Text Mining & Search
2020-2021

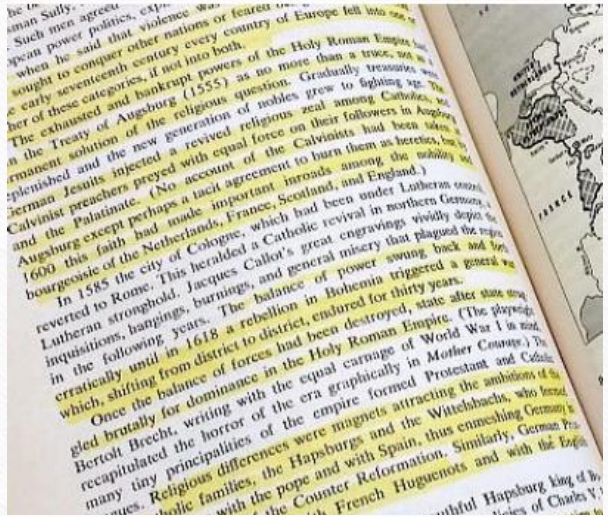
Text Summarization:
il dono della sintesi ai tempi odierni

Teresa Cigna – 813925

Davide Garavaldi – 818308

OBIETTIVO

- Ottenere riassunti di tipo estrattivo sia generici che query-based, con un alto grado di leggibilità.



DATI

~ 92’500 articoli della CNN con relativi riassunti non di tipo estrattivo

↓

12’000

Testo

(CNN) -- The 54 men and 14 boys rescued after being found chained this week at an Islamic religious school in Pakistan have been reunited with their families or placed in shelters, authorities said.The group was discovered in an underground room with heavy chains linking them together.The school, Al-Arabiya Aloom Jamia Masjid Zikirya, which also was a drug rehab clinic, is in Sohrab Goth, a suburb of Gadap in Karachi.All 14 boys were returned to their families, senior police official Ahsanullah Marwat told CNN.Of the adults, 47 had been released to their families, and seven were handed over to a shelter for the homeless, he said.Three people who worked at the facility were arrested, but the four men who ran the place were still at large, Marwat said.Officials said the facility was part madrassa and part drug-rehab facility, and the captives were chained at night apparently to prevent their escape."The operation was successful, and we plan on continuing our work to ensure that places like this are shut down," Marwat said.Many of the captives told police their families sent them there because they were recovering drug addicts. During the day, they worked and did religious studies.But the future of the rescued children was unclear.One woman told a local television station that she was willing to pay the police to keep her troublesome child. She said she would rather have the facility remain open, regardless of how it treated the children.Many others, however, said they were in shock and disbelief over the allegations.One man complained he was deep in debt after paying the school a large amount of money to board his son.

Riassunto di riferimento

@highlight
Captive boys and men were rescued from an Islamic religious school in Pakistan
@highlight
They were reunited with their families this week
@highlight
The facility was a school and drug rehab clinic
@highlight
Authorities say they're searching for the owners; three others arrested at the facility

PRE-PROCESSING

Preprocessing semplice:

- rimozione di link
- tokenizzazione
- sostituzione di numeri decimali con 'xxxx'
- rimozione del punto all'interno di acronimi (*e.g.* da *U.S.A.* a *USA*)
- rimozione di simboli
- sostituzione di spazi bianchi consecutivi con un unico spazio bianco

Semplice

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 .

Preprocessing approfondito:

- preprocessing semplice +
- rimozione di numeri
- trasformazione delle lettere in lettere minuscole
- lemmatizzazione
- rimozione di stopwords inglesi

Approfondito

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 .

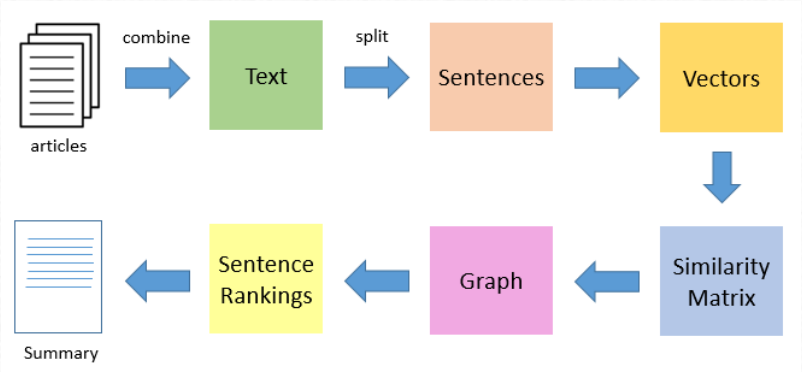
LEGGIBILITA'

- Calcoli su testo con preprocessing approfondito, ma inserimento delle corrispettive frasi con minimo preprocessing

- Re-ranking delle frasi inserite nel riassunto, secondo l'ordine di apparizione nel testo originale

TEXT-RANK

- Dividere il testo in frasi
- Trasformarle in vettori (e.g. word embedding, tf-idf, ecc.)
- Calcolarne la matrice di similarità
- Utilizzarla per costruire un grafo
- Assegnare un punteggio ad ogni frase
- Selezionare le frasi da inserire nel riassunto finale



TEXT-RANK - 1

- Divisione del testo in frasi

Testo → preprocessing semplice → *sent_tokenize* → preprocessing approfondito

- Trasformarle in vettori

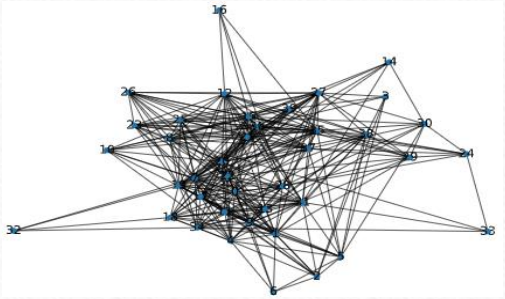
$$\mathbf{tfidf}_{i,j} = \mathbf{tf}_{i,j} \times \log\left(\frac{\mathbf{N}}{\mathbf{df}_i}\right)$$

- Calcolarne la matrice di similarità

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

TEXT-RANK - 2

- Costruire il grafo
  - Nodi → Frasi
  - Archi → Similarità



- Assegnare un punteggio ad ogni frase
  - Utilizzo dell'algoritmo PageRank
  - Selezione delle 10 frasi con score più alto

$$S(V_i) = (1-d) + d * \sum_{j \in In(v_i)} \frac{1}{|Out(V_j)|} S(V_j)$$

SELEZIONE DELLE FRASI

Maximal Marginal Relevance (MMR)

- Massimizza …
  - query-based → … rilevanza rispetto alla query
  - generico → … score di importanza
- Minimizza ridondanza tra le frasi nel riassunto

PRIMA FRASE

- Query-based → Frase più simile a query (e suoi sinonimi)
- Generico → Frase con score di importanza più alto

LEGGIBILITA'

- Calcoli su testo con preprocessing approfondito, ma inserimento delle corrispettive frasi con minimo preprocessing

- Re-ranking delle frasi inserite nel riassunto, secondo l'ordine di apparizione nel testo originale

VALUTAZIONE

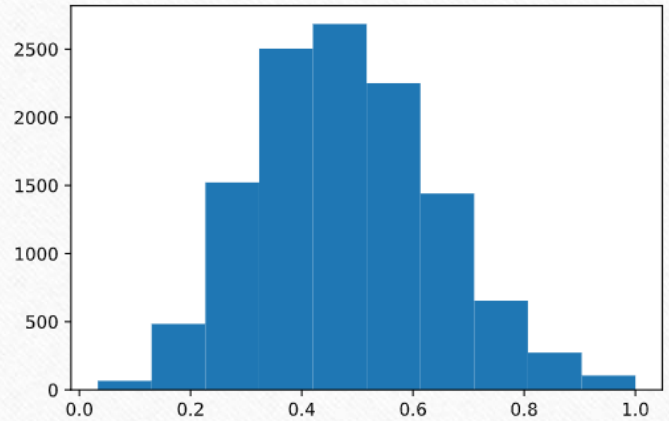
$$\text{ROUGE-1} = \frac{\text{n° parole comuni}}{\text{n° parole del riassunto di riferimento}}$$

Riassunto ottenuto → • Rimozione stopwords • Lemmatizzazione → POS tagging → Aggiunta di sinonimi solo per sostantivi, verbi e aggettivi

Riassunto fornito → • Rimozione stopwords • Lemmatizzazione

RISULTATI – RIASSUNTI GENERICI

- ROUGE-1 medio = 0.48



Istogramma dei valori di Rouge-1 ottenuti

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.

RISULTATI – RIASSUNTI QUERY-BASED

Riassunto query 'arson'	Riassunto query 'mental disorder'
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 .

RISULTATI – RIASSUNTI QUERY-BASED

Text summarization is the technique for generating a concise and precise summary of voluminous **texts** while focusing on the sections that convey useful information, and without losing the overall meaning. 15 apr 2019

https://blog.floydhub.com › gentle-introduction-to-text-su...

A Gentle Introduction to Text Summarization in Machine ...

> Have you ever summarized a lengthy document into a short paragraph? How long did you take? Manually generating a summary can be time consuming and tedious. Automatic text summarization promises to overcome such difficulties and allow you to generate the key ideas in a piece of writing easily.

Text summarization is the technique for generating a concise and precise summary of voluminous texts while focusing on the sections that convey useful information, and without losing the overall meaning.
Automatic text summarization aims to transform lengthy documents into shortened versions, something which could be difficult and costly to undertake if done manually. Machine learning algorithms can be trained to comprehend documents and identify the sections that convey important facts and information before producing the required summarized texts. For example, the image below is of this news article that has been fed into a machine learning algorithm to generate a summary.

Generico

```
unsorted_summ,sorted_summ = summarize(T, 10)
```

```
sorted_summ
```

```
['Machine learning algorithms can be trained to comprehend documents and identify the sections that con
vey important facts and information before producing the required summarized texts .']
```

Query- based

```
unsorted_summ,sorted_summ = summarize(T, 10, 'text summarization')
```

```
sorted_summ
```

```
['Text summarization is the technique for generating a concise and precise summary of voluminous texts
while focusing on the sections that convey useful information , and without losing the overall meaning
.']
```

CONCLUSIONI

- Approccio valido sia per risultati che per leggibilità
- Task soggettivo

SVILUPPI FUTURI

- Parole chiave con TextRank o estrazione di entità tramite NER per dare all'utente un insieme di parole da utilizzare come query