Relazione esercizio Segmentation

Il quarto esercizio ha visto la definizione di un algoritmo di Document Segmentation, vagamente ispirato all'implementazione del text-Tiling. Il document segmentation è uno dei task che abbiamo visto collocato nell'ambito del text mining. L'obiettivo è, a partire da un testo, individuare i punti di cambio del discorso, quindi, in direzione di questi punti, segmentare il documento e organizzare i suoi paragrafi tenendo conto del cambio di contesto.

L'euristica utilizzata per individuare correttamente i punti di taglio, ai fini di questa esercitazione è stata la seguente: "*siamo in corrispondenza di un cambio di discorso quando il lessico utilizzato muta notevolmente passando da un paragrafo all'altro del documento*". L'entry point del programma è contenuto nel file "**main.py**", il quale riceve da riga di comando 2 argomenti:

- Il path al file da segmentare
- La dimensione in termini di numero di frasi dei segmenti iniziali

Inizialmente, il programma organizza le frasi del documento in blocchi di una certa misura data in input, in seguito per ognuno di essi di esegue un bag of words e si recuperano per ciascun blocco l'insieme delle parole di contenuto. Per stimare il cambio di discorso è stata calcolata la similarità tra blocchi come **sovrapposizione lessicale** normalizzata alla dimensione del bag of words del blocco più piccolo. Quest'ultima operazione è gestita dalla funzione "**inter_block_similarity**" presente nel modulo "**segmentation**". I blocchi di frasi vengono analizzati in sequenza e le similarità vengono calcolate per ogni coppia di blocchi **Bi – Bi+1** (per ogni i da 1 a n blocchi).

Il risultato è uno score che segna di quanto il lessico dei due blocchi è simile. Se ci troviamo in corrispondenza di un valore di similarità alto vuol dire che i due blocchi molto probabilmente parlano della stessa cosa quindi è giusto che siano uniti in un solo segmento di frasi. Se, invece, la similarità è bassa allora potremmo essere di fronte ad un cambio di discorso. Per accertarci di ciò, le similarità vengono raccolte in una lista che in seguito viene analizzata in modo da individuare i punti di **minimo locale**, ovvero quei valori di similarità bassi circondati da valori alti. In questo caso vuol dire che abbiamo un cambio di discorso perché, ad esempio se abbiamo 4 blocchi B1, B2, B3, B4, tali che la similarità a coppie è espressa come:

- B1 – B2 : **0.5**
- B2 – B3 : **0.3**
- B3 – B4: **0.6**

Possiamo asserire che nel passaggio da B2 a B3 c'è un cambio di discorso semplicemente perchè abbiamo un cambio di uso di vocabolario, quindi da B3 potremmo cominciare a parlare di altre cose. Salvando il riferimento di B2, in seguito verranno creati **2 segmenti**: quello fino a **B2** e quello da B3 fino a **B4**. Agli score viene associato un indice che servirà per ricostruire i segmenti.

I valori di minimo locale vengono salvati in una lista a parte che in seguito verrà usata per rimodellare la posizione e la composizione dei blocchi di frasi ricavati dalla segmentazione iniziale. A questo punto, seguendo lo schema appena mostrato, si analizza nuovamente la lista di segmenti iniziali, ma questa volta con l'intento di fonderne alcuni secondo i nuovi intervalli ricavati dai minimi locali.

A termine dell'esecuzione, il programma stampa a video i risultati della nuova configurazione di segmenti e crea 2 file nella cartella **output**:

- **Initial_segments.txt**
- **Segmented_documents.txt**

Come si può intuire il primo conterrà la segmentazione originaria e il secondo quella ricavata dall'algoritmo.

Risultati

Per valutare il sistema è stato creato un file "corpus_3_fonti.txt" che racchiude alcuni paragrafi provenienti da documenti differenti e che parlano di "Northen Lights...", "I love my pet...", "the end of brexit...". Essendo corti i paragrafi, l'algoritmo è stato eseguito con una dimensione iniziale dei blocchi pari a 2. Di seguito i risultati, prima il documento originale poi quello segmentato.

```
The Northern Lights are the visible result of solar particles entering the earth's magnetic field and ionizing high i
Their intensity depends on the activity of the sun and the acceleration speed of these particles.
They appear as dancing lights high in the sky and vary in color.
The lights usually appear green, but occasionally also purple, red, pink, orange, and blue.
Their colors depend on the elements being ionized.
Solar activity is not regular, however.
Even if it is a dark, clear night, there could still be absolutely no chance of seeing the auroras due to a lack of s
It also means that the sky could be alive with Northern Lights on a midsummer day, but the sun's brightness obscures
Due to the nature of the earth's magnetic field, the auroras only appear at the poles.

I love to pet my cat while reading fantasy books.
The fur of my cat is so fluffy that chills me.
It entertain me, its tail sometimes goes over my pages and meows, also showing me the weird pattern in its chest's fu
Sometimes, it plays with me but hurts me with its claws and once it scratched me so bad the one drop of blood felt ov
Even so, its agile and soft body at least warms me and appease me with its cute fur, meow and purr, so I alwayse forg
Every time a cat is put aside of a character, i cannot prevent remembering when I found it malnourished and lacking i

The end of the Brexit transition period on 31 December is looming, and with it some major changes for anyone going fr
Some have been confirmed but others are dependent on continuing negotiations over a deal.
```

```
["The Northern Lights are the visible result of solar particles entering the earth's magnetic field and ionizing high in the atmosphe
re.", 'Their intensity depends on the activity of the sun and the acceleration speed of these particles.', 'They appear as dancing li
ghts high in the sky and vary in color.', 'The lights usually appear green, but occasionally also purple, red, pink, orange, and blue
.', 'Their colors depend on the elements being ionized.', 'Solar activity is not regular, however.', 'Even if it is a dark, clear nig
ht, there could still be absolutely no chance of seeing the auroras due to a lack of solar activity.', "It also means that the sky co
uld be alive with Northern Lights on a midsummer day, but the sun's brightness obscures them."]

["Due to the nature of the earth's magnetic field, the auroras only appear at the poles.", 'I love to pet my cat while reading fantas
y books.', 'The fur of my cat is so fluffy that chills me.', "It entertain me, its tail sometimes goes over my pages and meows, also
showing me the weird pattern in its chest's fur.", "Sometimes, it plays with me but hurts me with its claws and once it scratched me
so bad the one drop of blood felt over my favourite book's pages.", 'Even so, its agile and soft body at least warms me and appease m
e with its cute fur, meow and purr, so I alwayse forgive it and those little drops of damage.']

['Every time a cat is put aside of a character, i cannot prevent remembering when I found it malnourished and lacking in fur.', 'The
end of the Brexit transition period on 31 December is looming, and with it some major changes for anyone going from the UK to contine
ntal Europe – either to holiday, work or live.', 'Some have been confirmed but others are dependent on continuing negotiations over a
 deal.']
```

Il sistema ha tentato di dividere i paragrafi non riuscendoci, però. Un tentativo si può notare nel blocchi 2 e 3. In ambedue i casi, per una sola frase non è riuscito a separare correttamente. In sintesi, il sistema si presta ad essere migliorato ma si avvicina sufficientemente alla separazione per ambito tematico.