Relazione esercitazione – TM/TV

L'obiettivo di questa esercitazione è stato quello di utilizzare un algoritmo per il Topic Modelling che risulta essere un altro possibile task nell'ambito dell'NLP e in particolar modo in quella che viene chiamata **Semantica Documentale**. Lo scopo del Topic modelling è quello di partire da una collezione di documenti e individuare quelli che sono i topics trattati all'interno di essi. L'algoritmo utilizzato in questa esercitazione è stato l'**LDA (Latent Diritchlet Allocation)** che in sostanza è un'estensione di una versione probabilistica dell'LSA (Latent Semantic Analysis) che è anch'esso un algoritmo di Topic Modelling basato sulla fattorizzazione matriciale SVD. L'LDA invece, si basa sulla statistica Bayesiana e in particolare si basa sull'assunto che un documento è un insieme di topics in cui ogni parola ha un certo peso, che in questo caso è una probabilità, e che ci dice quanto è probabile che una certa parola compaia in ogni singolo topic. L'LDA ha bisogno di ricevere in input il numero di topics e una collezione di documenti. La prima cosa che fa è quella di assegnare in maniera casuale ciascuna parola dei documenti ad un topic qualsiasi. In questo modo si può ottenere una prima distribuzione di probabilità sia dei topics che delle parole presenti in essi; questo ci permette quindi di capire che ogni topic avrà una sua distribuzione di probabilità sul tutto il vocabolario di termini. Quello che fa successivamente l'algoritmo è aggiornare le probabilità assegnate alle parole per ogni topic, fino a quando non arriva ad una convergenza.

I passi seguiti dall'algoritmo implementato per l'esercitazione sono i seguenti:

1) Viene caricato il corpus di documenti che è stato creato utilizzando frasi prese da Wikipedia su 3 argomenti principali: **Phisycs, Music e Artificial Intelligence**. Il corpus è presente nella cartella chiamata "**Risorse_Topic_Modelling_Text_Visualization**" ed è chiamato "**documenti TM-TV.txt**". Ogni riga del corpus verrà visto dall'algoritmo come un singolo documento. Per questioni di maggiore chiarezza ogni singolo documento è stato inserito all'interno della cartella chiamata "**Documenti**".

2) Su ogni documento è stata eseguita una fase di pre-processing in cui:

   - Sono stati rimossi i segni di punteggiatura ed eventuali caratteri inutili.
   - E' stata effettuata l'eliminazione di eventuali stringhe vuote.
   - E' stata effettuata l'eliminazione delle stop words.
   - E' stata applicata la lemmatizzazione a tutte le parole presenti nel documento.
   - **Sono state eliminate dai documenti tutte le parole che comparivano una sola volta, poiché queste sono state ritenute poco rilevanti al fine della corretta individuazione dei topics.**

3) E' stato creato un dizionario per permettere di mantenere il mapping tra ogni parola e il suo id univoco.

4) Successivamente, ogni documento è stato trasformato in un vettore formato da coppie fatte in questo modo: [(0,1), (1,2), (2,1), ecc…]. In cui praticamente il primo valore di ogni coppia è l'id

univoco di una certa parola e il secondo valore corrisponde invece alla frequenza di tale parola nel documento considerato.

5) Dopodichè, tutti i documenti vettorizzati vengono dati in input alla funzione "**LdaModel" della libreria "Gensim"** che prende in input:

   - Documenti vettorizzati.
   - Dizionario delle parole.
   - Numero di topics da individuare.

   E che si preoccuperà di applicare l'algoritmo LDA.

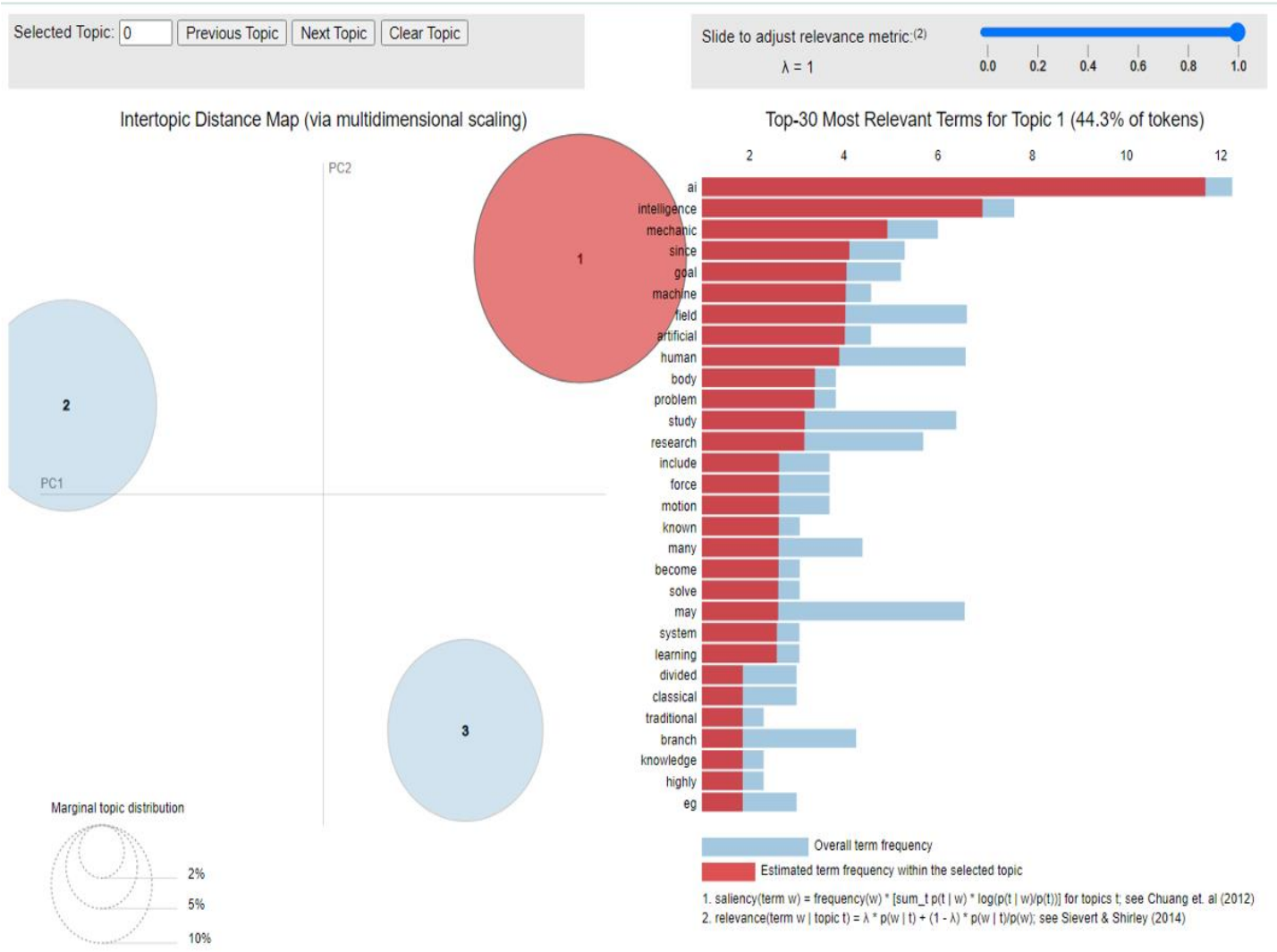
6) L'interpretazione e la visualizzazione dei risultati prodotti dalla funzione citata al passo 5, è stata eseguita mediante l'uso della libreria "**pyLDAvis**".

7) Infine, è stato eseguito il clustering sui documenti, utilizzando come centroidi di ciascun cluster l'insieme dei termini presenti all'interno di ciascun topic individuato dalla funzione LdaModel. In questo caso quindi, l'obiettivo è stato quello di associare ciascun documento ad un certo topic (o cluster) andando a calcolare l'overlap lessicale tra le parole di ogni documento e quello di ogni topic. Chiaramente, il cluster a cui verrà associato ciascun documento, sarà semplicemente il topic che otterrà il punteggio di overlap più alto.
   Nell'ultima pagina della relazione sono presenti i calcoli di alcune metriche di valutazione del clustering eseguito.

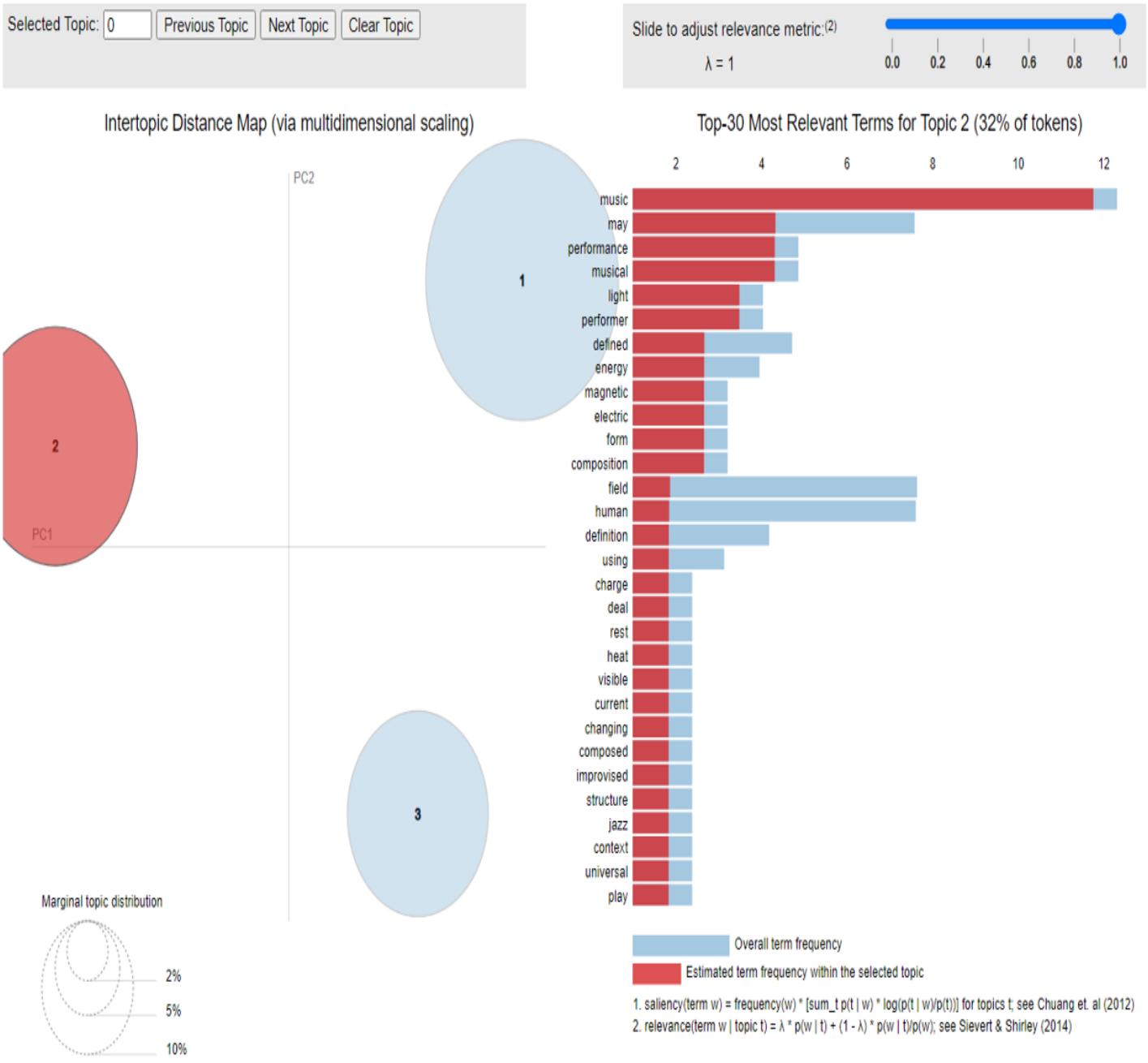
Visualizzazione e interpretazione dei risultati:

Di seguito vengono presentati in maniera grafica i risultati ottenuti dall'algoritmo LDA. Ogni cerchio rappresenta sostanzialmente un topic:

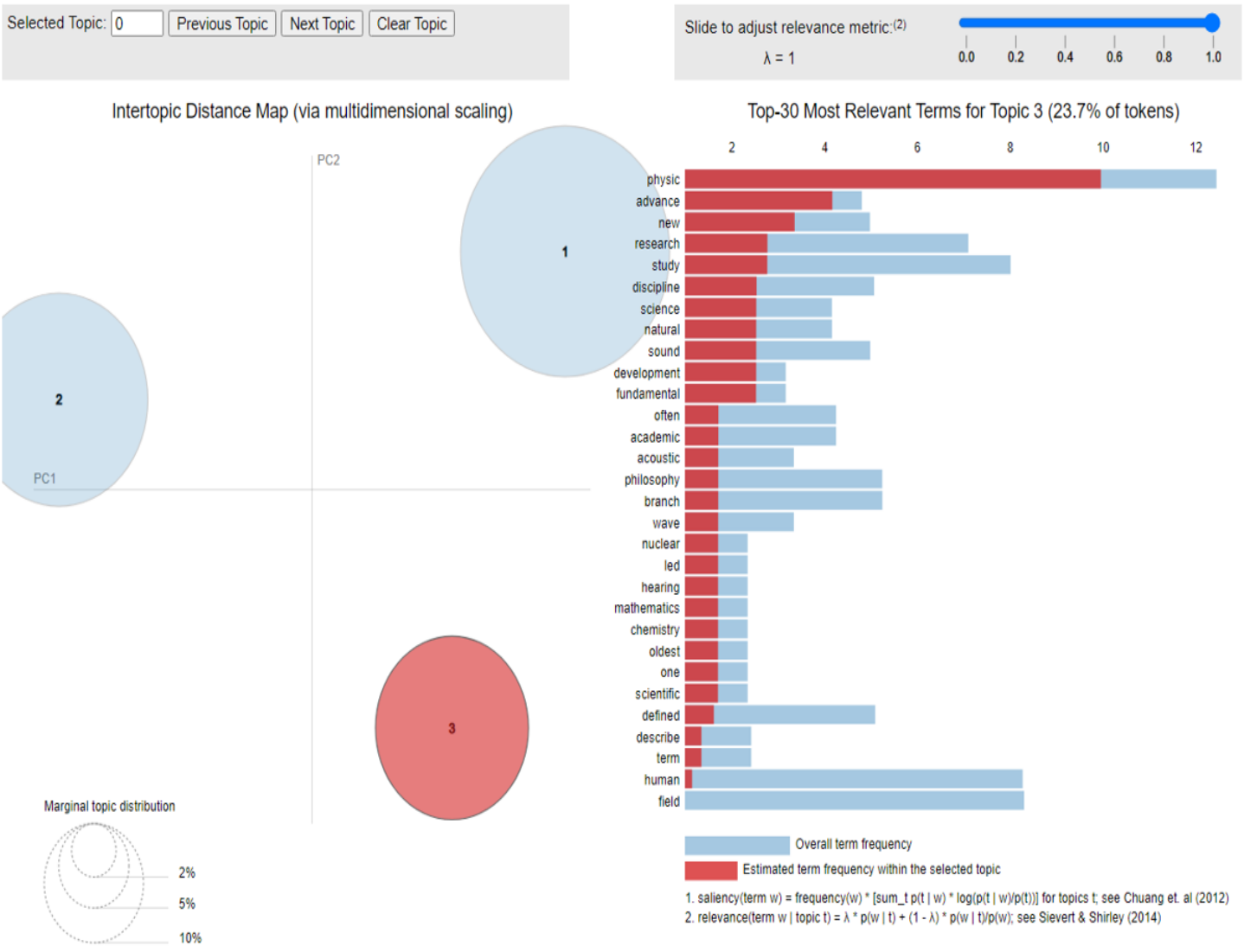
- **Il Topic-1 possiamo supporre che riguardi l'Artificial Intelligence.**
- **Il Topic-2 possiamo supporre che riguardi la Phisyc.**
- **Il Topic-3 possiamo supporre che riguardi la Music.**



Nel Topic-1 si può notare come la maggior parte dei termini presenti al suo interno riguardino per lo più l'AI, in particolar modo da notare l'elevata frequenza dei termini: ai, human, intelligence, machine, artificial.



Nel Topic-2 è invece da notare la presenza di parole come: music, musical, performance, performer, composition, play, jazz, ecc.. Legate comunque per lo più all'ambito musicale.



Nel Topic-3 si può notare l'elevata frequenza di termini come: physic, mechanic, science, research, sound, mathematics, wave, ecc… Per lo più parole che fanno riferimento comunque all'ambito della Fisica.

Dai 3 topics ottenuti è evidente però il fatto che comunque ogni topic, contenga delle parole con frequenza più o meno elevata che non sono poi così tanto legate all'argomento principale a cui possiamo pensare che quel determinato topic faccia riferimento. Un esempio è la presenza nel topic-1 delle parole "force" e "field" legate chiaramente più all'ambito della Fisica piuttosto che a quello dell'Intelligenza Artificiale.

Clustering e risultati ottenuti:

Centroidi dei 3 clusters individuati:

```
termini cluster_0:  ['ai', 'intelligence', 'mechanic', 'since', 'goal', 'machine', 'field', 'artificial', 'human', 'body']
termini cluster_1:  ['physic', 'advance', 'new', 'research', 'study', 'discipline', 'science', 'sound', 'natural', 'development']
termini cluster_2:  ['music', 'may', 'performance', 'musical', 'light', 'performer', 'defined', 'energy', 'magnetic', 'electric']
```

Di seguito vengono elencati i *clusters realmente corretti*, ciascun colore è associato ad un certo cluster ovvero ad un certo topic:

- **Cluster "Artificial Intelligence"**:
  Cluster_0 = [documento_8, documento_9, documento_10, documento_11]
- **Cluster "Physics"**: Cluster_1 = [documento_0, documento_1, documento_2, documento_3, documento_4]
- **Cluster "Music"**:
  Cluster_2 = [documento_5, documento_6, documento_7]

Di seguito invece, vengono riportati i *clusters ottenuti dall'algoritmo*:

- Cluster_0 = [documento_2, documento_8, documento_9, documento_10, documento_11]
- Cluster_1 = [documento_0, documento_1, documento_3]
- Cluster_2 = [documento_4, documento_5, documento_6, documento_7]

E' possibile notare come alcuni documenti riguardanti il topic "Physics" siano stati inseriti dall'algoritmo nei clusters sbagliati. Questo probabilmente è accaduto perché all'interno di tali documenti erano presenti un numero maggiore di termini associati al cluster_0 o al cluster_2 piuttosto che al cluster_1.

Di seguito vengono calcolate alcune *metriche di valutazione del clustering*:

$$\textbf{\textit{Accuracy clustering}} = \frac{\textbf{\textit{numero di documenti clusterizzati correttamente}}}{\textbf{\textit{numero di documenti totali}}}$$

$$\textbf{\textit{Precision cluster\_n}} = \frac{\textbf{\textit{TP}}}{\textbf{\textit{TP}} + \textbf{\textit{FP}}}$$

$$\textbf{\textit{Recall cluster\_n}} = \frac{\textbf{\textit{TP}}}{\textbf{\textit{numero totale di documenti del cluster\_n}}}$$

$$\textbf{\textit{Accuracy clustering}} = \frac{10}{12} \textbf{ = 0.83}$$

	Cluster_0	Cluster_1	Cluster_2
Precision	4/4 = **1**	3/(3+2) = **0.60**	3/3 **= 1**
Recall	4/4 = **1**	3/5 = **0.60**	3/3 **= 1**