Relazione esercitazione – Segmentation

L'obiettivo di questa esercitazione è stato quello di definire un algoritmo in grado di eseguire la Document Segmentation. In poche parole, quello che l'algoritmo dovrà fare, sarà cercare di individuare all'interno di un documento testuale le posizioni in cui è presente un possibile cambiamento di discorso. L'algoritmo implementato è leggermente ispirato al Text-Tiling.

Il corpus di documenti utilizzati in questa esercitazione sono stati 2, il primo è presente nella cartella chiamata "**paragrafi_Segmentation**" ed è chiamato "**3_paragrafi_Corpus.txt**". Esso è costituito da una serie di frasi prese da Wikipedia riguardanti 3 topics differenti che sono: **Animals, Football e Artificial Intelligence.** Il secondo invece, è chiamato "**2_paragrafi_corpus**" e contiene sempre un insieme di frasi prese da Wikipedia, riguardanti però 2 topics: **Economics e Finance**.

Il programma è stato sviluppato nel modo seguente:

1) Una volta caricato l'insieme delle frasi del documento, viene chiamata la funzione "**text_tiling**" a cui **vengono dati in input 3 parametri:** percorso documento, numero di linee di separazione da considerare e numero di minimi locali da considerare.

2) La funzione text_tiling per prima cosa attraverso la chiamata alla funzione "**get_overlap_lessicale_tra_le_varie_coppie_di_frasi_documento**", si fa restituire tutti i valori della sovrapposizione lessicale tra tutte le coppie consecutive di frasi presenti nel documento. L'overlap tra ogni coppia di frasi viene ottenuto dalla funzione appena citata tramite l'ausilio di un'altra funzione chiamata **"overlap_tra_frasi**" che per cercare come al solito di aumentare il materiale lessicale esegue innanzitutto una fase di pre-processing su ognuna delle due frasi ricevute in input (rimozione punteggiatura, rimozione stop words e lemmatizzazione). Dopodichè per ciascun lemma presente in ognuna delle due frasi si collega a Wordnet e una volta ottenuti tutti i possibili synsets associabili a quel lemma, applica la WSD per ottenere il synset migliore usando come contesto la frase in cui si trova il lemma e come signature varie informazioni relative al synset, come ad esempio la lista dei lemmi, la sua definizione e la lista di esempi (qualora fossero disponibili) ad esso associato. In questa esercitazione si è notato che per ottenere risultati leggermente migliori, una volta trovato il best synset per un certo lemma di una certa frase sarebbe stato meglio aggiungere come materiale lessicale alla frase iniziale non solo le informazioni riguardanti il best synset trovato ma anche tutti i lemmi associati a tutti gli iperonimi e iponimi del best synset. Ovviamente sul materiale aggiunto alle due frasi è stata sempre eseguita una fase di pre-processing come quella citata all'inizio del passo 2) in modo tale da mantenere sempre "pulite" le stringhe di testo. Infine, l'overlap tra le due frasi "aumentate" ottenute, viene trovato calcolando l'intersezione dei termini tra le due frasi e normalizzandola rispetto alla lunghezza della frase più piccola.

3) Dopodichè, una volta ottenuti tutti i valori degli overlaps **tra ogni coppia di frasi consecutive per le quali si memorizza sia l'overlap e sia l'indice della prima frase della coppia**, si procede all'ordinamento in modo crescente di queste coppie in base al valore di overlap.

4) A questo punto entra in gioco uno dei parametri dati in input alla funzione text_tiling che è il numero di minimi locali da considerare, che sostanzialmente servirà per decidere quanti minimi locali considerare, chiaramente se questo parametro assume ad esempio valore 3, allora i minimi locali che saranno considerati saranno solamente i primi 3 minimi ottenuti che sono i più piccoli. Dopo alcune sperimentazioni il valore migliore individuato per questo parametro è risultato essere 3-5. (**Chiaramente num. minimi locali >= num. barre altrimenti si rischierebbe di non sapere dove posizionare tutte le barre**).

5) Una volta ottenuti i minimi locali più piccoli da considerare e quindi anche le posizioni in cui questi si trovano nel documento, quello che fa l'algoritmo è considerare tutte le possibili combinazioni tra le posizioni in base anche al numero di barre di separazione (altro parametro di input).

   Quello appena scritto al passo 5), forse sarà più chiaro con un esempio pratico:

   Supponiamo che:

   - Num. Barre = 2
   - Num. di massimi locali da considerare = 3

   Allora, se ad esempio le posizioni dei primi 3 minimi locali più piccoli sono le seguenti:

   [22, 12, 37]

   Allora, poiché le barre da posizionare sono 2, le possibili combinazioni di posizionamento delle barre saranno le seguenti:

   [12,22] – [12,37] – [22,37]

6) Arrivati a questo punto, sempre all'interno della funzione text_tiling viene chiamata la funzione "**trova_barre_migliori**" che prendendo in input la lista dei primi n minimi locali da considerare, il numero delle linee di separazione, il numero di frasi del documento e l'insieme delle frasi del documento, restituirà le posizioni migliori in cui posizionare le barre di separazione. Per farlo, la funzione citata pocanzi, suddivide tutte le frasi in base ad ogni possibile combinazione di barre. Ogni volta che le barre verranno posizionate in determinate posizioni (ad esempio: [12,22]), otterremo dei clusters di frasi (nel nostro esempio i clusters ottenuti saranno 3) e per ciascun cluster verrà restituita la sua coesione media interna sempre tramite il calcolo dell'overlap tra tutte le coppie di frasi adiacenti presenti nel cluster. **Infine, per ogni suddivisione fatta per le frasi del documento (quindi sempre ad esempio [12,22]) verrà calcolata la coesione media su tutti i clusters ottenuti e alla fine, sempre la funzione "trova_barre_migliori", restituirà la suddivisione che ha ottenuto il valore più alto per tale media. In questo modo potremmo essere più sicuri del fatto che poiché le barre posizionate in quelle posizioni, ci permettono di ottenere l'insieme di clusters più coeso possibile, allora di conseguenza sarà più probabile che proprio in quelle posizioni sarà effettivamente avvenuto un cambiamento di discorso.**

Risultati ottenuti

Insieme di frasi considerate nel **documento orginale chiamato: "3_paragrafi_Corpus.txt**":

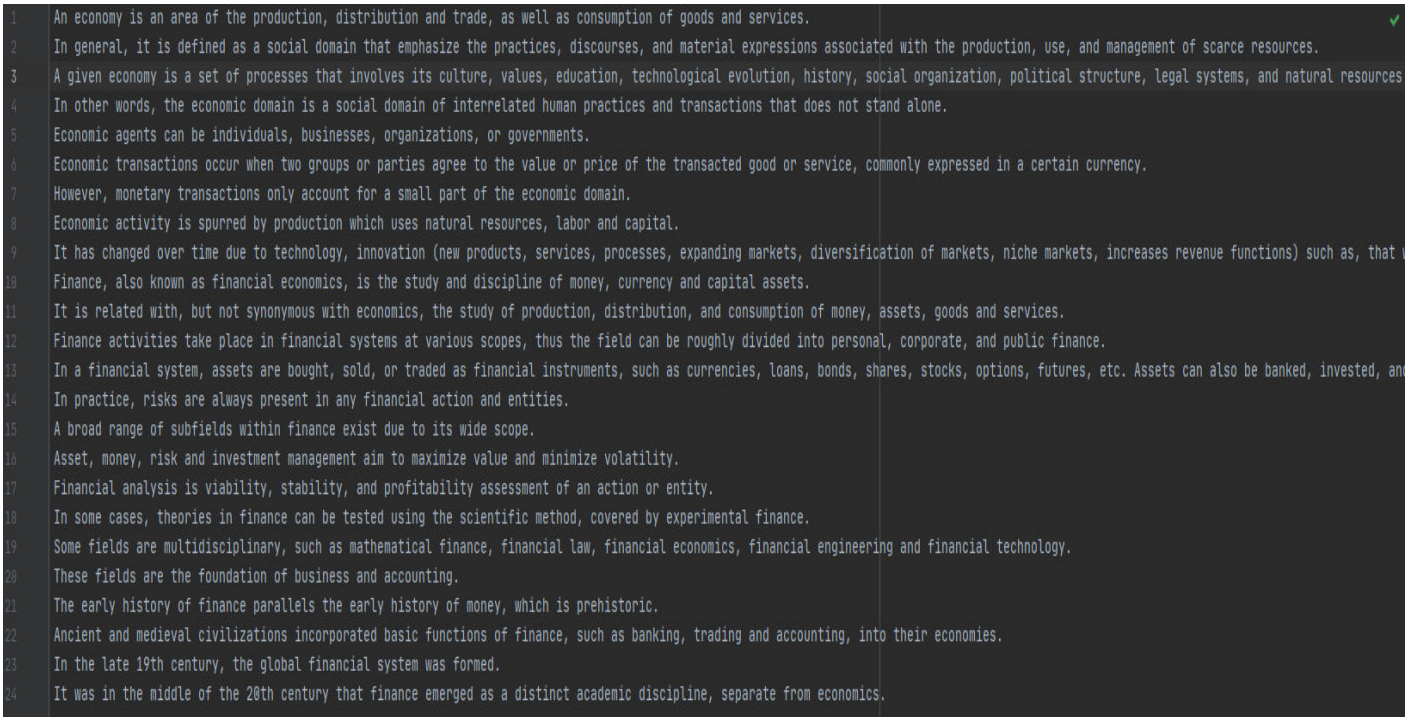
```
Animals are multicellular, eukaryotic organisms in the biological kingdom Animalia.
With few exceptions, animals consume organic material, breathe oxygen, are able to move, can reproduce sexually, and go through an ontogenetic stage in which their body consists of a hollow
The scientific study of animals is known as zoology.
Most living animal species are in Bilateria, a clade whose members have a bilaterally symmetric body plan.
The Bilateria include the protostomes, containing animals such as nematodes, arthropods, flatworms, annelids and molluscs, and the deuterostomes, containing the echinoderms and the chordate
Life forms interpreted as early animals were present in the Ediacaran biota of the late Precambrian.
Historically, Aristotle divided animals into those with blood and those without.
Carl Linnaeus created the first hierarchical biological classification for animals in 1758 with his Systema Naturae, which Jean-Baptiste Lamarck expanded into 14 phyla by 1809.
In 1874, Ernst Haeckel divided the animal kingdom into the multicellular Metazoa (now synonymous for Animalia) and the Protozoa, single-celled organisms no longer considered animals.
In modern times, the biological classification of animals relies on advanced techniques, such as molecular phylogenetics, which are effective at demonstrating the evolutionary relationships
Humans make use of many animal species, such as for food (including meat, milk, and eggs), for materials (such as leather and wool), as pets, and as working animals including for transport.
Dogs have been used in hunting, as have birds of prey, while many terrestrial and aquatic animals were hunted for sports.
Nonhuman animals have appeared in art from the earliest times and are featured in mythology and religion.
Football is a family of team sports that involve, to varying degrees, kicking a ball to score a goal.
Unqualified, the word football normally means the form of football that is the most popular where the word is used.
Sports commonly called football include association football (known as soccer in North America and Oceania); gridiron football (specifically American football or Canadian football); Austral
These various forms of football share to varying extent common origins and are known as football codes.
There are a number of references to traditional, ancient, or prehistoric ball games played in many different parts of the world.
Contemporary codes of football can be traced back to the codification of these games at English public schools during the 19th century.
The expansion and cultural influence of the British Empire allowed these rules of football to spread to areas of British influence outside the directly controlled Empire.
By the end of the 19th century, distinct regional codes were already developing: Gaelic football, for example, deliberately incorporated the rules of local traditional football games in ord
In 1888, The Football League was founded in England, becoming the first of many professional football associations.
During the 20th century, several of the various kinds of football grew to become some of the most popular team sports in the world.
Artificial intelligence (AI) is intelligence demonstrated by machines.
AI research has been defined as the field of study of intelligent agents, which refers to any system that perceives its environment and takes actions that maximize its chance of achieving i
The term "artificial intelligence" had previously been used to describe machines that mimic and display "human" cognitive skills that are associated with the human mind, such as "learning"
This definition has since been rejected by major AI researchers who now describe AI in terms of rationality and acting rationally, which does not limit how intelligence can be articulated.
AI applications include advanced web search engines (e.g., Google), recommendation systems (used by YouTube, Amazon and Netflix), understanding human speech (such as Siri and Alexa), self-d
As machines become increasingly capable, tasks considered to require "intelligence" are often removed from the definition of AI, a phenomenon known as the AI effect.
For instance, optical character recognition is frequently excluded from things considered to be AI, having become a routine technology.
Artificial intelligence was founded as an academic discipline in 1956, and in the years since has experienced several waves of optimism, followed by disappointment and the loss of funding (
AI research has tried and discarded many different approaches since its founding, including simulating the brain, modeling human problem solving, formal logic, large databases of knowledge
In the first decades of the 21st century, highly mathematical-statistical machine learning has dominated the field, and this technique has proved highly successful, helping to solve many ch
The various sub-fields of AI research are centered around particular goals and the use of particular tools.
The traditional goals of AI research include reasoning, knowledge representation, planning, learning, natural language processing, perception, and the ability to move and manipulate objects
General intelligence (the ability to solve an arbitrary problem) is among the field's long-term goals.
To solve these problems, AI researchers have adapted and integrated a wide range of problem-solving techniques - including search and mathematical optimization, formal logic, artificial neu
AI also draws upon computer science, psychology, linguistics, philosophy, and many other fields.
The field was founded on the assumption that human intelligence "can be so precisely described that a machine can be made to simulate it".
This raised philosophical arguments about the mind and the ethical consequences of creating artificial beings endowed with human-like intelligence; these issues have previously been explore
Computer scientists and philosophers have since suggested that AI may become an existential risk to humanity if its rational capacities are not steered towards beneficial goals.
```

Insieme di frasi considerate nel **documento segmentato chiamato: “3_paragrafi_corpus_risultati_segmentation.txt”**:

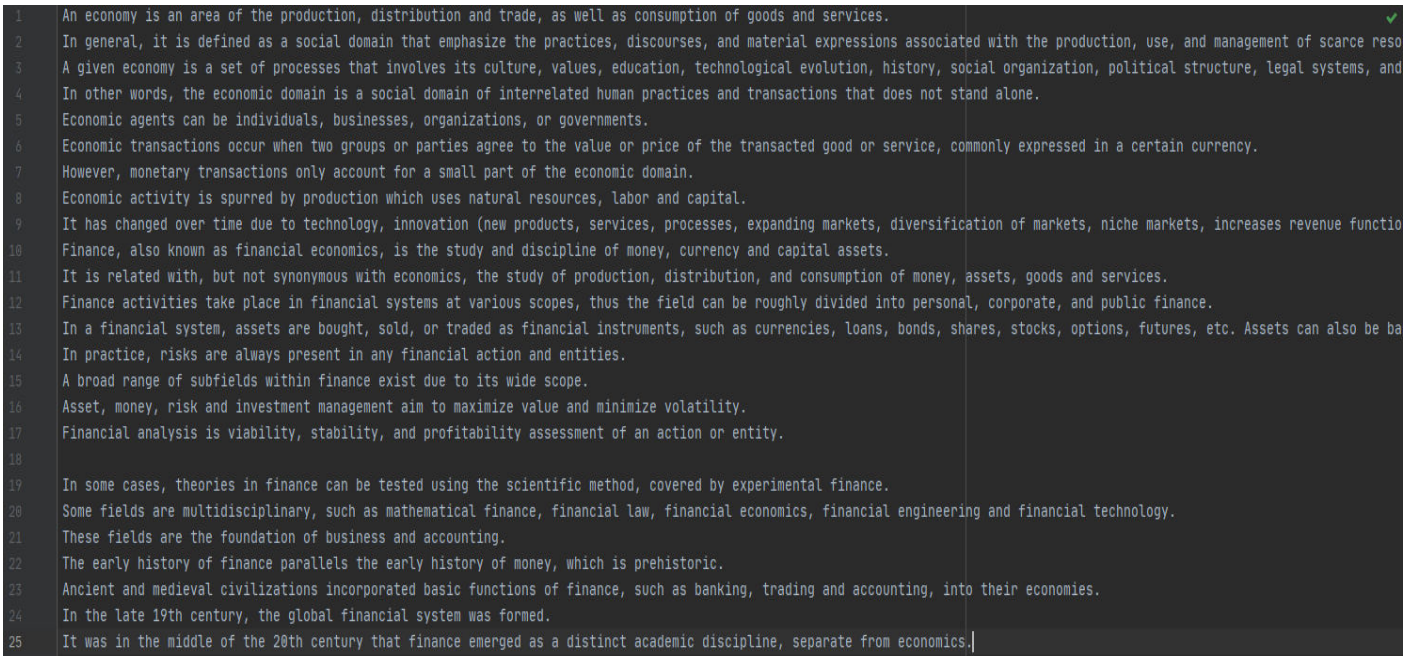
```
Animals are multicellular, eukaryotic organisms in the biological kingdom Animalia.
With few exceptions, animals consume organic material, breathe oxygen, are able to move, can reproduce sexually, and go through an ontogenetic stage in which their body consists of a hollow
The scientific study of animals is known as zoology.
Most living animal species are in Bilateria, a clade whose members have a bilaterally symmetric body plan.
The Bilateria include the protostomes, containing animals such as nematodes, arthropods, flatworms, annelids and molluscs, and the deuterostomes, containing the echinoderms and the chordate
Life forms interpreted as early animals were present in the Ediacaran biota of the late Precambrian.
Historically, Aristotle divided animals into those with blood and those without.
Carl Linnaeus created the first hierarchical biological classification for animals in 1758 with his Systema Naturae, which Jean-Baptiste Lamarck expanded into 14 phyla by 1809.
In 1874, Ernst Haeckel divided the animal kingdom into the multicellular Metazoa (now synonymous for Animalia) and the Protozoa, single-celled organisms no longer considered animals.
In modern times, the biological classification of animals relies on advanced techniques, such as molecular phylogenetics, which are effective at demonstrating the evolutionary relationships
Humans make use of many animal species, such as for food (including meat, milk, and eggs), for materials (such as leather and wool), as pets, and as working animals including for transport.
Dogs have been used in hunting, as have birds of prey, while many terrestrial and aquatic animals were hunted for sports.
Nonhuman animals have appeared in art from the earliest times and are featured in mythology and religion.

Football is a family of team sports that involve, to varying degrees, kicking a ball to score a goal.
Unqualified, the word football normally means the form of football that is the most popular where the word is used.
Sports commonly called football include association football (known as soccer in North America and Oceania); gridiron football (specifically American football or Canadian football); Austral
These various forms of football share to varying extent common origins and are known as football codes.
There are a number of references to traditional, ancient, or prehistoric ball games played in many different parts of the world.
Contemporary codes of football can be traced back to the codification of these games at English public schools during the 19th century.
The expansion and cultural influence of the British Empire allowed these rules of football to spread to areas of British influence outside the directly controlled Empire.
By the end of the 19th century, distinct regional codes were already developing: Gaelic football, for example, deliberately incorporated the rules of local traditional football games in ord
In 1888, The Football League was founded in England, becoming the first of many professional football associations.
During the 20th century, several of the various kinds of football grew to become some of the most popular team sports in the world.

Artificial intelligence (AI) is intelligence demonstrated by machines.
AI research has been defined as the field of study of intelligent agents, which refers to any system that perceives its environment and takes actions that maximize its chance of achieving i
The term "artificial intelligence" had previously been used to describe machines that mimic and display "human" cognitive skills that are associated with the human mind, such as "learning"
This definition has since been rejected by major AI researchers who now describe AI in terms of rationality and acting rationally, which does not limit how intelligence can be articulated.
AI applications include advanced web search engines (e.g., Google), recommendation systems (used by YouTube, Amazon and Netflix), understanding human speech (such as Siri and Alexa), self-d
As machines become increasingly capable, tasks considered to require "intelligence" are often removed from the definition of AI, a phenomenon known as the AI effect.
For instance, optical character recognition is frequently excluded from things considered to be AI, having become a routine technology.
Artificial intelligence was founded as an academic discipline in 1956, and in the years since has experienced several waves of optimism, followed by disappointment and the loss of funding (
AI research has tried and discarded many different approaches since its founding, including simulating the brain, modeling human problem solving, formal logic, large databases of knowledge
In the first decades of the 21st century, highly mathematical-statistical machine learning has dominated the field, and this technique has proved highly successful, helping to solve many ch
The various sub-fields of AI research are centered around particular goals and the use of particular tools.
The traditional goals of AI research include reasoning, knowledge representation, planning, learning, natural language processing, perception, and the ability to move and manipulate objects
General intelligence (the ability to solve an arbitrary problem) is among the field's long-term goals.
To solve these problems, AI researchers have adapted and integrated a wide range of problem-solving techniques - including search and mathematical optimization, formal logic, artificial neu
AI also draws upon computer science, psychology, linguistics, philosophy, and many other fields.
The field was founded on the assumption that human intelligence "can be so precisely described that a machine can be made to simulate it".
This raised philosophical arguments about the mind and the ethical consequences of creating artificial beings endowed with human-like intelligence; these issues have previously been explore
Computer scientists and philosophers have since suggested that AI may become an existential risk to humanity if its rational capacities are not steered towards beneficial goals.
```

E’ possibile notare come l’algoritmo sia riuscito, in questo caso, a separare correttamente i 3 paragrafi e quindi ad individuare correttamente le posizioni di cambiamento del discorso.

Insieme di frasi considerate nel **documento orginale chiamato:** “**2_paragrafi_corpus.txt**”:



Insieme di frasi considerate nel **documento segmentato chiamato:**
“2_paragrafi_corpus_risultati_segmentation.txt”



In questo caso si può invece notare come l’algoritmo non sia riuscito ad individuare l’effettivo cambio di discorso presente in realtà tra la riga 9 e 10. Probabilmente il motivo del fallimento è da attribuire al fatto che in questo secondo esperimento i topics trattati erano “Economics” e “Finance” che comunque risultano essere abbastanza simili tra loro. Quindi possiamo dire che l’algoritmo sembra

avere un comportamento piuttosto buono nel momento in cui i topics sono abbastanza diversi tra loro, mentre ha un comportamento molto meno preciso quando gli argomenti iniziano ad essere abbastanza simili tra loro.