Relazione esercizio automatic Summarization

L'esercizio 3 ha visto la definizione di un algoritmo di automatic summarization usando la risorsa vettoriale NASARI. L'algoritmo riceve in input un documento di testo e applicando un approccio shallow tenta di produrre in output un riassunto estrattivo riutilizzando porzioni di frasi e di paragrafi del documento originale. In particolare, nel testo riassunto troveremo un sottoinsieme dei paragrafi presenti nel testo iniziale. Il tasso di compressione, che esprime il rapporto tra la lunghezza dei 2 documenti, segna la discrepanza fra i due testi.

Fondamentale per l'implementazione di un algoritmo di automatic summarization è l'applicazione di criteri necessari per l'individuazione di un topic, ovvero la porzione del testo del documento originale nella quale risiedono le frasi e i termini più rilevanti. L'idea alla base di questo task è che, una volta capito dove risiedono le porzioni di testo più importanti, è possibile selezionare tali porzioni e restituirle in output come sintesi del documento. Nel caso di questa esercitazione si è deciso di implementare 2 criteri per l'individuazione dell'argomento:

- **titolo**: i termini più importanti sono le parole di contenuto presenti nel titolo, ovvero nella prima riga utile del documento.
- **cue phrases**: i termini rilevanti si trovano nei paragrafi con il più alto valore di score, calcolato in termini di maggior numero di bonus words e il minor numero di stigma words.

Il primo metodo è più impreciso rispetto al secondo, in quanto non tutti i testi iniziano necessariamente con un titolo, inoltre in alcuni casi il titolo potrebbe risultare poco informativo e troppo vago. Per tale ragione, è stato deciso di implementare un secondo approccio più preciso tenendo il metodo del titolo come baseline per fare confronti.

Tra i file di progetto, nella cartella util, sono presenti 2 documenti di testo:

- **bonus_words.txt**
- **stigma_words.txt**

Questi file contengono una lista di parole bonus e stigma, fra le più comuni presenti nel lessico inglese. Esempi di **stigma words** sono negazioni e pronomi come "I, you, he... mine, yours, hers... Invece fra le **bonus words** sono presenti superlativi e comparativi come ad esempio: "better, worse, best, worst..."

I termini più importanti vengono usati per valutare la coesione tra i paragrafi e il contesto di riferimento. Quest'ultimo viene costruito a partire dalla rappresentazione vettoriale, tramite i vettori NASARI, del significato dei termini rilevanti nel topic. Il topic è stato preprocessato rimuovendo simboli di punteggiatura, spazi multipli, stopwords e lemmatizzando i termini restanti, ottenendo un set di parole di contenuto. Le informazioni sui vettori NASARI sono recuperate dal file: "**dd-small-nasari-15.txt**" nella cartella "**util**", presente tra i file del progetto. I vettori del contesto vengono confrontati con i vettori costruiti a partire dalle parole di contenuto presenti nei

paragrafi preprocessati del documento, usando la metrica di similarità **weighted overlap**. Ovviamene il confronto tiene conto dei possibili termini polisemici, quindi, considerando una lista di vettori per ogni parola, la similarità tra 2 termini è calcolata come il massimo della weighted overlap per ogni coppia di vettori.

L'operazione di confronto avviene per ogni paragrafo. A termine di tutti i confronti utili tra coppie di vettori, le similarità trovate vengono sommate e divise per il numero di overlap trovati, ottenendo un valore di score che viene associato al paragrafo. Una volta calcolate tutte le similarità con il contesto, i paragrafi vengono ordinati per valore di score decrescente di modo che quelli con il punteggio più alto saranno considerati come paragrafi del documento sintetizzato. L'estrazione delle frasi per il riassunto avviene rispettando il loro ordinamento originale. Il fattore di compressione entra in gioco in questo istante. In base al suo valore vengono selezionati un sottoinsieme di **N – (N * F%)** paragrafi. Con **N = numero di paragrafi** e **F% = tasso di compressione**.

Nella cartella del progetto è presente il file **main.py** che contiene l'entry point dell'algoritmo. Il programma all'atto dell'esecuzione riceve in input diversi parametri:

- il documento da sintetizzare
- il fattore di compressione. Sono possibili 3 valori: 10%, 20%, 30%
- il metodo per la ricerca del topic. Sono ammessi 2 valori: "**title" e "cue"**

Seguendo il flusso di esecuzione del main, il documento viene letto riga per riga rimuovendo spazi multipli e simboli "\n". Successivamente viene recuperato l'insieme dei vettori NASARI, fondamentali per l'individuazione dei paragrafi rilevanti. Avendo a disposizione tutte le informazioni, il programma esegue la funzione **"automatic_summarization"** procedendo nei confronti tra topic e contesto come descritto in precedenza. A termine della procedura, viene restituito il documento sintetizzato.

Per valutare la qualità del riassunto sono state utilizzate 2 metriche:

- **BLEU** : |{relevant_document} & {retrieved_document}| / |{retrieved_document}|
- **ROUGE:** |{relevant_document} & {retrieved_document}| / |{relevant_document}|

BLEU valuta la precision mentre ROUGE misura la Recall. Le 2 metriche considerano l'intersezione tra i termini rilevanti estratti dal documento originale e le parole di contenuto presenti nel riassunto. Per individuare le parole più importanti presenti in tutto il documento si è usato un approccio semplicistico, andando a considerare tutte le keyword nel titolo e nelle conclusioni. Si nota che nei documenti in cui manca il titolo viene considerata l'introduzione. Le keyword vengono usate per puntare a frasi presenti nel testo. Il risultato finale sarà un bag of words delle parole di tutte le frasi puntate dalle keywords. L'intersezione con il bag of words del documento riassunto calcola il numeratore di entrambe le metriche. Il denominatore per BLEU sarà il numero di parole di contenuto nella sintesi e per ROUGE il numero di parole rilevanti del documento originale.

A titolo di esempio vengono riportati i risultati dell'automatic summarization eseguito sul documento **"Napoleon-wiki.txt**" ad un fattore di compressione del **30%** considerando entrambi gli approcci. A termine dell'esecuzione tali informazioni vengono salvate in file di testo sotto la cartella "**output**".

output > 30_Napoleon-wiki_cueTopic.txt

```
1  ----------------------------------------------------------------------
2  bilingual evaluation understudy (BLEU):  88.0%
3  Recall-Oriented Understudy for Gisting Evaluation (ROUGE):  69.0%
4  ----------------------------------------------------------------------
5
6  Napoleone Bonaparte.
7  Napoleon Bonaparte (born Napoleone di Buonaparte) was a French statesman and military leade
8  He was Emperor of the French as Napoleon I from 1804 until 1814 and again briefly in 1815 d
9  He won most of these wars and the vast majority of his battles, building a large empire tha
10 Napoleon's political and cultural legacy has endured as one of the most celebrated and cont
11 He was born in Corsica to a relatively modest Italian family from minor nobility. He was se
12 The French Directory eventually gave him command of the Army of Italy after he suppressed t
13 At age 26, he began his first military campaign against the Austrians and the Italian monar
14 In 1798, he led a military expedition to Egypt that served as a springboard to political po
15 Intractable differences with the British meant that the French were facing a Third Coalitio
16 In 1806, the Fourth Coalition took up arms against him because Prussia became worried about
17 Napoleon's influence on the modern world brought liberal reforms to the numerous territorie
18 British historian Andrew Roberts states: "The ideas that underpin our modern world—meritocr
```

output > 30_Napoleon-wiki_titleTopic.txt

```
1  ----------------------------------------------------------------------
2  bilingual evaluation understudy (BLEU):  86.0%
3  Recall-Oriented Understudy for Gisting Evaluation (ROUGE):  61.0%
4  ----------------------------------------------------------------------
5
6  Napoleone Bonaparte.
7  Napoleon Bonaparte (born Napoleone di Buonaparte) was a French statesman and military leade
8  He was Emperor of the French as Napoleon I from 1804 until 1814 and again briefly in 1815 d
9  He won most of these wars and the vast majority of his battles, building a large empire tha
10 Napoleon's political and cultural legacy has endured as one of the most celebrated and cont
11 He was born in Corsica to a relatively modest Italian family from minor nobility. He was se
12 The French Directory eventually gave him command of the Army of Italy after he suppressed t
13 At age 26, he began his first military campaign against the Austrians and the Italian monar
14 In 1798, he led a military expedition to Egypt that served as a springboard to political po
15 Intractable differences with the British meant that the French were facing a Third Coalitio
16 In 1806, the Fourth Coalition took up arms against him because Prussia became worried about
17 In 1809, the Austrians and the British challenged the French again during the War of the Fi
18 Napoleon then occupied the Iberian Peninsula, hoping to extend the Continental System and c
```

Le immagini hanno lo scopo di mettere in evidenza le differenze tra i due approcci, come tale non vengono riportate le frasi per intero ma solo quello che basta per capire cosa cambia. Il metodo cue è più preciso rispetto al metodo del titolo. Mentre BLEU è pressappoco simile, ROUGE segna una maggiore differenza. Questo è dato dal fatto che con la tecnica cue si individuano più intersezioni con il documento sintetizzato, segno che nel riassunto il rapporto tra il numero di parole rilevanti trovate e il totale delle parole importanti è più alto. La differenza di contenuto dei paragrafi è visibile leggendo le ultime 2 righe: 17 e 18.