered through Hollywood , West Hollywood and Sherman Oaks , according to authorities .	burkhart year old german national charged count arson following string fire los angeles . charge connection arson fire location scattered hollywood west hollywood sherman oak according authority .

Figura 2. Confronto tra i due tipi di preprocessing

3.1.2 Divisione del testo in frasi

Per la divisione in frasi si è utilizzata la funzione *sent_tokenize* che utilizza un algoritmo non supervisionato per la creazione di un modello in grado di capire le abbreviazioni delle parole, la loro collocazione all'interno del testo e le parole con cui solitamente inizia una frase. Successivamente questo modello viene utilizzato per delimitare le frasi. Perciò, visto il suo funzionamento, si è deciso di applicare questa funzione dopo il preprocessing superficiale, in modo da non eliminare parole e segni di punteggiatura potenzialmente utili al modello per capire la corretta divisione in frasi. Solamente dopo avere identificato le frasi si è applicato il preprocessing approfondito su ognuna di queste.

3.1.3 Vettori e similarità

Per calcolare la similarità tra le frasi si è scelto di utilizzare la *cosine similarity*, servendosi dei pesi *TF-IDF* per le parole. Il metodo *TF-IDF* è una tecnica per l'assegnazione dei pesi alle parole, atta a identificare l'importanza di queste ultime in un documento in base alla frequenza all'interno del documento stesso (*tf*) e al numero di documenti che le contengono (*df*). Viene assegnato un peso basso alle parole che appaiono in buona parte dei documenti (ritenute non discriminanti) e un peso alto a quelle che hanno un'alta frequenza all'interno del singolo documento ma una bassa presenza all'interno degli altri.

Una volta creati i vettori per ogni frase, aventi come elementi i pesi suddetti, si è utilizzata la *cosine similarity* per il calcolo della similarità tra le frasi.

$$\mathbf{tfidf}_{i,j} = \mathbf{tf}_{i,j} \times \log\left(\frac{\mathbf{N}}{\mathbf{df}_i}\right)$$

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

Figura 3. Formula del TF-iDF e della Cosine Similarity

3.1.4 Rappresentazione in grafo

Partendo dalla matrice di distanze, ogni testo è stato trasformato in un grafo, in cui ogni nodo rappresenta una frase e gli archi che connettono due nodi indicano di quanto le due frasi siano “simili” fra di loro.

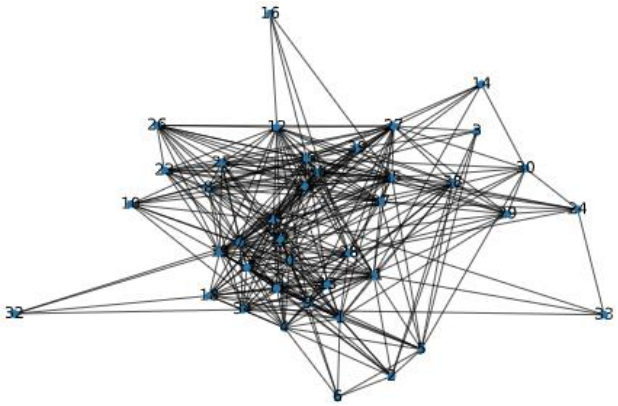

Figura 4. Esempio di testo di 40 frasi tradotto in grafo

3.1.5 Ranking delle frasi

In seguito, è’ stato assegnato uno score ad ogni frase in base all’algoritmo *PageRank*, che attribuisce un punteggio ad ogni nodo dando un peso maggiore a quelli più connessi all’interno del grafo ritenuti, quindi, importanti. Nel suo contesto originale (ranking di pagine web) questo punteggio riflette la probabilità di un utente di visitare la pagina in oggetto. Per i passi successivi si è scelto di tenere conto delle 10 frasi con un posizionamento più alto nel ranking.

3.2 Selezione delle frasi

Per selezionare le frasi da inserire all'interno delle sintesi si è scelto di utilizzare il criterio *Maximal Marginal Relevance (MMR)*. In questo approccio, i riassunti vengono formati una frase alla volta, cercando di massimizzare la rilevanza rispetto ad una query data dall'utente (o lo score di importanza nel caso in cui non ci sia una query specifica) e minimizzare la ridondanza, ovvero evitare di inserire frasi troppo simili a quelle già inserite.

- *Riassunti query-based*: la prima frase inserita all'interno del riassunto è stata quella con una similarità maggiore con la query definita dall'utente. Data la possibilità che quest'ultimo potesse inserire una parola non presente esattamente all'interno del testo, si è deciso di considerare come query non sono la parola impostata dall'utente, ma anche i suoi sinonimi, utilizzando *Wordnet* come risorsa lessicale.
- *Riassunti generici*: come prima frase da inserire è stata utilizzata quella con lo score di importanza più alto.

In entrambi i casi si è scelto di fermare l'algoritmo una volta raggiunto un numero di parole pari a 75.

3.3 Leggibilità

Sulla base di ciò che è stato precedentemente detto riguardo la leggibilità del riassunto finale, le similarità sono state calcolate sulle frasi appartenenti ai testi preprocessati in maniera approfondita, ma per la formazione del riassunto sono state poi prese in considerazione le corrispondenti frasi ottenute tramite il preprocessing semplice. Inoltre, utilizzando il metodo *MMR*, le frasi risultano ordinate in base a come sono state aggiunte all'interno del riassunto, pertanto non necessariamente seguendo l'ordine di apparizione nel testo originale; perciò, una volta ottenuto il riassunto, si è deciso di riordinarlo in modo tale da mantenere la cronologia di apparizione delle frasi per una lettura più lineare e una comprensione più semplice della sintesi.

3.4 Valutazione

Dato che la valutazione dei riassunti prodotti richiedeva il confronto con quelli forniti che non sono stati ottenuti in maniera estrattiva, si è scelto di utilizzare l'indicatore *ROUGE-n*; questo conta il numero di *n-grammi* condivisi tra il riassunto prodotto e quello di riferimento e lo rapporta al numero di *n-grammi* contenuti nel solo riassunto di riferimento, in modo tale da indicare la misura in cui quest'ultimo è contenuto nel riassunto prodotto. Nello specifico si è scelto di utilizzare il *ROUGE-1*, in cui vengono considerate le parole singole. È stata presa questa decisione in quanto, dovendo confrontare due tipologie di riassunti molto diverse tra loro, risulta meno probabile la corrispondenza di più parole consecutive. Per evitare che il valore venisse aumentato da termini poco significativi come le *stopwords*, è stata applicata la rimozione di queste ultime, oltre alla lemmatizzazione, sia sul riassunto prodotto sia su quello di riferimento. Inoltre, per far sì che le corrispondenze di contenuto venissero catturate da questo indicatore anche in caso di parole diverse che esprimessero lo stesso concetto, si è utilizzata la risorsa lessicale *WordNet*, in modo tale da considerare non solo le parole contenute nel riassunto prodotto, ma anche i loro sinonimi. A questo scopo si è prima applicato un *Part-Of-Speech Tagging*, per identificare solo nomi, aggettivi e verbi e, successivamente, si sono aggiunti i sinonimi di questi. Questa operazione è stata fatta unicamente sul riassunto prodotto poiché, se fosse stata applicata anche su quello di riferimento, il numero di parole al denominatore sarebbe aumentato, falsando i risultati di conseguenza.

4. Risultati

Come risultato, è stato ottenuto un *ROUGE-1* medio di 0.48. Si è notato che i valori più bassi si hanno in caso di articoli formati da elenchi (ad esempio liste di hotel da visitare, di videogame ecc…), per i quali risultano poco adatti i riassunti di tipo estrattivo. In tutti gli altri casi, invece, si è notata una buona capacità di sintesi, arrivando ad ottenere valori di *ROUGE-1* anche superiori a 0.9.

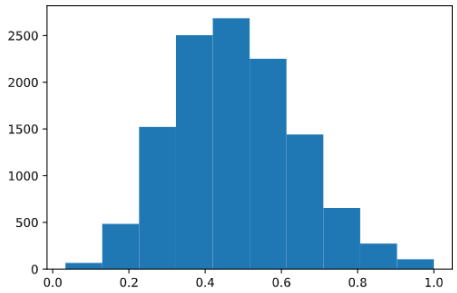

Figura 5. Istogramma dei valori di Rouge-1 ottenuti

Inoltre, anche nei casi in cui non si sono riscontrati alti valori di *ROUGE-1*, si è visto che i riassunti riescono ugualmente a cogliere il contenuto dell'articolo originale.

Riassunto fornito	Riassunto prodotto
NEW: A Canadian doctor says she was part of a team examining Harry Burkhart in 2010 . NEW: Diagnosis: "autism, severe anxiety, post-traumatic stress disorder and depression" . Burkhart is also suspected in a German arson probe, officials say. Prosecutors believe the German national set a string of fires in Los Angeles	'Burkhart, a 24 year old German national, has been charged with 37 counts of arson following a string of 52 fires in Los Angeles. Stancheva said the refugee applications by Burkhart and his mother were denied by the Canadian government, and she has not seen Burkhart since early March of 2010. The worst arson sprees in the citys history began last Friday morning with a car fire in Hollywood that spread to apartments above a garage, but no new fires have happened since Burkhart was arrested Monday, Los Angeles District Attorney Steve Cooley said.'

Figura 6. Confronto tra riassunto fornito e riassunto prodotto.
Il testo originale è consultabile qui

Per quanto riguarda i riassunti di tipo *query-based*, non avendo dati a disposizione per la valutazione, non è stato possibile dare un giudizio a livello analitico; tuttavia, si riportano alcuni esempi dei riassunti ottenuti, in modo da poter mostrarne i risultati.

Riassunto query 'arson'	Riassunto query 'mental disorder'
Los Angeles CNN A medical doctor in Vancouver, British Columbia, said Thursday that California arson suspect Harry Burkhart suffered from severe mental illness in 2010, when she examined him as part of a team of doctors. Burkhart, a 24 year old German national, has been charged with 37 counts of arson following a string of 52 fires in Los Angeles. It was revealed that Burkhart is also under investigation for arson and fraud in relation to a fire in Neukirchen, near Frankfurt, Germany. The worst arson sprees in the citys history began last Friday morning with a car fire in Hollywood that spread to apartments above a garage, but no new fires have happened since Burkhart was arrested Monday, Los Angeles District Attorney Steve Cooley said.	Los Angeles CNN A medical doctor in Vancouver, British Columbia, said Thursday that California arson suspect Harry Burkhart suffered from severe mental illness in 2010, when she examined him as part of a team of doctors. Stancheva said she and other doctors including a psychiatrist diagnosed Burkhart with autism, severe anxiety, post traumatic stress disorder and depression. Cooley called it almost attempted murder, because people were sleeping in apartments above where Burkhart allegedly set cars on fire with incendiary devices placed under their engines. Dorothee Burkhart was arrested a day before on an international arrest warrant issued by a district court in Frankfurt, Germany, said federal court spokesman Gunther Meilinger.'

Figura 7. Confronto tra riassunto ottenuto tramite query 'arson' e tramite query 'mental disorder'. Il testo originale è consultabile qui

5. Conclusioni e sviluppi futuri

In conclusione, si può affermare che l'approccio utilizzato risulta molto valido nel produrre riassunti efficaci e leggibili, ottenendo non solo un Rouge-1 medio pari a 0.48, ma anche valori superiori a 0.9 in alcuni articoli. Ad ogni modo, è opportuno notare come il compito della sintesi risulti essere un task estremamente soggettivo, quindi rende difficile una valutazione analitica precisa. Possibili sviluppi potrebbero riguardare l'identificazione di parole chiave tramite *TextRank* o di entità tramite *NER* per poter presentare ad un possibile utente un insieme di parole da poter utilizzare come query per ottenere dei riassunti *query-based* ancor più efficaci..

Bibliografia

An Introduction to Text Summarization using the TextRank Algorithm (with Python implementation) - Prateek Joshi
https://www.analyticsvidhya.com/blog/2018/11/introduction-text-summarization-textrank-python/

The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries – Jaime Carbonell, Jade Goldstein
https://www.cs.cmu.edu/~jgc/publication/The_Use_MMR_Diversity_Based_LTMIR_1998.pdf

Understand TextRank for Keyword Extraction by Python - Xu LIANG
https://towardsdatascience.com/textrank-for-keyword-extraction-by-python-c0bae21bcec0

NLP — Sentence Extraction using NLTK: TextRank Algorithm - Akash Panchal
https://medium.com/analytics-vidhya/sentence-extraction-using-textrank-algorithm-7f5c8fd568cd

Use TextRank to Extract Most Important Sentences in Article – Ceshine Lee
https://medium.com/the-artificial-impostor/use-textrank-to-extract-most-important-sentences-in-article-b8efc7e70b4

Keyword and Sentence Extraction with TextRank (pytextrank) – David Ten
https://xang1234.github.io/textrank/

Text Preprocessing in Python: Steps, Tools, and Examples -
https://medium.com/@datamonsters/text-preprocessing-in-python-steps-tools-and-examples-bf025f872908

Summarization of Icelandic Texts - Karin Christiansen
https://www.ru.is/faculty/hrafn/students/msc_karin_sumoficetext_paper.pdf

TF-IDF and similarity scores - Chanseok Kang
https://goodboychan.github.io/chans_jupyter/python/datacamp/natural_language_processing/2020/07/17/04-TF-IDF-and-similarity-scores.html