

TV SERIES ANALYSIS: THE BIG BANG THEORY – THE VAMPIRE DIARIES

INDICE

INTRODUZIONE

PUNTO 1: ESTRAZIONE SCRIPT DELLE PUNTATE

1.1 Estrazione script delle puntate: "The Vampire Diaries"

1.2 Estrazione script delle puntate: "The Big Bang Theory"

PUNTO 2: SENTIMENT ANALYSIS DELLE SERIE

2.1 Calcolo del sentiment: metodologia

2.2 Il codice

2.3 Analisi dei risultati ottenuti

 2.3.1. Analisi dei risultati: "The Vampire Diaries"

 2.3.2. Analisi dei risultati: "The Big Bang Theory"

2.4 Sentiment Analysis delle serie a confronto

PUNTO 3: SENTIMENT ANALYSIS DEI PERSONAGGI

3.1 Il codice

3.2 Analisi dei risultati ottenuti

 3.2.1 Analisi dei risultati: "The Vampire Diaries"

 3.2.2 Analisi dei risultati: "The Big Bang Theory"

PUNTO 4: TOPIC MODELING

4.1 Modello LDA e NMF

 4.1.1 Il codice

 4.1.2 Analisi dei topic: "The Vampire Diaries"

 4.1.3 Analisi dei topic: "The Big Bang Theory"

4.2 Visualizzazioni HTML

 4.2.1 Il codice

 4.2.2 Analisi delle visualizzazioni: "The Big Bang Theory"

 4.2.3 Analisi delle visualizzazioni: "The Vampire Diaries"

PUNTO 5: CONFRONTO DEI TOPICS TRA LE SERIE IN ANALISI

5.1: Analisi con WordCloud

CONCLUSIONE

INTRODUZIONE

Il progetto si propone di analizzare due serie televisive distinte: "The Big Bang Theory," una sitcom americana trasmessa tra il 2007 e il 2019, e "The Vampire Diaries," una serie televisiva americana caratterizzata dai generi dell'orrore, del fantasy, del dramma e del sentimentale, trasmessa tra il 2009 e il 2017.

"The Big Bang Theory" presenta in modo ironico le vicende quotidiane di un gruppo di scienziati, Leonard, Sheldon, Howard e Raj, che lavorano al California Institute of Technology. Nonostante questi siano tra le menti più brillanti del Paese, la loro intelligenza straordinaria li rende socialmente inadeguati. La serie televisiva presenta il tema predominante la quotidianità, i legami relazionali e d'amicizia, il tema scientifico-accademico e quello ricreativo.

"The Vampire Diaries" racconta la storia di Elena, una ragazza adolescente che vive nella cittadina fittizia di Mystic Falls, in Virginia. La sua vita viene sconvolta quando scopre che il suo fidanzato, Stefan Salvatore, protagonista della serie, e suo fratello Damon Salvatore, sono vampiri. I due fratelli si innamorano di Elena ed entrano a far parte attivamente della sua vita. I temi di questa serie sono molto vari, principalmente relativi al tema sentimentale e a quello cupo e misterioso dell'orrore e del fantasy, caratterizzato dalla presenza di creature sovraumane.

Queste due serie, scelte appositamente per le loro profonde differenze in termini di generi e trame, saranno oggetto di un'analisi completa. Utilizzando diversi metodi di analisi del testo, ci proponiamo di esplorarle in modo approfondito, concentrandoci prima sugli aspetti complessivi e successivamente approfondendo l'analisi dei personaggi. L'obiettivo finale è determinare le effettive differenze tra le due serie.

Gli aspetti principali di questa analisi includono:

- 1 L'estrazione di tutti gli script delle puntate per entrambe le serie;
- 2 L'effettuazione di un'analisi del sentimento al fine di determinare il sentimento medio di ciascun episodio e, successivamente, il sentimento medio di ciascuna stagione;
- 3 L'analisi del sentimento di tre personaggi principali per ciascuna serie (Stefan Salvatore, Damon Salvatore ed Elena Gilbert per The Vampire Diaries; Sheldon Cooper, Leonard Hofstadter e Penny per The Big Bang Theory) per calcolare il sentimento medio di ognuno di essi in ciascun episodio e, successivamente, il sentimento medio per ogni stagione;
- 4 Lo sviluppo di una Topic Modeling per identificare i principali argomenti trattati in ciascuna stagione e, infine, nell'intera serie.
- 5 Il confronto dei temi emersi tramite la Topic Modeling tra le due serie televisive, seguito dall'elaborazione delle conclusioni.

In questo modo, miriamo a esaminare in dettaglio le due serie, mettendo in evidenza sia le loro differenze che eventuali punti in comune nell'approccio narrativo e tematico, basandoci sugli stessi dialoghi dei personaggi.

PUNTO 1: ESTRAZIONE SCRIPT DELLE PUNTATE

1.1 Estrazione script delle puntate: “The Vampire Diaries”

Il codice sviluppato per l'estrazione degli script di questa serie è riportato nel file ‘1_Extraction_tvd.py’.

La logica del codice di questo file consiste nel creare un corpus di testi che contenga tante directory quante sono le stagioni e che all'interno della directory di ciascuna stagione salvi i file di testo di tutti gli episodi della stagione corrispondente.

Prima di tutto, però, è stato necessario trovare una pagina web che fornisse tutti i testi di ciascuna puntata di ogni stagione, e la fonte scelta è stata la pagina ‘https://vampirediaries.fandom.com/wiki/Category:Episode_Transcripts’, che corrisponde del file python variabile “url”.

Analizzando più nello specifico la struttura della pagina, questa in realtà contiene oltre agli script delle puntate di ‘The Vampire Diaries’, anche quelli della serie spin-off ‘The Originals’ che non è di nostro interesse.

Dunque, una volta esplorata la pagina principale, viene costruita una funzione che abbia come output una struttura dati che contenga le stagioni, e per ogni stagione i link delle rispettive puntate.

La funzione dedicata è “link_extractor(url)”, che prende in ingresso la variabile che contiene la pagina web principale prima citata, e da questa, utilizzando le librerie “requests” e “BeautifulSoup”, effettua lo scraping della pagina, per andare a ricercare i tag HTML in cui si trovano i link delle puntate. La funzione itera tra le stagioni, contenute nel tag “div” e per ogni stagione ricerca i tag “a”, ognuno dei quali contiene una parte del link per accedere a una puntata. Attraverso un ciclo for, si itera tra i vari episodi e attraverso una condizione if si inseriscono soltanto i dati che sono di nostro interesse, escludendo quelli relativi alla serie tv “The Originals” e quelli che non riguardano gli script delle puntate.

Pertanto, la funzione restituisce un dizionario in cui le chiavi sono i nomi delle stagioni e i valori sono delle liste contenenti tutti i link degli episodi di quella stagione.

Dopo aver ottenuto tutti i link di tutte le puntate, per la creazione del corpus di testi era dunque necessario estrarre il testo di ciascun episodio e salvarlo all'interno di un file di testo.

La funzione che assolve questo compito è “extracting_texts(episode,season_dir,episode_number)” che prende in input, il link dell'episodio di cui si vuole estrarre il testo, la directory in cui inserirlo, e il numero dell'episodio.

Questa funzione, sempre con le stesse librerie sopra citate, esplora la pagina dell'episodio e per quell'episodio crea un file di testo all'interno del quale riportare il testo della puntata. Nella funzione vengono inserite delle condizioni affinché venga estratto soltanto il testo di interesse, modificando ed escludendo caratteri non rilevanti o che rendevano differente la formattazione del testo tra i vari episodi.

All'interno delle precedenti funzioni sono state gestite le eccezioni, con i blocchi try ed except relativamente allo scraping delle pagine.

Infine, è stato possibile creare la funzione per creare il corpus di testi, “create_corpus(d)”, che prende in input il risultato della funzione “link_extractor(url)”, crea, se non esiste, il corpus di testi “corpus the vampire diaries”, itera tra la struttura dati in input, dunque tra le stagioni della serie, creando una directory per ogni stagione, e infine itera tra i link degli episodi eseguendo la funzione “extracting_texts(episode_url, season_dir,episode_number)” creando così il corpus di testi descritto all'inizio del paragrafo.

In questo modo è stata creata una struttura dati ordinata, che permette di accedere separatamente alle varie stagioni ed episodi della serie e che costituisce la base per effettuare le successive analisi.

1.2 Estrazione script delle puntate: “The Big Bang Theory”

Il codice sviluppato per l'estrazione degli script di questa serie è riportato nel file ‘1_Extraction_tbbt.py’

La logica del codice di questo file, come fatto precedentemente per il file ‘1_Extraction_tvd.py’ consiste nel creare un corpus di testi che contenga tante directory quante sono le stagioni e che all'interno della directory di ciascuna stagione salvi i file di testo di tutti gli episodi della stagione corrispondente. Tuttavia, a causa della diversa struttura della pagina web, il codice per l'estrazione degli script è diverso da quello predisposto per The Vampire Diaries.

Una volta trovata ed esplorata la pagina web che fornisce tutti i testi di ciascuna puntata di ogni stagione (“<https://bigbangtrans.wordpress.com>”) corrisposta alla variabile “url” del file python, è stata costruita una funzione avente come output una struttura dati che contenga le stagioni, e per ogni stagione i link delle rispettive puntate.

La funzione dedicata è “link_extractor(url)”, che prende in ingresso la variabile che contiene la pagina web principale prima citata, e da questa, utilizzando le librerie “requests” e “BeautifulSoup”, effettua lo scraping della pagina, per andare a ricercare i tag HTML in cui si trovano i link delle puntate restituendo un dizionario in cui le chiavi sono i nomi delle stagioni (ad esempio, 'Season01') e i valori sono liste di link agli episodi per ogni stagione (ossia “season_links”: questo dizionario memorizzerà i collegamenti degli episodi raggruppati per stagione; la lista ‘links’ invece viene utilizzata per memorizzare temporaneamente i collegamenti degli episodi per ogni stagione). Tra le variabili si trova anche la richiesta HTTP all'url fornito memorizzata nella variabile page; il parser che analizza il contenuto della pagina web utilizzando BeautifulSoup; ‘episode_links’ che con la funzione ‘parser.find_all('a', href=True)’ trova tutti gli elementi HTML ‘<a>’ che hanno l'attributo ‘href’ estraendo tutti i collegamenti dalla pagina.

A questo punto vengono estratte le informazioni sulla stagione e sull'episodio: se il collegamento rappresenta una stagione (individuata dalla presenza della parola 'Series'), la funzione estraе il numero della stagione e crea un gruppo di collegamenti per quella stagione nel dizionario ‘season_links’. Se la stagione è già presente nel dizionario, appende semplicemente il collegamento alla lista dei links.

Dopo aver ottenuto tutti i link di tutte le puntate, per la creazione del corpus di testi era dunque necessario estrarre il testo di ciascun episodio e salvarlo all'interno di un file di testo.

La funzione che assolve questo compito è “extracting_texts(episode,season_dir,episode_number)” che prende in input il link dell'episodio di cui si vuole estrarre il testo, la directory in cui inserirlo, e il numero dell'episodio.

Questa funzione, sempre con le stesse librerie sopra citate, dopo aver fatto la richiesta HTTP all'URL specifico dell'episodio e memorizzata la risposta con la variabile ‘page=requests.get(episode)’, ha fatto il parsing con BeautifulSoup per analizzare il contenuto HTML della pagina dell'episodio.

In seguito per estrarre il testo è stata usata la funzione ‘parser.find_all('div', class_="entrytext")’ che trova tutti gli elementi HTML `<div>` che hanno la classe ‘entrytext’, e per ogni `<div>` trovato si estraggono tutti i paragrafi `<p>` all'interno del `<div>`.

È stata quindi creata una stringa ‘script_text’ per memorizzare il testo del copione, includendo solo i paragrafi che contengono il carattere ‘:’ (che presumibilmente rappresentano i dialoghi) e inoltre, se il paragrafo contiene delle ‘(’ nei dialoghi, anche questo viene aggiunto a script_text.

Infine, viene eseguita la pulizia del testo per ogni riga nella stringa `script_text`, sostituendo alcuni caratteri non desiderati con spazi o altri caratteri.

Viene quindi restituito il testo pulito (`cleaned_text`) scritto in un file di testo specifico per quell'episodio, situato nella directory della stagione, con un nome file che contiene il numero dell'episodio.

Infine, è stata definita la funzione per creare il corpus di testi, “`create_corpus(d)`”, che accetta il dizionario dei link agli episodi per ogni stagione e che crea la struttura di directory “corpus the big bang theory”. Iterando su ogni stagione presente nel dizionario (`d`) crea una sottodirectory, qualora non esista, all'interno della directory del corpus, denominandola con il numero della stagione. Viene poi richiamata la funzione `extracting_texts(episode_url, season_dir, episode_number)` che itera su ogni episodio all'interno della stagione estraendone il testo, in modo da scaricare e salvare lo script per ogni episodio.

In questo modo è stata creata una struttura dati ordinata, che permette di accedere separatamente alle varie stagioni ed episodi della serie e che costituisce la base per effettuare le successive analisi.

PUNTO 2: SENTIMENT ANALYSIS DELLE SERIE

2.1 Calcolo del sentiment: metodologia

L'obiettivo principale di questa analisi è esplorare in profondità i corpus testuali creati delle due serie TV, al fine di identificare e valutare il sentiment presente in ciascun episodio. Questo processo consente di ottenere una comprensione più approfondita delle dinamiche emotive delle serie, rivelando il tono prevalente in ogni episodio e apripendo la strada a un'analisi comparativa tra le diverse stagioni.

La Sentiment Analysis ci permette di determinare non solo il sentiment individuale degli episodi, ma anche di aggregare questi risultati per ottenere una visione d'insieme del sentiment nelle stagioni e di identificare tendenze e pattern lungo il corso delle serie, mettendo in luce episodi con sentiment più positivi o negativi.

Infine, è possibile effettuare un confronto sul sentiment tra le due diverse serie, per scoprire se le differenze che le caratterizzano, a livello di trama e di genere, si riflettono anche sul sentiment.

Nel presente report vengono presentate le analisi di sentiment effettuate sui dialoghi delle due serie, con l'obiettivo di valutare il sentiment delle battute dei personaggi all'interno di ogni episodio e aggregato per stagione. Sono stati utilizzati due metodi di analisi: Vader e TextBlob, per poter valutare eventuali differenze tra i risultati restituiti dai due metodi, e scegliere quello più adatto all'analisi.

2.2 Il codice

Per questa analisi è stato creato il file “2_SentAnalysis.py”.

Per preparare i dati all'analisi del sentiment, è stato essenziale effettuare una pulizia del testo.

La funzione “clean(episode,custom_words)” è stata sviluppata per svolgere questa operazione, essa riceve in input il percorso dell'episodio e una lista di parole personalizzate da escludere dall'analisi, che variano tra le due serie. La procedura di pulizia include:

- La conversione del testo in caratteri minuscoli;
- L'estrazione delle battute dai dialoghi, eliminando la parte relativa al personaggio che dice la battuta;
- La rimozione della punteggiatura;
- L'eliminazione delle stopwords inglesi;
- La rimozione di eventuali caratteri numerici;
- L'eliminazione delle parole con meno di tre caratteri;
- L'esclusione delle parole personalizzate specifiche per ciascuna serie tv;
- L'eliminazione delle battute vuote.

La funzione “clean” restituirà una lista, che comprende le battute “ripulite” di un episodio.

Per il calcolo del sentiment sono state create due funzioni specifiche:

- 1) “Sentiment_for_episode_vader(cleaned_text)”: questa funzione, funzione utilizza Vader per calcolare il sentiment di ciascuna battuta all'interno di un episodio e ne calcola la media per l'intero episodio;
- 2) “Sentiment_for_episode_textblob(cleaned_text)”: questa funzione, invece, utilizza TextBlob per il medesimo scopo, dunque calcolare il sentiment di ciascuna battuta di un episodio e calcolarne la media per l'intero episodio.

Una volta che sono state costruite queste funzioni, l'analisi del sentiment è stata condotta in due modi, prima episodio per episodio e successivamente per stagione.

L'analisi del sentiment per episodio viene eseguita dalla funzione “visualize_episodes_sentiment(root_dir, serie name, custom_words)”.

Questa funzione, itera attraverso il corpus principale, composto di tutte le stagioni e le puntate, costruito nel punto precedente, e calcola attraverso le funzioni di calcolo del sentiment, il sentiment di ciascun episodio. I risultati dell'analisi vengono riportati in un dataframe che contiene le seguenti colonne:

- “Season”: la stagione dell’episodio;
- “Episode”: il numero dell’episodio;
- “Vader_sent_label”: l’etichetta di sentiment calcolata con Vader;
- “Textblob_sent_label”: l’etichetta di sentiment calcolata con TextBlob;
- “Vader_sent_numeric”: il valore numerico di sentiment calcolato con Vader, arrotondato a quattro cifre decimali;
- “Textblob_sent_numeric”: il valore numerico di sentiment calcolato con TextBlob, arrotondato a quattro cifre decimali.

I valori numerici di sentiment sono quelli che vengono direttamente restituiti dai suoi strumenti di analisi del sentiment, mentre i valori etichetta vengono attribuiti mediante la funzione “assign_sentiment_label(value)”.

Questa funzione restituisce valore “positive” quando il valore del sentiment è maggiore di zero, “negative” quando è minore di zero e infine “neutral” quando è uguale a zero.

L'analisi del sentiment per stagione viene invece eseguita dalla funzione “visualize_season_sentiment(root_dir, serie name, custom_words)”. Questa funzione, a pari modo della precedente, itera attraverso il corpus principale, e dunque tra tutte le stagioni e gli episodi, ma oltre a calcolare il sentiment per ciascun episodio, calcola il sentiment aggregato di ogni stagione, facendo la media tra i sentiment degli episodi. Anche stavolta i risultati vengono raccolti in un dataframe che comprende le seguenti colonne:

- “Season”: la stagione;
- “Vader_sent_label”: l’etichetta di sentiment aggregata calcolata con Vader;
- “Textblob_sent_label”: l’etichetta di sentiment aggregata calcolata con TextBlob;
- “Vader_sent_numeric”: il valore numerico di sentiment aggregato calcolato con Vader, arrotondato a quattro cifre decimali;
- “Textblob_sent_numeric”: il valore numerico di sentiment aggregato calcolato con TextBlob, arrotondato a quattro cifre decimali.

Anche qui i valori etichetta sono quelli restituiti dalla funzione “assign_sentiment_label(value)”.

2.3 Analisi dei risultati ottenuti

Dopo aver effettuato l'analisi del sentiment di entrambe le serie e aver raccolto i risultati nei dataframe, che sono stati salvati nella directory “df” all'interno degli appositi file csv, utilizziamo questi file per eseguire un'accurata e specifica esaminazione dei risultati, anche attraverso dei grafici che permettano di rendere l'analisi più comprensibile e informativa.

2.3.1. Analisi dei risultati: “The Vampire Diaries”

L'analisi del sentiment della serie “The Vampire Diaries” è appunto stata svolta su due livelli, prima in modo più specifico, esaminando il sentiment di ogni singolo episodio e successivamente, a livello più aggregato calcolando il sentiment di ogni stagione.

Per prima cosa concentriamo l'analisi sul sentiment di ogni episodio.

Per visualizzare i risultati è possibile aprire il file “2_episode_result.tvd.csv”, contenuto nella directory “df”, nel quale è salvato l’output restituito dalla funzione “visualize_episodes_sentiment”, che riporta per ogni episodio di ogni stagione di The Vampire Diaries il valore di sentiment espresso sia come etichetta che come valore numerico, sia per Vader che per TextBlob.

Il file è composto da 171 righe e 6 colonne, in quanto ogni stagione contiene 22 episodi, tranne l’ottava che ne contiene 16, per un totale di 170 episodi, viene di sotto mostrata un’anteprima di questo file.

Season	Episode	Vader_sent_label	Textblob_sent_label	Vader_sent_numeric	Textblob_sent_numeric
Season01	episode_1.txt	positive	positive	0.0295	0.0155
Season01	episode_2.txt	positive	positive	0.0059	0.0222
Season01	episode_3.txt	positive	positive	0.0619	0.053
Season01	episode_4.txt	positive	positive	0.0431	0.0342
Season01	episode_5.txt	positive	positive	0.0466	0.0149
Season01	episode_6.txt	negative	positive	-0.0081	0.0296
Season01	episode_7.txt	negative	positive	-0.013	0.0056
Season01	episode_8.txt	positive	positive	0.0513	0.0211
Season01	episode_9.txt	negative	negative	-0.0147	0.0046
Season01	episode_10.txt	negative	positive	-0.0104	0.0056
Season01	episode_11.txt	positive	positive	0.0634	0.0207
Season01	episode_12.txt	positive	positive	0.0349	0.0073
Season01	episode_13.txt	positive	positive	0.0324	0.0109
Season01	episode_14.txt	positive	positive	0.0198	0.0006
Season01	episode_15.txt	positive	positive	0.0443	0.0208
Season01	episode_16.txt	positive	positive	0.0488	0.0291

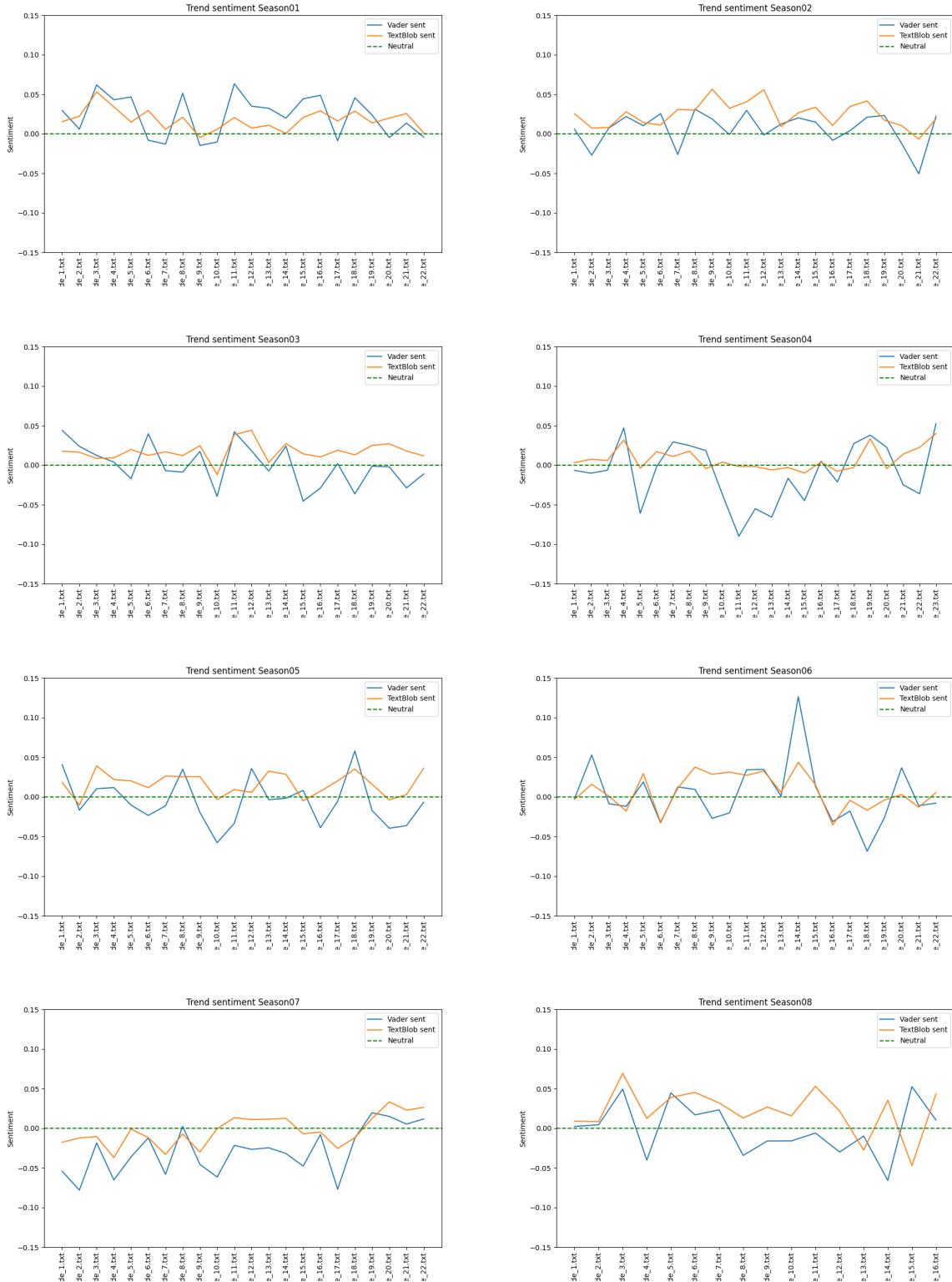
Analizzando ed esplorando i dati ottenuti, si è ritenuto che i risultati ottenuti con i due metodi, Vader TextBlob seguano lo stesso andamento, anche se i risultati ottenuti con il metodo Vader sembra che riescano a rispecchiare meglio il sentiment degli episodi, riuscendo a captare delle sfumature, che TextBlob non rileva; pertanto, nell’analisi che segue saranno utilizzati maggiormente i dati relativi al calcolo di Vader.

Per giustificare e spiegare questa scelta all’interno del file “2_Sentiment_graphs.py”, è stata definita una funzione che permette di effettuare il calcolo della correlazione tra i due metodi, “calculate_correlation(df)”, che utilizza per il calcolo i dati presenti nel file csv su riportato.

La correlazione restituita dalla funzione è di 0.51153, arrotondata a cinque cifre decimali. Questo valore indica che è presente una correlazione positiva, e dunque quando i valori di Vader tendono ad aumentare aumentano anche quelli di TextBlob e viceversa; quindi, queste variabili tendono a seguire lo stesso andamento, però non è una correlazione forte. Perciò nonostante i valori dei due metodi tendono a seguire la stessa tendenza nella gran parte dei casi, nella restante parte si discostano con delle differenze.

Per estrarre informazioni da questo file sugli episodi, e per comprendere l’andamento del sentiment nel corso delle puntate di ogni stagione, e per scoprire se è presente uno specifico trend si utilizza prima di tutto un grafico a linee, che risulta essere particolarmente informativo.

Attraverso la funzione “plot_episodes_line(df)”, che riporta sulle ascisse gli episodi della stagione e sulle ordinate il valore numerico del sentiment, viene creato un grafico per ogni stagione, in cui le due linee, una che utilizza i valori di Vader (linea blu) e l’altra i valori di Text Blob (linea arancione), mostrano l’andamento del sentiment. In questo modo si rende visibile anche la differenza nei valori dei due metodi. La linea orizzontale tratteggiata verde mostra il valore di neutralità, in modo da rendere più chiaro quando i sentiment oscillano tra positivo e negativo.



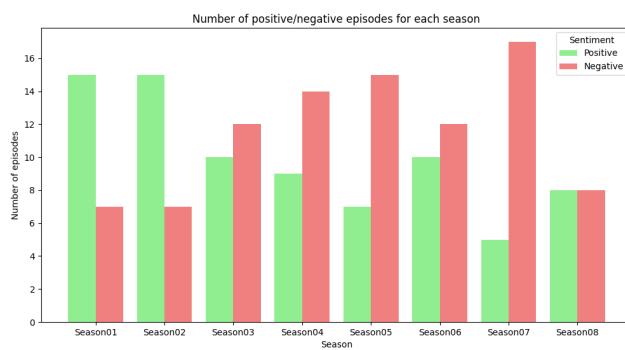
Si può notare, che come prima detto, le due linee tendono, con delle differenze, a seguire lo stesso andamento, anche se TextBlob tende a mantenersi su valori più alti e positivi.

Si può sicuramente osservare l'andamento oscillante del sentiment tra un episodio e l'altro, che è possibile notare quasi per tutte le stagioni, anche se sicuramente, la prima stagione è caratterizzata una maggiore positività del sentiment; per i valori di TextBlob le prime due stagioni risultano interamente positive, mentre si nota che la sesta stagione sembra fortemente caratterizzata da un sentiment negativo.

In generale, considerando i valori di Vader, in tutte le stagioni si alternano episodi positivi e negativi senza la possibilità di individuare uno specifico trend. Possiamo sicuramente dire che il sentiment, è molto variabile.

Osservano questi line plot si può sicuramente dire che le stagioni con sentiment più positivo sono la prima e la seconda, anche se non appare chiaro quale tra le due lo sia maggiormente, mentre non è altrettanto visibile e chiarissimo quale sia quella con il sentiment più negativo, ossia quella con maggior numero di episodi che riportano un sentiment negativo, in quanto le stagioni successive alla seconda riportano tutte una forte oscillazione dei valori.

Pertanto, per cercare di rendere più chiaro questo aspetto, è stata creata la funzione “plot_sentiment_for_episode_bar(df)” che crea un grafico a barre, il quale riporta sulle ascisse le stagioni della serie e sulle ordinate il conteggio degli episodi delle varie stagioni. Per ogni stagione, la barra in verde indica quanti episodi che riportano un sentiment positivo sono presenti all'interno della stagione mentre la barra rossa, quanti sono quelli dal sentiment negativo. Questa funzione permette, guardando un solo grafico, di comprendere il trend del numero di episodi positivi e negativi nel corso dell'intera serie, ma considerando la totalità degli episodi.



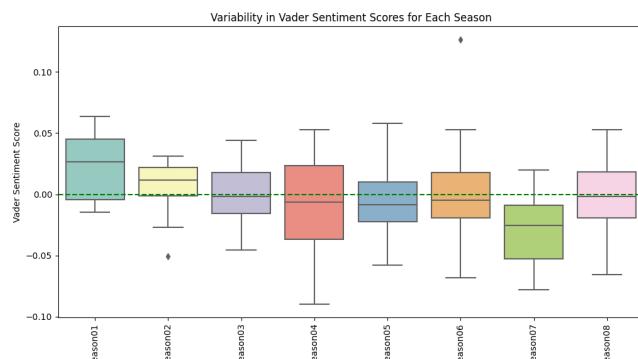
Ciò che emerge, è che, come abbiamo detto prima le prime due stagioni sono quelle che riportano maggiormente un sentiment positivo, le quali riportano esattamente lo stesso numero di episodi positivi e negativi. Successivamente, come ci si addentra nel corso delle stagioni, l'intreccio delle trame e degli eventi fanno crescere il sentiment negativo delle varie stagioni, che raggiunge il picco nella stagione sette. Nonostante nelle stagioni successive alla seconda il sentiment negativo è sempre predominante rispetto al positivo, tranne nell'ottava, i cui episodi si dividono esattamente a metà tra positivi e negativi, una buona parte di episodi negativi è presente anche nelle prime due stagioni.

Si può comunque affermare che ciò è pienamente giustificato dalla trama, dai temi e dalle vicende che scandiscono gli episodi di questa serie, un quanto costellata di combattimenti tra vampiri, presenza di personaggi fantastici misteriosi e cupi come licantropi e ibridi, sacrifici, maledizioni e rituali per scongiurarle. Pertanto, questo andamento del sentiment nel corso dello sviluppo della serie appare coerente.

Come abbiamo detto, la settima stagione è quella che appare più negativa, e per giustificare questo risultato, è possibile dire che i temi chiave di questa stagione sono: una grande battaglia che coinvolge molti personaggi con il conseguente esilio di Stefan (personaggio che sarà oggetto della sentiment analysis dei personaggi più avanti) da Mystic Falls e la depressione di Damon (altro personaggio che sarà analizzato più avanti) che lo porta alla decisione di chiudersi per anni all'interno di una bara. Tuttavia, sono presenti anche eventi positivi, come la storia d'amore di Stefan.

La caratteristica di questa serie, che è possibile sicuramente riscontrare nei dati ottenuti, è che ogni stagione è ricca di eventi, che si succedono molto rapidamente facendo variare continuamente e repentinamente il sentiment dei personaggi, senza che ci siano degli effettivi momenti di stallo che rendano stabile il sentiment per un certo numero di episodi, creando nello spettatore una tensione costante.

È stato dunque detto che la serie presenta una forte variabilità del sentiment, e per dimostrare ciò graficamente viene utilizzata la funzione “plot_vader_variation_by_season(df)” per produrre dei boxplot che mostrino la variabilità dei valori degli episodi per ogni stagione. In questo grafico vengono riportate sulle ascisse le stagioni della serie, mentre sulle ordinate i valori del sentiment, e viene riportata una linea verde tratteggiata orizzontale che permette di individuare il valore neutro, al di sopra della quale il sentiment assume valore positivo, mentre al di sotto negativo.



Questo grafico appare molto informativo e permette di confermare quanto detto precedentemente, ossia che le prime due stagioni sono quelle con valore del sentiment più positivo, in quanto presentano la maggior parte delle osservazioni al di sopra della linea della neutralità, mentre la scatola del boxplot della stagione sette si trova completamente nell’area al di sotto della linea della neutralità, con solo qualche osservazione al di sopra della stessa.

Notiamo invece che è la quarta stagione a riportare la maggiore variabilità nei valori numerici di sentiment, mentre per l’ottava e la terza stagione, è possibile dire che gli episodi si dividono quasi a metà tra osservazioni positive e negative.

In generale, comunque, questi box plot permettono di confermare l’ampia variabilità nei valori di sentiment, che si è riscontrata sia considerando i valori numerici sia i valori etichetta.

Alla luce di questi risultati, ora concentriamo l’analisi sul sentiment aggregato per stagione.

Per analizzare questi risultati ora utilizziamo il file “2_season_results.tvd.csv”, conservato sempre all’interno della directory “df”. In questo file è salvato il risultato restituito dalla funzione “visualize_season_sentiment(root_dir)” che riporta per ogni stagione di The Vampire Diaries il valore di sentiment espresso sia come etichetta che come valore numerico, sia per Vader che per TextBlob.

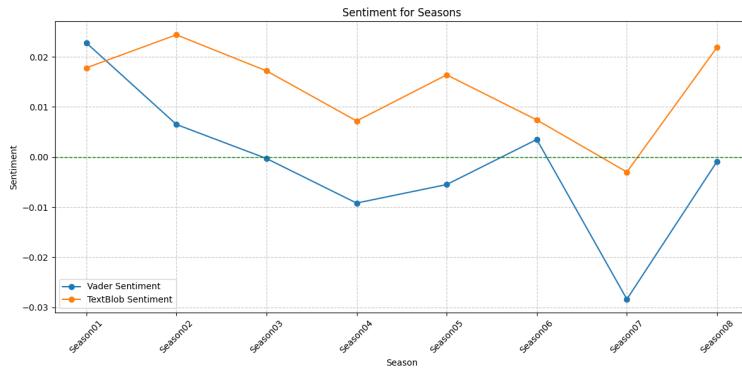
Il file è composto da 9 righe e 5 colonne, in quanto la serie è composta da otto stagioni; viene di sotto mostrata un’anteprima del file.

Season	Vader_sent_label	Textblob_sent_label	Vader_sent_numeric	Textblob_sent_numeric
Season01	positive	positive	0.0228	0.0178
Season02	positive	positive	0.0065	0.0244
Season03	negative	positive	-0.0003	0.0172
Season04	negative	positive	-0.0092	0.0072
Season05	negative	positive	-0.0055	0.0164
Season06	positive	positive	0.0035	0.0074
Season07	negative	negative	-0.0284	-0.003
Season08	negative	positive	-0.0009	0.0219

È stata usata la funzione “calculate_correlation(df)”, definita e spiegata precedentemente, per calcolare la correlazione dei valori del sentiment ottenuti con i due metodi contenuti in questo file.

L'output della funzione indica un valore della correlazione di 0.70365, dunque un risultato maggiore rispetto a quello ottenuto per il file suddiviso per episodi, che riflette una correlazione positiva; tuttavia, risulta facile trovare la causa della crescita di questo valore nell'aggregazione dei risultati.

Per analizzare l'andamento del sentiment aggregato nel corso delle stagioni, viene utilizzato un grafico a linee, prodotto dalla funzione “plot_sentiment_for_season_line(df)”, il quale riporta sulle ascisse le stagioni della serie e sulle ordinate i valori del sentiment e che permette di visualizzare lo sviluppo del sentiment calcolato con Vader (linea blu) e con TextBlob (linea arancione).

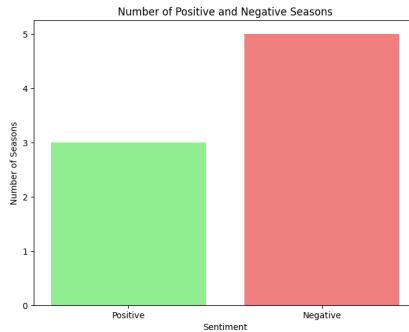


Vediamo che la differenza sostanziale tra i due metodi è la maggiore positività nei valori di TextBlob.

Prestando maggiore attenzione alla linea prodotta dai valori di Vader, notiamo e scorgiamo un trend ben definito: a partire dalla prima stagione, che presenta il valore del sentiment aggregato maggiore, il valore del sentiment aggregato decresce, diventando negativo, ma prossimo al valore neutro nella terza stagione e continuando a decrescere fino alla quinta stagione; successivamente risalire assumendo valore positivo nella sesta stagione e infine decresce nuovamente, raggiungendo un picco di negatività nella settima stagione per poi risalire nell'ultima stagione assumendo un valore negativo ma prossimo alla neutralità.

Ovviamente, le differenze che si riscontrano tra questi risultati e quelli ottenuti nell'analisi per episodio, sono dovuti al fatto che aggregando tutti gli episodi si perde quella specificità e capacità esplicativa che si aveva nell'analisi episodio per episodio. Inoltre, mentre i dati numerici assumono valori, che possono essere più o meno elevati, la trasformazione in etichetta positiva e negativa rende equivalenti stagioni, che non lo sono altrettanto considerando i valori numerici. Infatti, mentre precedentemente abbiamo rilevato che le prime due stagioni presentavano lo stesso numero di episodi positivi ed episodi negativi, e nonostante comunque avessimo riscontrato la maggiore positività nei valori della prima stagione, questo grafico che considera i valori numerici aggregati permette di visualizzare a colpo d'occhio le differenze nel sentiment, fornendo una nuova visione globale e d'insieme della serie.

Infine, per visualizzare globalmente quante stagioni presentano un sentiment aggregato positivo e quante un sentiment aggregato negativo, utilizziamo la funzione “plot_season_bar(df)” che produce un grafico a barre che riporta questo semplice conteggio.



Quindi, l'analisi del sentiment condotta ci fa notare che la serie presenta tre stagioni con valore del sentiment aggregato positivo e cinque stagioni con valore del sentiment aggregato negativo.

Dunque, per concludere, dall'analisi del sentiment della serie “The Vampire Diaries” emergono una forte variabilità nei valori del sentiment che comportano un’oscillazione dello stesso lungo tutte le stagioni della serie, e la forte presenza di episodi con sentiment negativo che successivamente a livello aggregato, creano come risultato una maggioranza di stagioni con sentiment negativo.

Questi risultati sono ampiamente spiegati dalla tipologia di serie tv in analisi, che, come detto in precedenza, ha l’obbiettivo di creare suspense insidiando nello spettatore una costante tensione causata dal cambiamento repentino di eventi e situazioni. Essendo una serie soprannaturale, misteriosa e horror, gli eventi che scandiscono la serie, come le lotte e gli scontri costanti, producono come risultato maggiori valori del sentiment negativo, d’altra parte non manca il verificarsi di episodi che generano invece del sentiment positivo, soprattutto legati all’intrecciarsi delle relazioni sentimentali che accompagnano tutta la serie e hanno estrema rilevanza in quanto questa nasce e si sviluppa attorno alla storia romantica che vede i due vampiri protagonisti Stefan e Damon Salvatore innamorati della medesima ragazza Elena Gilbert. Questi tre personaggi saranno i protagonisti della Sentiment Analysis dei personaggi su “The Vampire Diaries” che verrà sviluppata successivamente.

2.3.2. Analisi dei risultati: “The Big Bang Theory”

L'analisi del sentiment della serie “The Big Bang Theory” è appunto stata svolta, come per “The Vampire Diaries” su due livelli, prima in modo più specifico, esaminando il sentiment di ogni singolo episodio e successivamente, a livello più aggregato calcolando il sentiment di ogni stagione.

Per prima cosa concentriamo l'analisi sul sentiment di ogni episodio.

Per visualizzare i risultati è possibile aprire il file “2_episode_result.tbbt.csv”, contenuto nella directory “df”, nel quale è salvato l’output restituito dalla funzione “visualize_episodes_sentiment”, che riporta per ogni episodio di ogni stagione della serie televisiva il valore di sentiment espresso sia come etichetta che come valore numerico, sia per Vader che per TextBlob.

Il file è composto da 232 righe e 6 colonne, in quanto ogni stagione contiene 24 episodi, tranne la seconda e la terza che ne contengono 23, e la prima che ne contiene solo 17, per un totale di 231 episodi, viene di sotto mostrata un’anteprima di questo file.

```

Season,Episode,Vader_sent_label,Textblob_sent_label,Vader_sent_numeric,Textblob_sent_numeric
Season01,episode_1.txt,positive,positive,0.891,0.085
Season01,episode_2.txt,positive,positive,0.0564,0.0561
Season01,episode_3.txt,positive,positive,0.0712,0.0662
Season01,episode_4.txt,positive,positive,0.1116,0.0782
Season01,episode_5.txt,positive,positive,0.1466,0.0987
Season01,episode_6.txt,positive,positive,0.0827,0.0468
Season01,episode_7.txt,positive,positive,0.0133,0.0533
Season01,episode_8.txt,positive,positive,0.0989,0.0512
Season01,episode_9.txt,positive,positive,0.0827,0.0583
Season01,episode_10.txt,positive,positive,0.0211,0.0308
Season01,episode_11.txt,positive,negative,0.0341,-0.0145
Season01,episode_12.txt,positive,positive,0.0893,0.0564

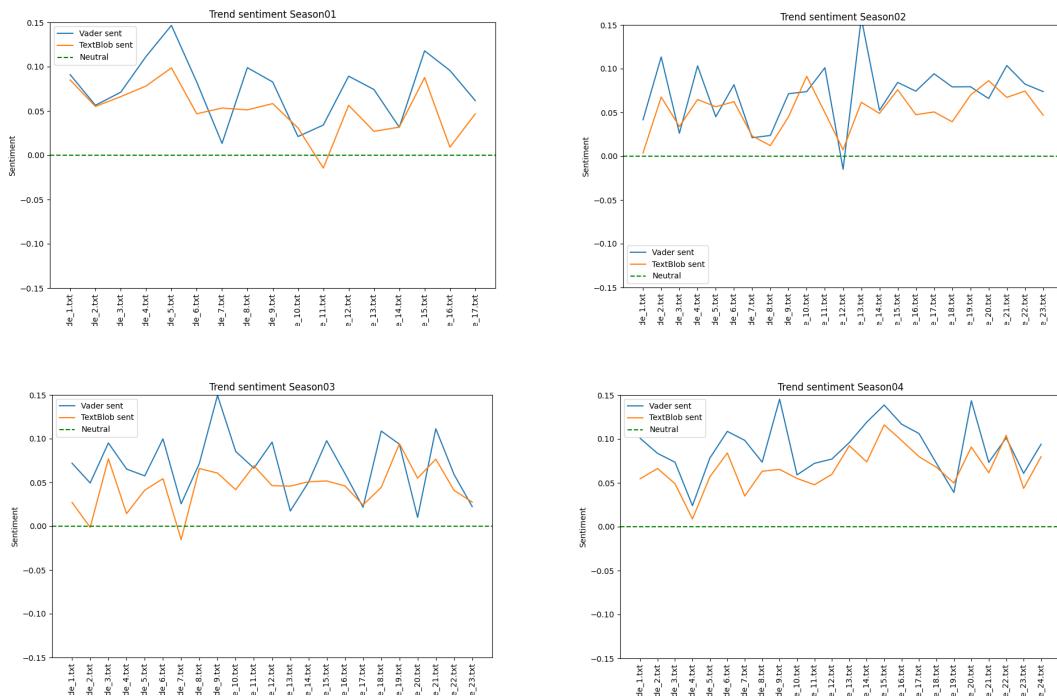
```

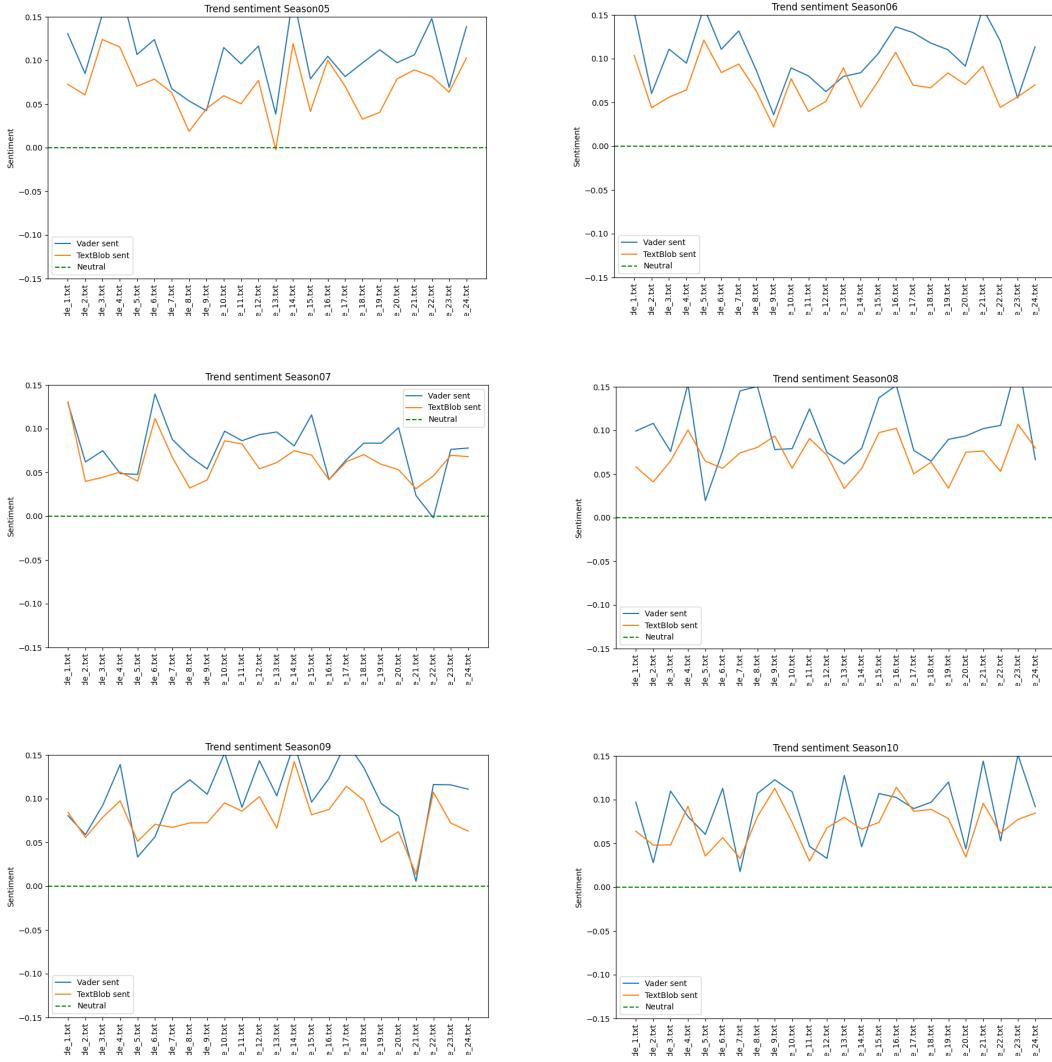
Analizzando ed esplorando i dati ottenuti, si è ritenuto che i risultati ottenuti con i due metodi, Vader e TextBlob seguano lo stesso andamento positivo.

All'interno del file “2_Sentiment_graphs.py”, la funzione “calculate_correlation(df)” effettua il calcolo della correlazione tra i due metodi utilizzando per il calcolo i dati presenti nel file csv su riportato, restituendo una correlazione pari a 0.71941, arrotondata a cinque cifre decimali. Questo valore indica che è presente una correlazione tendenzialmente positiva, e dunque quando i valori di Vader tendono ad aumentare aumentano anche quelli di TextBlob e viceversa; sottolineando come queste variabili tendano a seguire lo stesso andamento.

Per estrarre informazioni da questo file sugli episodi, e per comprendere l'andamento del sentimento nel corso delle puntate di ogni stagione, e per scoprire se è presente uno specifico trend si utilizza prima di tutto un grafico a linee, che risulta essere particolarmente informativo.

Come fatto per “The Vampire Diaries”, attraverso la funzione “plot_episodes_line(df)” che riporta sulle ascisse gli episodi della stagione e sulle ordinate il valore numerico del sentimento, viene creato un grafico per ogni stagione, in cui le due linee, una che utilizza i valori di Vader (linea blu) e l'altra i valori di Text Blob (linea arancione), mostrano l'andamento del sentimento. In questo modo si rende visibile anche la differenza nei valori dei due metodi. La linea orizzontale tratteggiata verde mostra il valore di neutralità, in modo da rendere più chiaro quando il sentimento oscilla tra positivo e negativo.





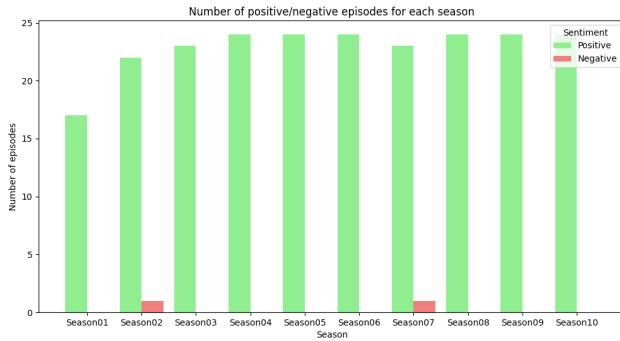
Si può notare, che come prima detto, le due linee tendono, con delle differenze, a seguire lo stesso andamento, anche se Vader tende a mantenersi su valori decisamente più alti e positivi, contrariamente a quanto accadeva nella precedente analisi descritta per “The Vampire Diaries”, la cui linea arancione di TextBlob tendeva ad essere maggiormente positiva.

Si può sicuramente osservare l’andamento oscillante del sentiment tra un episodio e l’altro, sempre rimanendo al di sopra della linea di neutralità per tutte le stagioni. Come si può notare dai grafici, tuttavia, le prime stagioni sono caratterizzate da singoli episodi negativi, di fatto TextBlob presenta l’ep.11 della stagione 1 e l’ep.7 della stagione 3, con valori al di sotto la linea verde della neutralità, mentre Vader scende sotto la linea neutrale solo nell’ep.12 stagione 2. Tutte le restanti puntate per tutte le stagioni sono sempre categorizzate dalla positività in entrambi i metodi.

Si può dire quindi che il trend sentiment è decisamente positivo in quanto i valori degli episodi in ogni stagione sono sempre più o meno positivi.

Osservando questi line plot si può notare che le stagioni con sentiment meno positivo sono le prime tre, secondo i risultati restituiti da TextBlob, mentre quella con il sentiment più positivo, ossia quella con maggior numero di episodi che riportano un sentiment positivo, sembra essere la quinta, la sesta e la nona, in quanto si può notare che la linea blu di Vader (ma anche quella arancione di TextBlob rispetto alle prime stagioni) è spostata molto verso l’alto, ma soprattutto si trovano molto più distanti dalla linea verde di neutralità; tuttavia in generale la linea arancione di TextBlob che si trova quasi sempre più vicina alla neutralità e a valori positivi più bassi.

Per avere una visione più generale delle stagioni è stata creata la funzione “plot_sentiment_for_episode_bar(df)” che crea un grafico a barre, il quale riporta sulle ascisse le stagioni della serie e sulle ordinate il conteggio degli episodi delle varie stagioni. Per ogni stagione, la barra in verde indica quanti episodi che riportano un sentiment positivo sono presenti all'interno della stagione mentre la barra rossa, quanti sono quelli dal sentiment negativo. Questa funzione permette, guardando un solo grafico, di comprendere il trend del numero di episodi positivi e negativi nel corso dell'intera serie, ma considerando la totalità degli episodi.



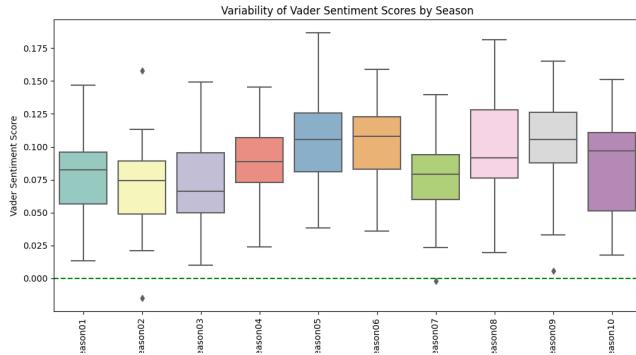
Ciò che emerge, è che, come abbiamo detto prima, quasi tutte le stagioni sono positive, in linea con il genere leggero della sitcom. Dal grafico emergono a parità di episodi negativi, che solo pochi di questi vengono catalogati come negativi nelle stagioni 2 e 7. Questo può essere giustificato in quanto all'inizio della serie (e anche nella seconda stagione quindi) Leonard e Penny affrontano alcune sfide nella loro relazione a causa delle loro differenze, attraversando spesso momenti di separazione. Nella settima stagione invece il grafico può essere giustificato dalla pausa temporanea nella relazione tra Sheldon e Amy; Howard invece si trova a gestire la paura del volo quando parte per una missione spaziale, mentre Leonard affronta problemi professionali che influiscono sulla fiducia in sé stesso.

Tuttavia, come si può notare in generale dal grafico, il trend è decisamente positivo, ciò è giustificato dal fatto che essendo The Big Bang Theory una sitcom, l'obiettivo principale è far ridere il pubblico; la comicità, le battute, e le situazioni esilaranti contribuiscono infatti a generare un'atmosfera generale positiva e leggera.

È stato dunque detto che la serie presenta una poca variabilità del sentiment, che rimane quasi sempre positiva, come si può vedere nel boxplot prodotto dalla funzione “plot_vader_variation_by_season(df)”. In questo grafico vengono riportate sulle ascisse le stagioni della serie, mentre sulle ordinate i valori del sentiment, e viene riportata una linea verde tratteggiata orizzontale che permette di individuare il valore neutro, al di sopra della quale il sentiment assume valore positivo, e specularmente al di sotto diventa negativo.

Questo grafico permette di confermare quanto detto precedentemente, ossia che le stagioni sono tutte positive, di fatto i box delle stagioni si trovano tutti sopra la linea verde di neutralità. Solo la seconda e la settima sono quelle con valore del sentiment positivo più basso (come rilevato prima nel grafico a barre), di fatto dopo la seconda stagione i box incrementano il proprio punteggio di positività per poi decrescere nella stagione sette e riprendere a risalire nell'ottava.

Dai box sembra che siano le stagioni 5 e 8 ad avere maggior positività e variabilità di punteggio. Come nei grafici precedenti anche qui si confermano outlier negativi nella stagione 2 e nella 7.



Alla luce di questi risultati, ora concentriamo l'analisi sul sentiment aggregato per stagione.

Per analizzare questi risultati ora utilizziamo il file “2_season_results.tbbt.csv”, conservato sempre all'interno della directory “df”. In questo file è salvato il risultato restituito dalla funzione “visualize_season_sentiment” che riporta per ogni stagione di “The Big Bang Theory” il valore di sentiment espresso sia come etichetta che come valore numerico, sia per Vader che per TextBlob.

Il file è composto da 11 righe e 5 colonne, in quanto la serie è composta da dieci stagioni; viene di sotto mostrata un'anteprima del file.

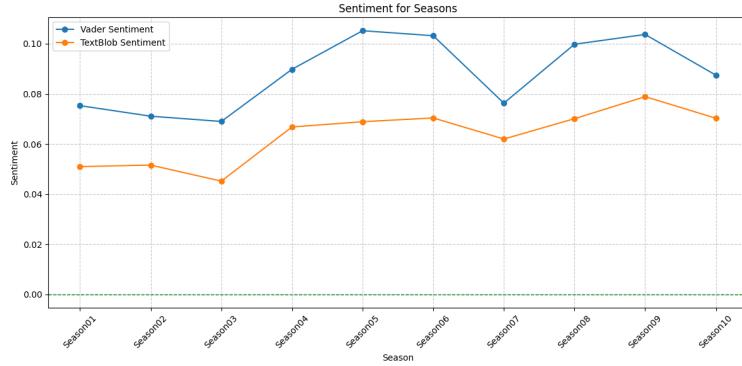
Season	Vader_sent_label	Textblob_sent_label	Vader_sent_numeric	Textblob_sent_numeric
Season01	positive	positive	0.0753	0.051
Season02	positive	positive	0.0711	0.0516
Season03	positive	positive	0.069	0.0452
Season04	positive	positive	0.0898	0.0668
Season05	positive	positive	0.1052	0.0689
Season06	positive	positive	0.1832	0.0704
Season07	positive	positive	0.0763	0.062
Season08	positive	positive	0.0998	0.0701
Season09	positive	positive	0.1037	0.0789
Season10	positive	positive	0.0875	0.0703

È stata usata la funzione “calculate_correlation(df)”, definita e spiegata precedentemente, per calcolare la correlazione dei valori del sentiment ottenuti con i due metodi contenuti in questo file.

L'output della funzione indica un valore della correlazione di 0.89088, dunque un risultato maggiore rispetto a quello ottenuto per il file suddiviso per episodi, che riflette una correlazione positiva; tuttavia, risulta facile trovare la causa della crescita di questo valore nell'aggregazione dei risultati.

Per analizzare l'andamento del sentiment aggregato nel corso delle stagioni e considerando i valori numerici aggregati, il grafico permette di visualizzare a colpo d'occhio le differenze nel sentiment, fornendo una nuova visione globale e d'insieme della serie. Quindi viene utilizzato un grafico a linee, prodotto dalla funzione “plot_sentiment_for_season_line(df)”, il quale riporta sulle ascisse le stagioni della serie e sulle ordinate i valori del sentiment e che permette di visualizzare lo sviluppo del sentiment calcolato con Vader (linea blu) e con TextBlob (linea arancione).

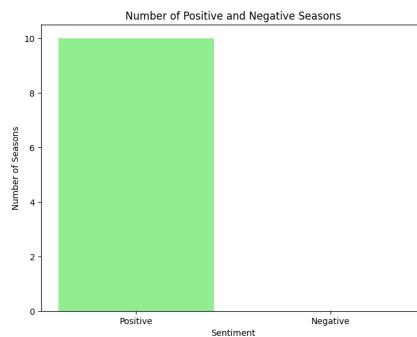
Vediamo che la differenza sostanziale tra i due metodi è la maggiore positività nei valori di Vader.



Prestando maggiore attenzione alla linea prodotta dai valori di Vader, notiamo e scorgiamo un trend ben definito: a partire dalla prima stagione, il valore del sentiment aggregato decresce fino alla terza stagione per poi risalire fino al la quinta, che, come visto nei precedenti grafici, sembra essere la più positiva, per poi decrescere nuovamente fino alla settima, come visto essere tra le meno positive, risalire subito nell'ottava e nella nona per poi decrescere nella decima. Anche la linea arancione di TextBlob segue lo stesso identico andamento pur rimanendo con valori leggermente più bassi.

Si può notare in questo grafico, coerentemente con gli altri grafici, che le stagioni meno positive risultino ancora essere la seconda e la settima, mentre quelle più positive sono la quinta e la nona.

Infine, per visualizzare globalmente quante stagioni presentano un sentiment aggregato positivo e quante un sentiment aggregato negativo, è stata definita la funzione “plot_season_bar(df)” che produce un grafico a barre che riporta questo semplice conteggio. Come previsto, l’analisi del sentiment condotta ci porta a concludere che la serie presenta mediamente tutte le stagioni con un valore del sentiment aggregato totalmente positivo, e nessuna con valore del sentiment aggregato negativo.



Dunque, per concludere, dall’analisi del sentiment della serie “The Big Bang Theory”, emergono deboli variabilità nei valori del sentiment, con una forte presenza di episodi con sentiment positivo che successivamente a livello aggregato, creano come risultato una totalità di stagioni con sentiment positivo.

Come per “The Vampire Diaries” questi risultati sono ampiamente giustificati dalla tipologia della serie, che in questo caso ha l’obbiettivo di creare leggerezza e divertire gli spettatori, con temi di vita quotidiana, scientifica e d’amicizia, dando un intrattenimento piacevole senza drammi o tensioni pesanti, al contrario di “The Vampire Diaries”.

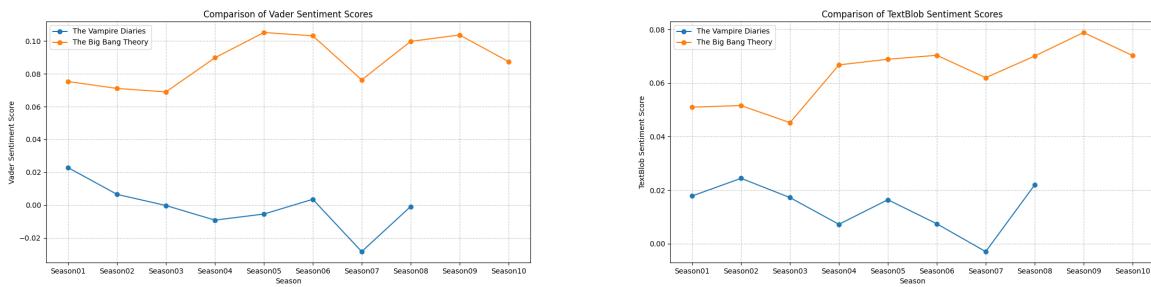
Il tema dell’amicizia, in particolare, è un tema centrale in “The Big Bang Theory”, la differenza tra gli stili di vita scientifici di Sheldon e Leonard e quello comune di Penny è molto trattato all’interno della serie. Questi tre personaggi, infatti, saranno i protagonisti della Sentiment Analysis riguardante i personaggi della serie che verrà sviluppata successivamente.

2.4 Sentiment Analysis delle serie a confronto

Dopo aver esaminato e analizzato il sentiment delle due serie tv nello specifico e aver dato spiegazione ai risultati ottenuti, possiamo ora effettuare un confronto tra le due, per comprendere meglio se emergono delle rilevanti differenze e per quale ragione.

Per far ciò utilizziamo i file “2_season_results_tvd.csv” e “2_season_results.tbbt.csv” che contengono i valori del sentiment aggregati per ogni stagione.

Pertanto confrontiamo l’andamento del sentiment delle due serie, utilizzando le funzioni “seasons_tvd_vs_tbbt_vader(df_tvd,df_tbbt)” e “seasons_tvd_vd_tbbt_textb(df_tvd,df_tbbt)” che utilizzano i valori numerici la prima di Vader e l’altra di TextBlob per creare dei grafici a linee che permettano di mettere a confronto l’andamento del sentiment aggregato lungo le stagioni delle due serie. La linea blu mostra l’evoluzione del sentiment per “The Vampire Diaries” mentre la linea arancione rappresenta quella per “The Big Bang Theory”.



Si può subito dire che tra i due metodi non si notano grandi differenze, anzi, per valori aggregati riportano degli andamenti molto simili.

Concentrandosi invece sui valori di sentiment delle due serie la differenza che emerge è ben visibile: mentre “The Vampire Diaries” presenta sia valori, e dunque stagioni, che stanno sia al di sopra che al di sotto dello zero, e dunque presenta sia valori positivi che negativi, la linea di “The Big Bang Theory” rimane ben al di sopra sia dello zero sia dell’intera linea di “The Vampire Diaries”, dunque anche i valori positivi di quest’ultima sono in ogni caso inferiori a quelli riportati da “The Big Bang Theory”.

Nonostante ci si potesse aspettare questa differenza nei risultati, trattandosi di serie tv dai generi che potremmo definire opposti, questa analisi ha permesso di condurre un’esplorazione delle serie ancora più in profondità portando a risultati interessanti.

Le due serie, infatti, sono intrinsecamente differenti e richiedono anche uno spettatore differente. Mentre per “The Vampire Diaries” lo spettatore deve essere attento e concentrato per districarsi tra gli innumerevoli eventi e accadimenti che costellano tutte le stagioni, “The Big Bang Theory” si presenta come una serie leggera, dai temi comici e per i quali non è richiesta la piena attenzione dello spettatore per non perdersi all’interno della storia. Ulteriore differenza sta nella “staticità” delle due serie, “The Vampire Diaries” si svolge in luoghi sempre diversi, nuove città, nuovi luoghi particolari e misteriosi e dunque i personaggi sono in costante movimento; al contrario “The Big Bang Theory” presenta la maggior parte degli episodi che si svolgono sempre negli stessi luoghi, senza che ci sia particolare dinamicità.

La differenza principale, tuttavia, è il genere di appartenenza della serie, che ci porta a concludere, dati i risultati, come questo influenzi il sentiment. “The Vampire Diaries” infatti risulta essere la serie con il sentiment più negativo, dati i temi principali della morte, del dolore e del conflitto, causanti eventi drammatici, perdite di personaggi amati o situazioni pericolose, generando di conseguenza tensioni e angoscia emotiva. Tutto ciò ovviamente si riflette nel sentiment dei personaggi. Dall’altra parte invece “The Big Bang Theory” è la serie caratterizzata dalla predominanza del sentiment positivo, risultato comprensibile dato dal genere sitcom della commedia, presentata con situazioni strampalate e interazioni esilaranti tra i personaggi, creati al fine di far

divertire gli spettatori. Il genere comico e leggero della serie, quindi, contribuisce alla definizione di un sentimento decisamente positivo.

In conclusione, la Sentiment Analysis delle serie nella loro totalità ha permesso di individuare come si muove il sentimento all'interno delle varie stagioni e successivamente di avere una panoramica globale del sentimento.

Il sentimento ora verrà analizzato nel dettaglio prendendo in analisi tre personaggi per ciascuna serie, Stefan, Damon e Elena per "The Vampire Diaries" e Sheldon, Leonard e Penny per "The Big Bang Theory", scelti specularmente in questo caso: due vampiri e un'umana nella prima serie, due scienziati e un'attrice nella seconda, al fine di visualizzare meglio la differenza del sentimento tra i personaggi anche in relazione alle loro caratteristiche personali anche all'interno della propria serie stessa.

PUNTO 3: SENTIMENT ANALYSIS DEI PERSONAGGI

Dopo aver condotto un'analisi accurata e dettagliata del sentiment nelle due serie TV nel loro complesso, con l'obiettivo di identificare trend ed evoluzioni nel corso delle stagioni, ci apprestiamo ora a esplorare un livello di analisi più approfondito. In particolare, ci concentreremo sull'analisi dei sentiment dei personaggi principali delle due serie al fine di comprendere come si evolvono nel corso della narrazione e se emergono differenze significative tra di loro. Questa analisi ci consentirà di gettare ulteriore luce sulla complessità delle dinamiche dei personaggi e sul modo in cui le loro esperienze influenzano il loro stato emotivo durante il corso delle stagioni, arricchendo così la nostra comprensione delle serie TV in esame.

3.1 Il codice

Per questa analisi è stato creato il file “3_SentAnalysis_characters.py”

Per condurre questa analisi, è necessario calcolare il sentiment di ciascun personaggio tenendo in considerazione esclusivamente le battute di quel personaggio.

Pertanto, è stato creato un nuovo corpus, per ogni serie tv, che contiene per ogni personaggio, tutte le battute estratte del personaggio, per tutte le stagioni e tutti gli episodi, in modo tale da avere una struttura dati che permetta di agevolare l'analisi.

Questo ruolo è svolto dalla funzione “create_character_corpus(character_1,character_2,character_3,corpus)”, che prende in ingresso appunto i nomi dei tre personaggi, e il corpus originale nel quale iterare per estrarre le battute dei personaggi.

Ottenuta la struttura dati principale, viene svolta l'analisi del sentiment per ogni personaggio prima episodio per episodio e successivamente in modo aggregato per stagione, così come è stato fatto nel punto precedente.

Per ottenere il sentiment episodio per episodio viene usata la funzione “character_episodes_sentiment(root_dir,characters,custom_words,serie_name)”, che itera tra i personaggi e nel corpus creato dei personaggi, per calcolare il sentiment, sia numerico, sia categorico sia per Vader che per TextBlob, utilizzando la stessa modalità e le stesse funzioni del punto precedente:

- ‘Clean(episode,custom_words)’: per effettuare la pulizia del testo, trasformare il testo in minuscolo, e eliminare punteggiatura, stopwords, ‘custom_words’, parole con meno di tre caratteri, le battute vuote e infine utilizzare un lemmatizzatore;
- ‘Sentiment_for_episode_vader(cleaned_text)’: che calcola il valore del sentiment di un episodio utilizzando Vader;
- ‘Sentiment_for_episode_textblob(cleaned_text)’: che calcola il valore del sentiment di un episodio utilizzando TextBlob;
- ‘Assign_sentiment_label(value)’: che trasforma il valore numerico del sentiment restituito dalle due funzioni precedenti in valore categorico.

Utilizzando queste funzioni all'interno di “character_episodes_sentiment(root_dir,characters,custom_words,serie_name)”, è stato possibile calcolare dunque il sentiment episodio per episodio e successivamente salvare i risultati ottenuti all'interno di un dataframe che viene inserito nella directory “df”.

Questo dataframe contiene le colonne:

- “Character”: il personaggio;
- “Season”: la stagione che contiene l'episodio del personaggio;
- “Episode”: l'episodio;

- “Vader_sent_label”: etichetta di sentiment dell’episodio per il personaggio indicato, calcolata con Vader;
- “Textblob_sent_label”: etichetta di sentiment dell’episodio per il personaggio indicato, calcolata con TextBlob;
- “Vader_sent_numeric”: valore numerico di sentiment dell’episodio per il personaggio indicato, calcolato con Vader;
- “Textblob_sent_numeric”: valore numerico di sentiment dell’episodio per il personaggio indicato, calcolato con TextBlob;

Calcolato il sentiment per ogni personaggio episodio per episodio, si è proceduto calcolando il sentiment di ogni personaggio a livello aggregato per ogni serie tv.

La funzione “character_season_sentiment(root_dir,characters,custom_words,serie_name)” assolve a questo compito, iterando tra i personaggi e nel corpus dei personaggi, calcola il sentiment per ogni episodio e successivamente calcola la media per stagione.

Tutti i risultati vengono salvati in un dataframe che viene memorizzato all’interno delle directory “df”, contenenti le colonne:

- “Character”: il personaggio;
- “Season”: la stagione;
- “Vader_sent_label”: etichetta di sentiment della stagione per il personaggio indicato, calcolata con Vader;
- “Textblob_sent_label”: etichetta di sentiment della stagione per il personaggio indicato, calcolata con TextBlob;
- “Vader_sent_numeric”: valore numerico di sentiment della stagione per il personaggio indicato, calcolato con Vader;
- “Textblob_sent_numeric”: valore numerico di sentiment della stagione per il personaggio indicato, calcolato con TextBlob.

3.2 Analisi dei risultati ottenuti

Una volta completata la fase di raccolta dei risultati della Sentiment Analysis mediante il codice precedentemente descritto, e dopo aver raccolto i risultati dell’analisi all’interno di appositi file csv archiviati nella directory “df”, sfruttiamo questi file per condurre un’analisi dettagliata e accurata dei risultati ottenuti. Verranno utilizzati anche dei grafici per rendere l’analisi più chiara, comprensibile e informativa. Questa approfondita esplorazione dei dati ci consentirà di acquisire una comprensione più approfondita dei sentiment dei vari personaggi e di individuare eventuali pattern significativi nei loro sviluppi emotivi lungo l’arco delle stagioni.

3.2.1 Analisi dei risultati: “The Vampire Diaries”

Anche per questo punto l’analisi sarà svolta su più livelli prima in modo più specifico, esaminando il sentiment di ogni singolo episodio e successivamente, a livello più aggregato calcolando il sentiment di ogni stagione.

Conduciamo ora l’analisi del sentiment dei tre personaggi Stefan, Damon ed Elena considerando i risultati sul sentiment che sono stati calcolati episodio per episodio.

Per far ciò viene utilizzato il file “3_episode_results_tvd.csv” che riporta il valore del sentiment sia espresso in valore numerico sia come etichetta per ogni episodio di ogni stagione e per ogni personaggio, sia calcolato con Vader che con TextBlob.

Questo file ha 513 righe e 7 colonne, e ne viene riportata sotto un’anteprima:

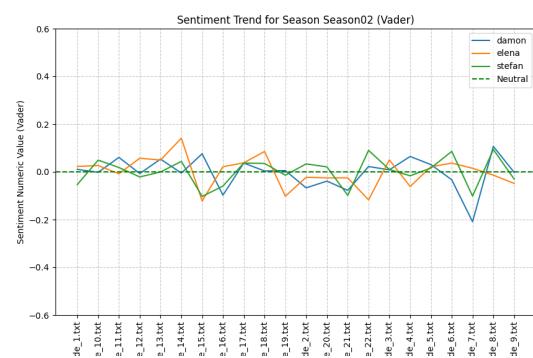
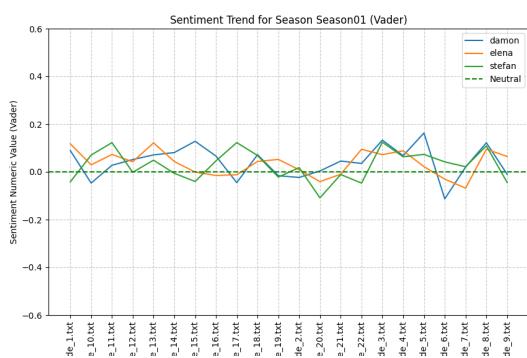
Character	Season	Episode	Vader_sent_label	Textblob_sent_label	Vader_sent_numeric	Textblob_sent_numeric
stefan	Season01	episode_1.txt	negative	-0.0418	-0.0238	
stefan	Season01	episode_2.txt	positive	0.0178	-0.0009	
stefan	Season01	episode_3.txt	positive	0.1242	0.1614	
stefan	Season01	episode_4.txt	positive	0.0631	0.0139	
stefan	Season01	episode_5.txt	positive	0.0732	-0.005	
stefan	Season01	episode_6.txt	positive	0.0418	0.0164	
stefan	Season01	episode_7.txt	positive	0.0215	0.0251	
stefan	Season01	episode_8.txt	positive	0.1087	0.0441	
stefan	Season01	episode_9.txt	negative	0.0443	0.0131	
stefan	Season01	episode_10.txt	positive	0.0763	-0.0062	
stefan	Season01	episode_11.txt	positive	0.1224	0.059	
stefan	Season01	episode_12.txt	negative	0.0019	-0.019	
stefan	Season01	episode_13.txt	positive	0.0491	0.0179	
stefan	Season01	episode_14.txt	negative	0.0062	-0.0247	
stefan	Season01	episode_15.txt	negative	-0.0404	-0.011	

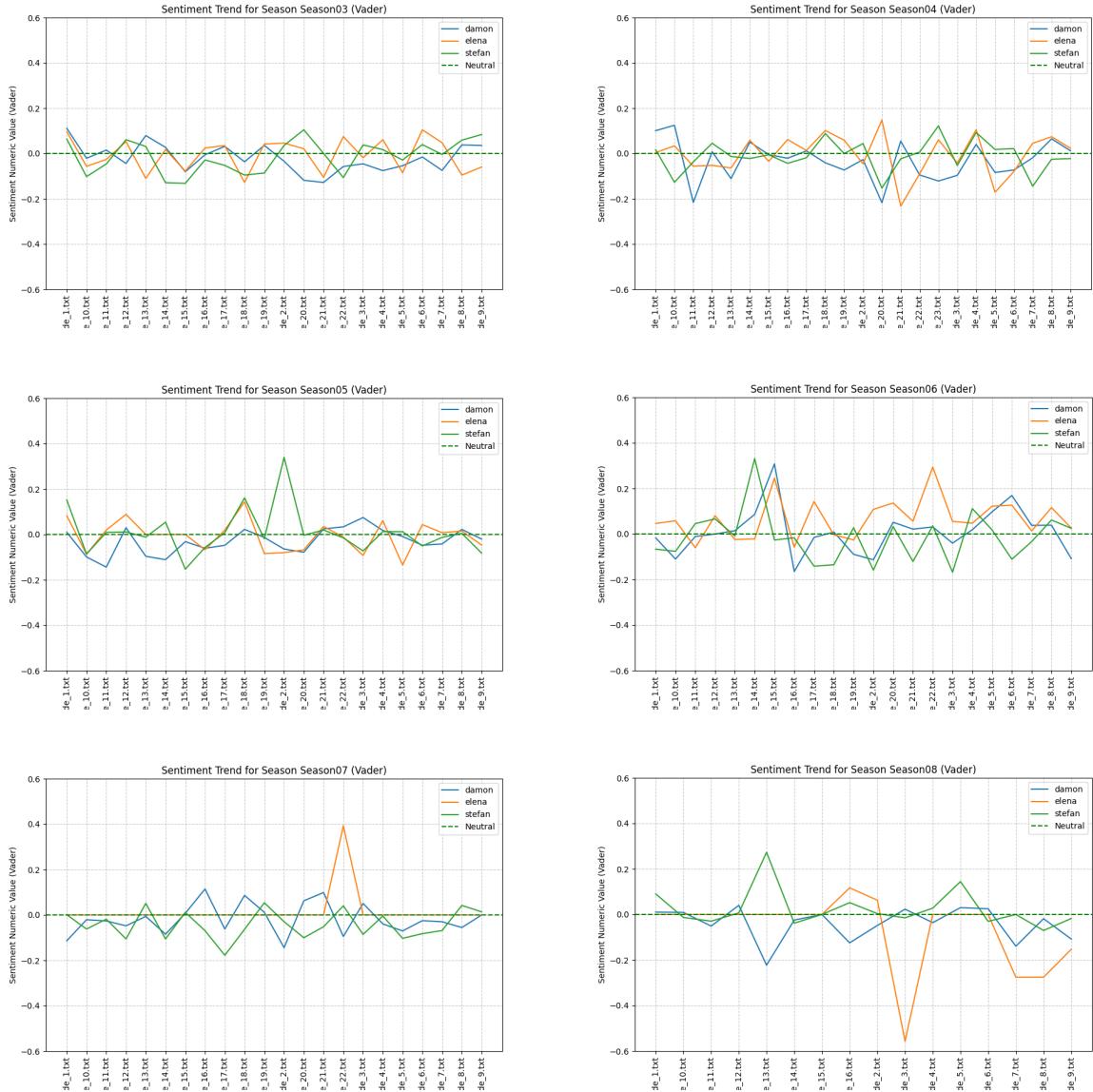
In questo studio, è stato ritenuto importante valutare la correlazione tra i valori restituiti da Vader e quelli da TextBlob per determinare se entrambi i metodi mostrano un andamento simile nell'analisi del sentiment. Per effettuare questa analisi, è stata utilizzata la funzione "calculate_correlation(df)" contenuta nel file "3_Sentiment_characters_graphs.py", che ha restituito un valore di correlazione pari a 0.36674. Questo valore indica una correlazione positiva moderata tra le due variabili prese in considerazione. In altre parole, c'è una tendenza positiva tra i risultati ottenuti dai due metodi: quando uno dei metodi rileva un aumento nel sentiment, anche l'altro tende ad aumentare, e viceversa. Tuttavia, questa associazione non è così forte da essere considerata altamente predittiva.

È interessante notare che tra i due metodi, riteniamo che i valori restituiti da Vader riflettano una maggiore variazione del sentiment. Questo è attribuibile al fatto che Vader è in grado di rilevare sfumature più sottili nel sentiment rispetto a TextBlob, il quale tende a fornire valori di sentiment più neutri e meno sensibili alle sfumature linguistiche.

Successivamente, l'analisi verrà condotta attraverso il supporto di grafici, le cui funzioni per realizzarli sono contenute all'interno del file "3_Sentiment_characters_graphs.py".

Per ottenere una visione dettagliata dell'andamento del sentiment dei personaggi episodio per episodio, abbiamo utilizzato la funzione "sentiment_line_for_episodes(df)". Questa funzione ha permesso di creare grafici a linee per ciascuna stagione, utilizzando i valori numerici del sentiment calcolati con Vader. Sugli assi dei grafici sono stati riportati gli episodi di ogni stagione (ascisse) e i valori del sentiment (ordinate). Ogni linea nel grafico rappresenta l'evoluzione del sentiment nel corso della stagione per un personaggio specifico: la linea blu è associata a Damon, la linea arancione a Elena e la linea verde a Stefan. Inoltre, è presente anche una linea tratteggiata orizzontale verde che coincide con la neutralità, e permette di comprendere quali episodi sono al di sopra della linea e dunque positivi. Questa rappresentazione visiva ci consente di analizzare in modo approfondito come il sentiment dei personaggi si sviluppa nel corso della serie, offrendo un'ulteriore comprensione delle dinamiche emotive dei protagonisti.





Dai grafici presentati, emergono alcune osservazioni chiave riguardo al sentimento dei tre personaggi principali, ovvero Damon, Elena e Stefan, nel corso delle stagioni di "The Vampire Diaries". In generale, si nota che il sentimento è caratterizzato da una notevole oscillazione, e non è possibile individuare un trend specifico nel comportamento emotivo dei personaggi nel corso della serie. Tuttavia, ci sono alcune tendenze e punti salienti che possono essere identificati.

Nelle prime due stagioni della serie, il sentimento dei personaggi oscilla in un range di valori più contenuto rispetto alle stagioni successive. In queste prime fasi, le fluttuazioni dei valori di sentimento sembrano essere meno estreme.

Nelle stagioni successive, soprattutto nelle ultime, si osservano picchi sia positivi che negativi nei valori di sentimento. Questo suggerisce una maggiore complessità nelle emozioni dei personaggi e negli eventi che li coinvolgono.

Un dettaglio interessante è riscontrabile nella settima stagione, in cui la linea di Elena coincide in gran parte con il valore neutro. Questo accade perché Elena non è presente in tutte le puntate di questa stagione, essendo stata messa in uno stato di sonno magico attraverso un incantesimo. Elena esce da questo stato solo una volta che la maledizione viene spezzata.

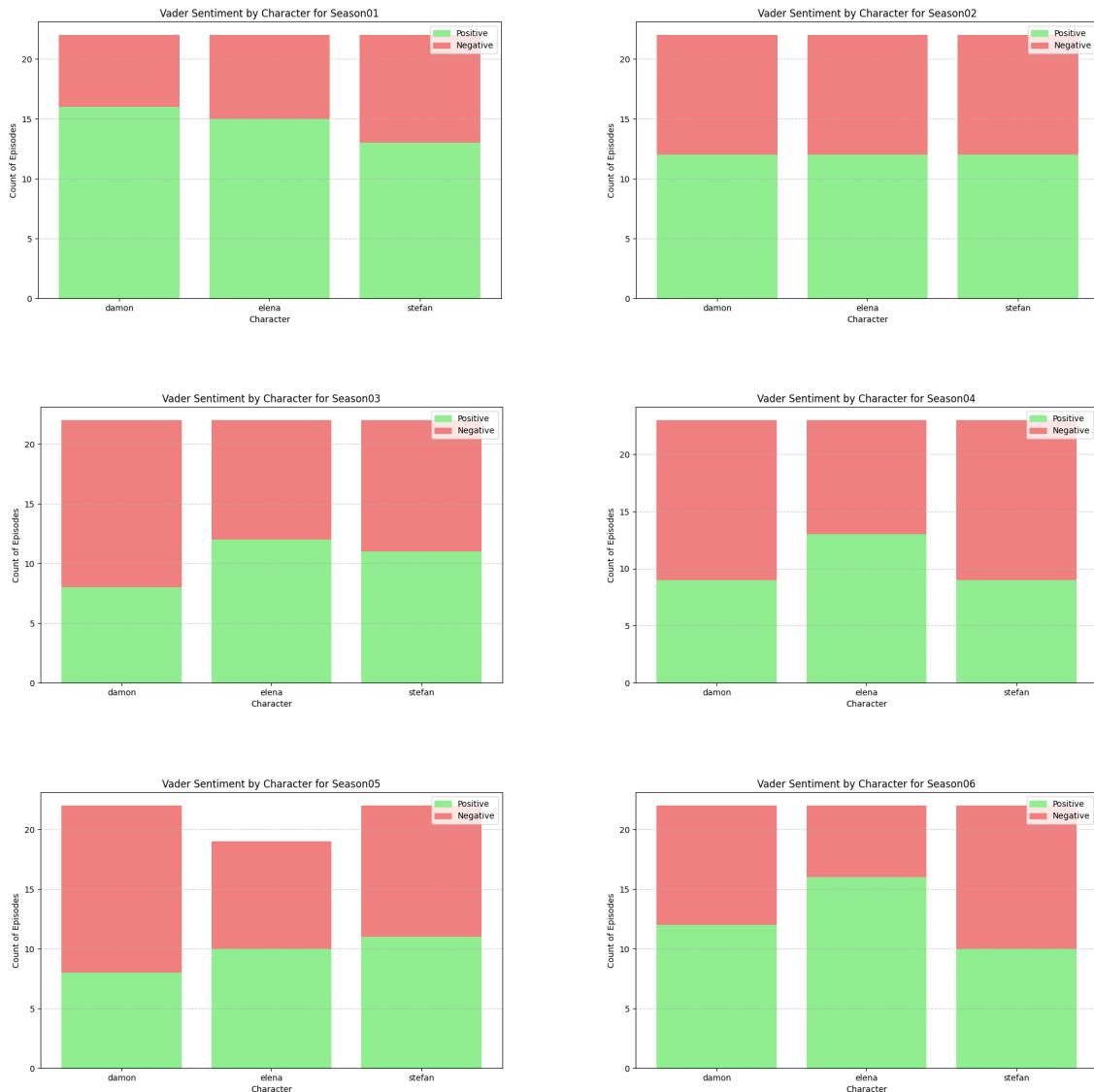
Inoltre, nella stagione finale, l'ottava, si nota un valore di sentimento particolarmente negativo per il personaggio di Elena, indicando la presenza eventi drammatici nel suo percorso narrativo.

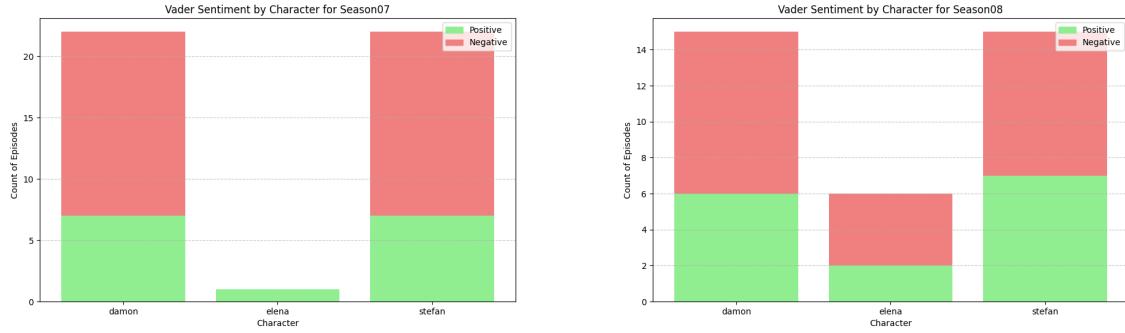
Tuttavia, questi grafici forniscono una panoramica generale e non consentono di determinare quanti episodi presentano un sentimento positivo o negativo in modo specifico.

Per questo motivo, per ottenere una visione più dettagliata del sentimento dei personaggi principali nelle diverse stagioni di "The Vampire Diaries", viene utilizzata la funzione "vader_barplot_by_seasons(df)", la quale genera un grafico a barre per ciascuna stagione della serie.

Nel grafico a barre, sull'asse delle ascisse sono elencati i nomi dei tre personaggi analizzati: Damon, Elena e Stefan. Sull'asse delle ordinate, invece, vengono riportati i conteggi degli episodi. Questo tipo di visualizzazione consente di vedere chiaramente quante puntate presentano un sentimento positivo o negativo per ciascun personaggio in ogni stagione.

In questo modo è possibile ottenere un'analisi dettagliata della distribuzione dei sentimenti nei vari episodi e comprendere come essi si evolvano nel corso delle diverse stagioni della serie.





L'analisi dei grafici a barre rivela interessanti tendenze. Innanzitutto, è importante notare che in tutte le stagioni della serie, troviamo episodi con sentiment sia positivo che negativo per i personaggi in analisi. Questo sottolinea la complessità delle loro esperienze emotive nel corso della storia.

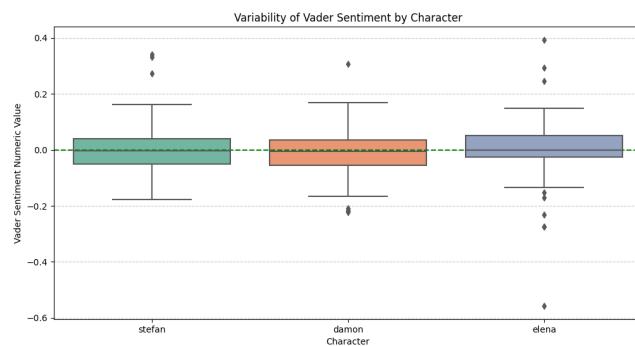
Tuttavia, nella settima stagione, il personaggio di Elena è notevolmente limitato, in quanto appare solo in un singolo episodio, come è stato rilevato anche precedentemente.

Un punto di interesse è la sesta stagione, che si distingue per il fatto che per tutti e tre i personaggi la maggioranza degli episodi è caratterizzata da un valore di sentiment positivo; questo riflette la predominanza di eventi gioiosi in questa stagione, come lo sviluppo della relazione che coinvolge Stefan e la vampira Caroline.

In generale, nelle altre stagioni, il numero di episodi con sentiment positivo e negativo sono equilibrati, con le barre divise quasi a metà. Ciò riflette il fatto che la serie intreccia costantemente momenti di felicità e momenti di tensione che coinvolgono i personaggi principali.

Infine, si può notare che la settima stagione sembra essere particolarmente positiva per i personaggi di Stefan e Damon, con una prevalenza di episodi caratterizzati da sentiment positivo. In conclusione, l'analisi dei grafici a barre offre una panoramica dettagliata delle dinamiche emotive dei personaggi principali in diverse stagioni di "The Vampire Diaries", evidenziando le variazioni nelle esperienze emotive e consentendo di apprezzare meglio l'evoluzione dei personaggi nel corso della serie.

Un ulteriore modo per esplorare il sentiment dei personaggi è capire quanto il loro valore del sentiment sia variabile nel corso delle stagioni, anche per determinare se vi sono differenze significative tra i personaggi oggetto di analisi. Per far ciò si utilizza la funzione "box_plot_episodes(df)" che produce un boxplot per ogni personaggio permettendo di capire aspetti interessanti. Questo grafico, riporta sulle ascisse i nomi dei vari personaggi, mentre sulle ordinate il valore del sentiment calcolato con Vader.

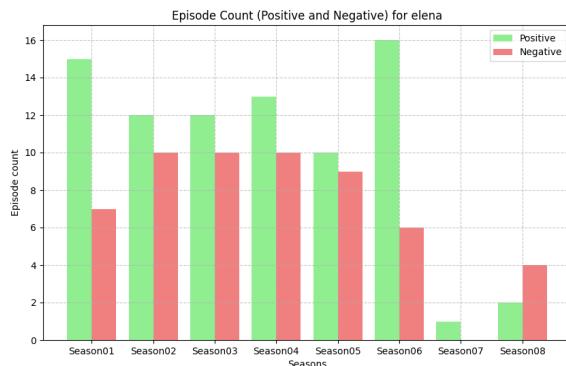


Ciò che subito è possibile notare, è che in termini di variabilità del sentiment non ci sono grandi differenze tra i tre personaggi. Stefan e Damon presentano le osservazioni centrali all'incirca egualmente divise tra osservazioni negative e osservazioni positive, anche se la scatola del boxplot di Damon è leggermente traslata verso il basso, rispetto a quella di Stefan. Elena presenta invece la maggior parte delle osservazioni centrali al

di sopra della linea della neutralità, anche se è possibile notare, che è il personaggio che riporta più valori anomali, che è stato notato anche precedentemente osservando i grafici a linee.

Dopo aver utilizzato vari grafici che permettessero di visualizzare i risultati inerenti ai diversi personaggi all'interno di uno stesso grafico, ora si utilizza una funzione che permette di costruire un grafico a barre separatamente per ogni personaggio, in modo tale da concentrare l'attenzione e comprendere meglio gli sviluppi e l'andamento del sentimento specificamente per ogni personaggio.

Si utilizza la funzione “barplot_for_character(df)”, la quale riportando sulle ascisse le varie stagioni e sulle ordinate il conteggio degli episodi permette di rendere visibile l'andamento del numero di episodi positivi e negativi nel corso delle stagioni, che non era ancora chiaro e visibile dai grafici precedentemente illustrati.

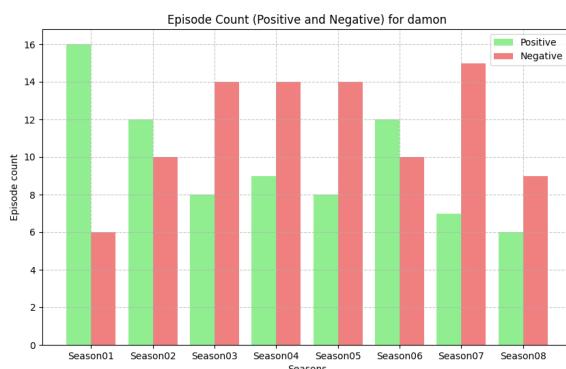


Concentrando l'attenzione sul personaggio di Elena Gilbert, si evidenzia l'evoluzione nel suo sentimento nel corso delle stagioni della serie: dopo una prima stagione in cui il suo sentimento è positivo, si osserva un calo significativo nelle stagioni successive; tuttavia, nella quarta stagione, si verifica una leggera ripresa, seguita da un aumento sostanziale nella sesta stagione.

Questo andamento può essere giustificato dalla trama della serie. Nella sesta stagione, Elena riscopre completamente la sua relazione con Damon, dopo aver avuto i ricordi di lui cancellati a causa del dolore insopportabile che provava per la sua presunta perdita. Inoltre, la quinta stagione presenta una mescolanza di episodi con sentimento positivo e negativo, indicando una complessa evoluzione emotiva del personaggio.

Infine, nella ottava stagione, sebbene Elena appaia solo in pochi episodi, si osserva una prevalenza di sentimento negativo. Questo può essere spiegato dal gran numero di eventi drammatici e tragici che caratterizzano questa stagione, come la morte e la perdita di Stefan, che ha un impatto emotivo devastante su Elena.

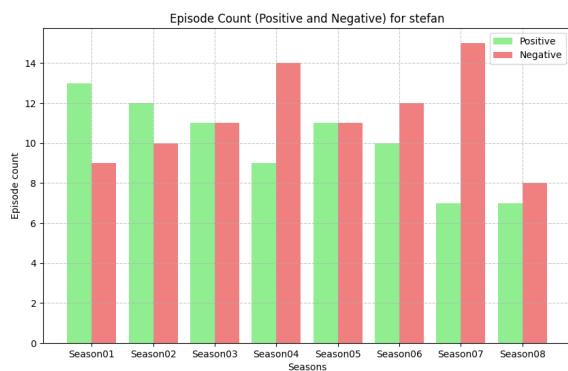
In sintesi, l'andamento del sentimento di Elena Gilbert riflette le sfumature e le evoluzioni della sua storia e delle sue relazioni nella serie, culminando in una conclusione emotivamente intensa nella stagione finale.



Possiamo osservare che il sentimento di Damon presenta un andamento notevolmente diverso rispetto a quello di Elena. Mentre Elena è una ragazza dal cuore gentile e tende a mostrare prevalentemente sentimenti positivi, la situazione è diversa per Damon, un vampiro noto per il suo animo vendicativo, spesso coinvolto in situazioni spiacevoli e osteggiato da vari nemici nel corso della serie.

Queste differenze si riflettono nei risultati dei sentimenti. Nelle stagioni tre, quattro e cinque, notiamo un numero simile di episodi con sentimento positivo e negativo per Damon, il che può essere attribuito all'intrecciarsi delle trame e delle vicende magiche presenti nella serie. Tuttavia, nella settima stagione, vediamo un notevole aumento degli episodi con sentimento negativo per Damon. Questo aumento può essere spiegato dalla situazione emotivamente intensa del personaggio durante questa stagione. Infatti, a causa di un incantesimo, Elena era sprofondata in uno stato di sonno eterno, e Damon aveva temuto di perderla per sempre. Questo lo aveva portato a prendere la drastica decisione di non voler più vivere, chiudendosi in una bara e addormentandosi profondamente. Tuttavia, in seguito verrà risvegliato dal fratello Stefan ma alla fine della stagione scomparirà misteriosamente all'interno di una cripta.

Anche nell'ottava stagione, osserviamo un'alta prevalenza di sentimento negativo per Damon, poiché la stagione è caratterizzata da decisioni di vita difficili, dure scelte, scontri e dal sacrificio del fratello Stefan. Questi eventi hanno contribuito a creare un'atmosfera di negatività intorno al personaggio di Damon durante l'ultima stagione della serie.

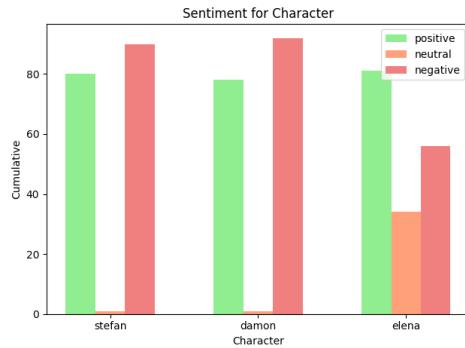


Anche Stefan presenta un notevole numero di episodi caratterizzati da sentimento negativo in ogni stagione. Tuttavia, a differenza di Elena e anche di Damon, che hanno stagioni in cui prevalgono i sentimenti positivi, come ad esempio la prima stagione, per Stefan il numero di episodi con sentimenti positivi e negativi è spesso molto simile. Ci sono alcune eccezioni, come nella quarta e nella settima stagione, in cui si registrano più episodi con sentimenti negativi.

Un evento drammatico nella quarta stagione è stato il momento in cui Elena ha dovuto fare una scelta tra i due fratelli, rifiutando Stefan e portandolo alla decisione di abbandonare la città. Successivamente, Stefan è stato imprigionato da una creatura magica, contribuendo così a rendere la quarta stagione una delle più ricche di episodi con sentimenti negativi sia per lui che per Damon. Come già menzionato, la settima stagione è stata caratterizzata da una serie continua di eventi drammatici e situazioni cariche di tensione, che hanno influenzato profondamente il sentimento dei personaggi.

Infine, per concludere l'analisi stagione per stagione, e per avere una panoramica un po' più globale per ogni personaggio, si utilizza la funzione "plot_combined_sentiment(df)" che costruisce un grafico a barre, in cui nelle ascisse si riportano i nomi dei personaggi e sulle ordinate il conteggio degli episodi. Questo grafico a barre, per ogni personaggio riporta tre barre: una che conteggia gli episodi con sentimento positivo, una che conteggia gli episodi con sentimento negativo e una che conteggia gli episodi con sentimento neutro, ossia il numero di episodi in cui il personaggio non compare.

In questo modo possiamo vedere la situazione dei personaggi, nella totalità della serie senza fare distinzioni tra stagioni.



Possiamo dunque notare che i personaggi Stefan e Damon, seppur l'analisi grafica precedentemente svolta ci mostri le differenze che emergono tra i due personaggi all'interno delle stagioni, a livello cumulato sono molto simili. Entrambi i personaggi sono presenti in quasi tutti gli episodi della serie televisiva e condividono una serie di esperienze ed eventi che influenzano profondamente i loro sentimenti nel corso delle stagioni.

D'altra parte, il personaggio di Elena Gilbert si distingue nettamente dagli altri due. Elena è un personaggio intrinsecamente più positivo, con valori di sentimento negativo sensibilmente inferiori rispetto a Stefan e Damon. Questa differenza può essere attribuita alla sua natura compassionevole e alla sua propensione per il bene. Inoltre, l'evoluzione del suo sentimento nel corso della serie è influenzata in modo significativo dalla sua relazione con entrambi i fratelli Salvatore, il che aggiunge un elemento di complessità alla sua storia.

In definitiva, le analisi dei sentimenti dei personaggi principali di "The Vampire Diaries" ci forniscono una prospettiva interessante sulle loro evoluzioni emotive nel corso delle otto stagioni della serie. Stefan e Damon mostrano similitudini sorprendenti, mentre Elena si distingue per la sua inclinazione verso un sentimento più positivo. Questi risultati ci aiutano a comprendere meglio i personaggi e le dinamiche della storia narrata nella serie televisiva.

Dopo aver analizzato i sentimenti dei personaggi stagione per stagione, concentriamoci ora su un livello più globale per ottenere una panoramica generale delle loro evoluzioni emotive. Per fare ciò, utilizzeremo il file "3_season_results_tvd.py" contenuto all'interno della directory "df". Questo file riporta per ciascun personaggio il valore del sentimento aggregato per ogni stagione, sia in forma numerica che con etichette, utilizzando sia Vader che TextBlob.

Questo approccio ci consentirà di comprendere ancora meglio, e da un punto di vista diverso, come i sentimenti dei personaggi principali si siano evoluti nell'arco delle otto stagioni della serie televisiva "The Vampire Diaries".

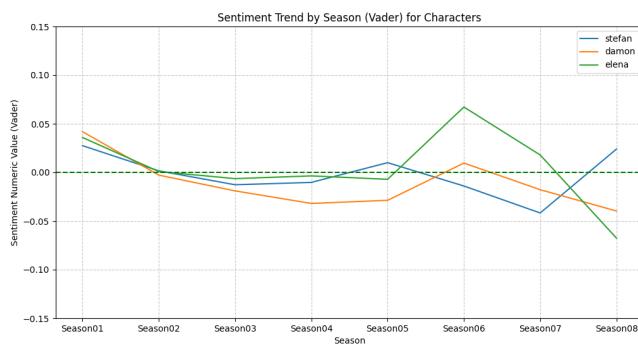
Questo file ha 24 righe e 6 colonne e ne viene di sotto mostrata un'anteprima:

Character	Season	Vader_sent_label	Textblob_sent_label	Vader_sent_numeric	Textblob_sent_numeric
stefan	Season01	positive	positive	0.0275	0.0122
stefan	Season02	positive	positive	0.0016	0.026
stefan	Season03	negative	positive	-0.0127	0.0297
stefan	Season04	negative	positive	-0.0103	0.0218
stefan	Season05	positive	positive	0.01	0.0512
stefan	Season06	negative	negative	-0.014	-0.0121
stefan	Season07	negative	negative	-0.0417	-0.0074
stefan	Season08	positive	positive	0.024	0.0235
damon	Season01	positive	positive	0.0419	0.038
damon	Season02	negative	positive	-0.0028	0.0472
damon	Season03	negative	positive	-0.019	0.0359
damon	Season04	negative	positive	-0.0319	0.0233
damon	Season05	negative	positive	-0.0287	0.0012
damon	Season06	positive	positive	0.0096	0.0263
damon	Season07	negative	positive	-0.0178	0.0153

Per condurre questa analisi ci si serve delle funzioni contenute all'interno del file “3_Sentiment_characters_graphs.py”.

Per prima cosa, per completezza, anche in questo caso si calcola la correlazione dei valori restituiti dai due metodi per verificare se a livello globale vi è una variazione del risultato rispetto al valore ottenuto precedentemente nell'analisi stagionale. Il valore della correlazione restituito dalla funzione “calculate_correlation(df)” è 0.32558, perciò non si verificano differenze con il valore ottenuto per il file con i valori episodio per episodio. Questo valore riflette una correlazione leggermente positiva ma non dal valore altamente predittivo. Tuttavia, si sceglie, come di utilizzare i valori restituiti da Vader, in quanto si riscontra che per la stagione “The Vampire Diaries” riesca a captare meglio delle sfumature di sentiment.

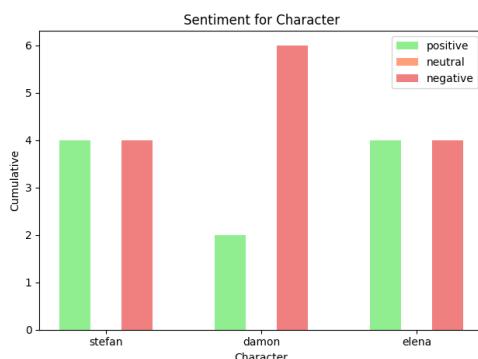
Per visualizzare il trend del sentiment aggregato per stagione di ogni personaggio, si utilizza la funzione “plot_line_for_seasons(df)” che costruisce un singolo grafico a linee che permette di mostrare simultaneamente l'andamento in analisi per tutti e tre i personaggi. Sulle ascisse si riportano le stagioni, mentre sulle ordinate i valori del sentiment.



Questo grafico permette di notare che i tre personaggi presentano un'evoluzione del sentiment pressoché simile, che tende inizialmente a diminuire per poi differenziarsi maggiormente a partire dalla quinta stagione.

La linea di Elena, confermando ciò che abbiamo detto precedentemente, è quella che mantiene valori maggiormente positivi, per subire una brusca discesa nell'ottava stagione. Le linee di Stefan e Damon, seppur con delle evidenti differenze, si mantengono molto simili, anche se la linea di Damon tende a mantenere un andamento leggermente meno positivo rispetto a quella di Stefan.

Infine, per visualizzare lo stesso risultato ma utilizzando i valori espressi in etichetta, calcolati sempre con Vader, si utilizza la funzione “plot_combined_sentiment_seasons(df)” che crea un grafico a barre, riportando per ogni personaggio il conteggio delle stagioni con sentiment positivo, neutro e negativo. Sulle ascisse vengono riportati i nomi dei personaggi mentre sulle ordinate il conteggio delle stagioni.



Possiamo notare che nessun personaggio ha la barra delle stagioni neutre, perché ogni personaggio, anche se per pochi episodi, come Elena, compare sempre in tutte le stagioni. Un aspetto degno di nota è la distribuzione equilibrata di sentiment positivo e negativo nelle stagioni di Elena e Stefan. Questo suggerisce una sorta di equilibrio emotivo nei loro percorsi, nonostante le sfide e gli eventi drammatici che affrontano. D'altra parte, Damon emerge come il personaggio con più stagioni caratterizzate da un sentiment negativo. Questa differenza rispecchia le loro personalità e le loro scelte di vita. Mentre Stefan cerca di vivere in armonia con la sua natura vampira e di mantenere un certo equilibrio, Damon, con il suo animo vendicativo e la tendenza a coinvolgersi in situazioni oscure, accumula più momenti di negatività. Pertanto, a livello aggregato, i personaggi Stefan ed Elena appaiono molto più simili di quanto sembrino effettuando l'analisi episodio per episodio. Sicuramente, utilizzando i dati a livello aggregato si perdono diverse informazioni, ma è indubbio che il personaggio maggiormente positivo tra i tre sia Elena, e che quello più negativo, al contrario sia Damon. Infatti, nonostante i fratelli Salvatore siano stati descritti come personaggi dalle molte similitudini, Stefan tende ad essere meno vendicativo e più equilibrato rispetto a Damon, per quanto si stia parlando sempre di creature sovrannaturali.

In conclusione, l'analisi del sentiment dei personaggi in "The Vampire Diaries" ha fornito informazioni intriganti e ci ha aiutato a comprendere meglio come evolvano le emozioni dei protagonisti nel corso delle otto stagioni della serie. Ciò dimostra quanto sia interessante esplorare le sfumature emotive dei personaggi nella serie e come ciò possa arricchire la nostra comprensione delle loro storie e personalità.

3.2.2 Analisi dei risultati: "The Big Bang Theory"

Anche per questo punto l'analisi sarà svolta su più livelli prima in modo più specifico, esaminando il sentiment di ogni singolo episodio e successivamente, a livello più aggregato calcolando il sentiment di ogni stagione.

Conduciamo ora l'analisi del sentiment dei tre personaggi Sheldon, Leonard e Penny considerando i risultati sul sentiment che sono stati calcolati episodio per episodio.

Per far ciò viene utilizzato il file "3_episode_results_tbtt.csv" che riporta il valore del sentiment sia espresso in valore numerico sia come etichetta per ogni episodio di ogni stagione e per ogni personaggio, sia calcolato con Vader che con TextBlob.

Questo file ha 694 righe e 7 colonne, e ne viene riportata sotto un'anteprima:

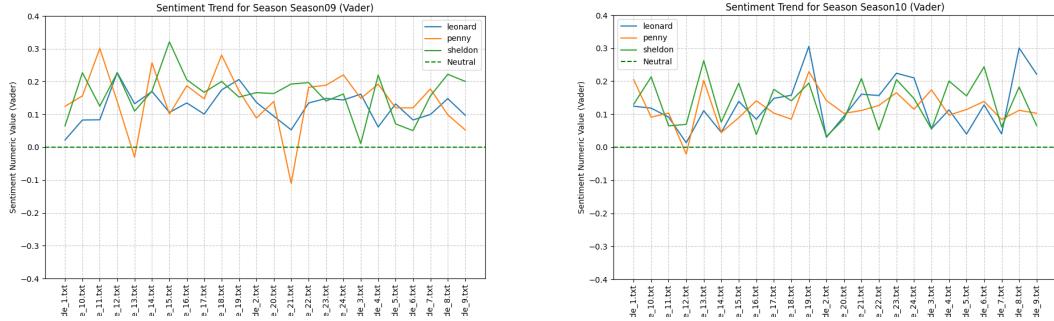
Character	Season	Episode	Vader_sent_label	Textblob_sent_label	Vader_sent_numeric	Textblob_sent_numeric
sheldon	Season01	episode_1.txt	positive	positive	0.0618	0.0326
sheldon	Season01	episode_2.txt	positive	positive	0.0224	0.0245
sheldon	Season01	episode_3.txt	negative	negative	-0.0095	-0.0041
sheldon	Season01	episode_4.txt	positive	positive	0.103	0.0601
sheldon	Season01	episode_5.txt	positive	positive	0.0862	0.0317
sheldon	Season01	episode_6.txt	positive	positive	0.0695	0.024
sheldon	Season01	episode_7.txt	negative	positive	-0.0515	0.0092
sheldon	Season01	episode_8.txt	positive	positive	0.0545	0.044
sheldon	Season01	episode_9.txt	positive	positive	0.0592	0.0417
sheldon	Season01	episode_10.txt	negative	negative	-0.06	-0.0056
sheldon	Season01	episode_11.txt	positive	negative	0.0124	-0.0437
sheldon	Season01	episode_12.txt	positive	positive	0.0944	0.0611
sheldon	Season01	episode_13.txt	positive	positive	0.0197	0.0522

In questo studio, è stato ritenuto importante valutare la correlazione tra i valori restituiti da Vader e quelli da TextBlob per determinare se entrambi i metodi mostrano un andamento simile nell'analisi del sentiment. Per effettuare questa analisi, è stata utilizzata la funzione "calculate_correlation(df)" contenuta nel file "3_Sentiment_characters_graphs.py", che ha restituito un valore di correlazione pari a 0.52539. Questo valore indica una correlazione positiva tra le due variabili prese in considerazione. In altre parole, c'è una tendenza positiva tra i risultati ottenuti dai due metodi: quando uno dei metodi rileva un aumento nel sentiment, anche l'altro tende ad aumentare, e viceversa.

L'analisi verrà quindi condotta attraverso il supporto di grafici, le cui funzioni per realizzarli sono contenute all'interno del file "3_Sentiment_characters_graphs.py".

Per ottenere una visione dettagliata dell'andamento del sentiment dei personaggi episodio per episodio, abbiamo utilizzato la funzione "sentiment_line_for_episodes(df)". Questa funzione ha permesso di creare grafici a linee per ciascuna stagione, utilizzando i valori numerici del sentiment calcolati con Vader. Sugli assi dei grafici sono stati riportati gli episodi di ogni stagione (ascisse) e i valori del sentiment (ordinate). Ogni linea nel grafico rappresenta l'evoluzione del sentiment nel corso della stagione per un personaggio specifico: la linea blu è associata a Leonard, la linea arancione a Penny e la linea verde a Sheldon. Inoltre, è presente anche una linea tratteggiata orizzontale verde che coincide con la neutralità. Questa rappresentazione visiva ci consente di analizzare in modo approfondito come il sentiment dei personaggi si sviluppa nel corso della serie, offrendo un'ulteriore comprensione delle dinamiche emotive dei protagonisti.





Dai grafici si nota che il sentimento è caratterizzato da oscillazioni, generalmente sempre positive, ma che non identificano un trend specifico nel comportamento emotivo dei personaggi nel corso della serie, ma indicano una variazione del sentimento sempre più o meno positiva fino all'ultima stagione.

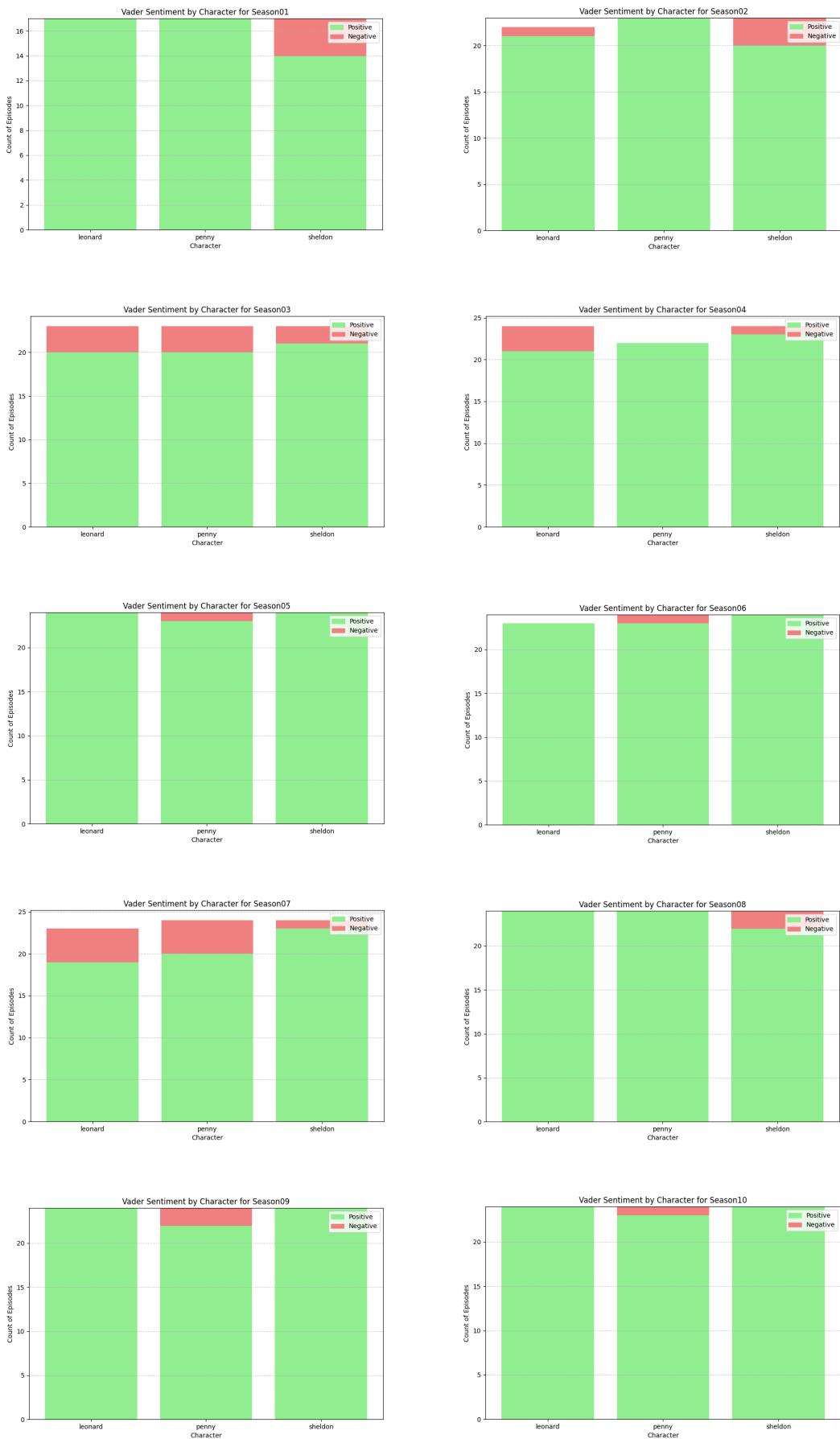
Nella prima stagione della serie, il sentimento dei personaggi oscilla negli episodi in modo più contenuto rispetto alle stagioni successive. Dopo la prima stagione, infatti, si osservano oscillazioni più frequenti con picchi positivi e relativamente pochi picchi negativi nei valori del sentimento. Questo suggerisce una leggera variabilità nelle emozioni dei personaggi.

La linea arancione di Penny nelle prime stagioni è positiva, trovandosi sopra la linea di neutralità, mentre dalla terza stagione comincia ad avvicinarsi alla linea di neutralità abbassando quindi i valori positivi e arrivando a toccare picchi di valori negativi come succede nelle stagioni 3, 5, 7 e 9. Questo può essere giustificato dalle difficoltà finanziarie e lavorative di Penny, difficoltà nella relazione con Leonard, e incertezze sul suo futuro da attrice. Anche la linea di Leonard generalmente tende ad essere molto vicina alla linea di neutralità, con valori di positività più bassi, e specie nella settima stagione ci sono alcuni episodi in cui il suo valore è negativo. Ciò è giustificato dalle difficoltà nella relazione con Penny e nella propria carriera professionale, che spesso contribuiscono a sensazioni di insoddisfazione. Sheldon invece è il personaggio la cui linea tocca la neutralità più spesso, nelle prime quattro stagioni sono presenti picchi di valori negativi, giustificati dal suo carattere saccante, e dalle sue difficoltà nei rapporti interpersonali; i suoi valori di positività aumentano nelle successive stagioni, pur stando sempre vicino alla linea di neutralità, con maggiori valori positivi nella decima stagione, stagione in cui sposa Amy: qui infatti la sua linea è maggiormente distante dalla linea di neutralità rispetto alle altre stagioni.

Tuttavia, va notato che questi grafici forniscono una panoramica generale e non consentono di determinare quanti episodi presentano un sentimento positivo o negativo in modo specifico.

Per questo motivo, per ottenere una visione più dettagliata del sentimento dei personaggi principali nelle diverse stagioni di "The Big Bang Theory", viene utilizzata la funzione "vader_barplot_by_seasons(df)", che genera un grafico a barre per ciascuna stagione della serie.

Nel grafico a barre, sull'asse delle ascisse sono elencati i nomi dei tre personaggi analizzati: Leonard, Penny e Sheldon. Sull'asse delle ordinate, invece, vengono riportati i conteggi degli episodi. Questo tipo di visualizzazione consente di vedere chiaramente quante puntate presentano un sentimento positivo o negativo per ciascun personaggio in ogni stagione, così è possibile ottenere un'analisi dettagliata della distribuzione dei sentimenti nei vari episodi e comprendere come essi si evolvano nel corso delle diverse stagioni della serie.

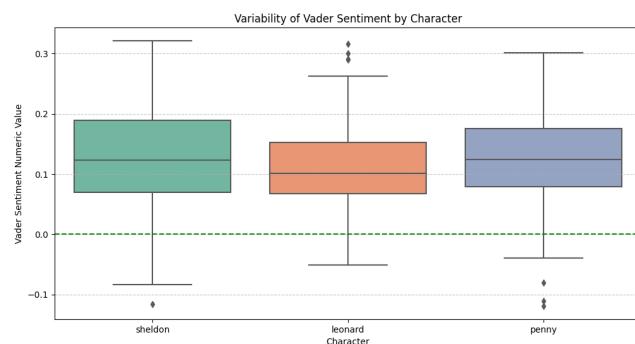


L'analisi dei grafici a barre rivela che in tutte le stagioni della serie si trovano maggiormente episodi con sentimento positivo piuttosto che negativo per i tre personaggi in analisi. Questo può essere giustificato dalla normalità delle scene di vita quotidiana presenti nella serie.

Un punto di interesse sono la settima e la terza stagione, che si distinguono per il fatto che per tutti e tre i personaggi sono presenti episodi caratterizzati da un valore di sentimento negativo; questo riflette la predominanza di eventi difficili in queste stagioni. Nella terza stagione, infatti, Leonard e Penny cercano di riaccendere la loro relazione in crisi per problemi causati dalle loro differenze. Sempre in questa stagione Sheldon sperimenta un momento di crisi identitaria quando si rende conto che potrebbe non raggiungere mai un Premio Nobel. Questo lo porta a una fase di introspezione e ad affrontare le incertezze sulla sua ricerca scientifica. Penny inoltre scopre che i genitori stanno per divorziare, mettendo in discussione la sua visione del matrimonio.

Come visto nei grafici a linee Sheldon sembra essere più negativo nelle prime tre stagioni, aumentando la sua positività nelle restanti stagioni, nelle stagioni centrali inizia una relazione con Amy che culminerà nella proposta di matrimonio nella decima stagione.

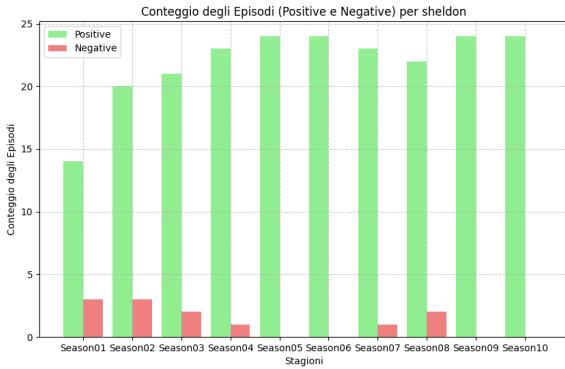
Un ulteriore modo per esplorare il sentimento dei personaggi è capire quanto il loro valore del sentimento sia variabile nel corso delle stagioni, anche per determinare se vi sono differenze significative tra i personaggi oggetto di analisi. Per far ciò si utilizza la funzione “box_plot_episodes(df)” che produce un boxplot per ogni personaggio permettendo di capire aspetti interessanti. Questo grafico, riporta sulle ascisse i nomi dei vari personaggi, mentre sulle ordinate il valore del sentimento calcolato con Vader.



Ciò che subito è possibile notare, è che in termini di variabilità del sentimento non ci sono grandi differenze tra i tre personaggi. Tutti e tre i box dei personaggi sono positivi, Sheldon sembra essere il personaggio che presenta più variabilità nei suoi valori, seguito da Penny e da Leonard, il cui box è quello più traslato verso il basso dei tre personaggi. Si possono osservare maggiori outlier negativi per Penny, e alcuni positivi per Leonard. Dal boxplot sembra che il personaggio meno positivo in generale sia Leonard.

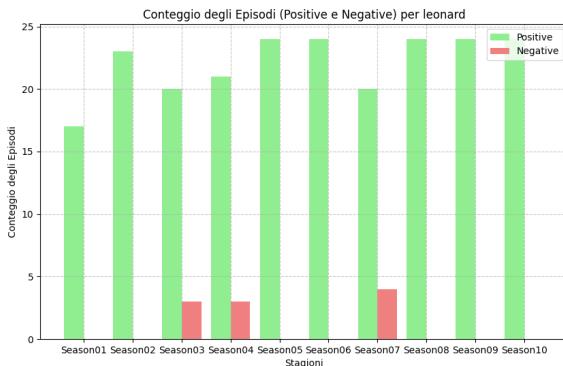
Dopo aver utilizzato vari grafici che permettessero di visualizzare i risultati inerenti ai diversi personaggi all'interno di uno stesso grafico, ora si utilizza una funzione che permette di costruire un grafico a barre separatamente per ogni personaggio, in modo tale da concentrare l'attenzione e comprendere meglio gli sviluppi e l'andamento del sentimento specificamente per ogni personaggio.

Si utilizza la funzione “barplot_for_character(df)”, la quale riportando sulle ascisse le varie stagioni e sulle ordinate il conteggio degli episodi permette di rendere visibile l'andamento del numero di episodi positivi e negativi nel corso delle stagioni, che non era ancora chiaro e visibile dai grafici precedentemente illustrati.



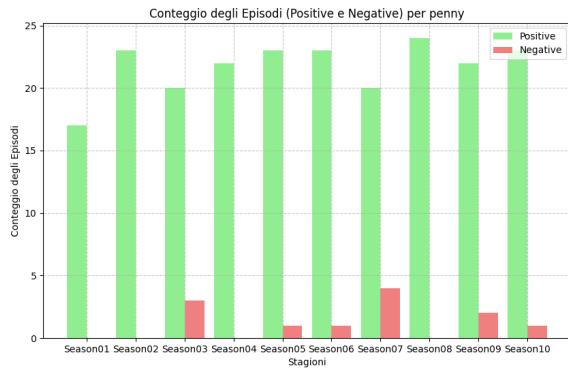
Concentrando l'attenzione sul personaggio di Sheldon Cooper, si evidenzia l'evoluzione nel suo sentimento nel corso delle stagioni della serie: nelle prime due stagioni sono presenti alcune puntate con sentimento negativo, con un calo di queste fino alla quarta stagione; la quinta e la sesta risultano essere totalmente positive, riprendendo poi una negatività che aumenta fino all'ottava stagione, terminando con le restanti due stagioni in positivo.

Sheldon è un personaggio estremamente razionale e logico, spesso basa le sue decisioni su fatti scientifici e logici piuttosto che sull'emozione, con una mancanza di comprensione delle interazioni e delle norme sociali, spesso comportandosi in modo schivo e insensibile nei confronti degli altri, per questo è comprensibile che il sentimento dei suoi dialoghi venga catalogato come "negativo" proprio a causa del suo atteggiamento arrogante e sprezzante verso coloro che ritiene inferiori dal punto di vista intellettuale o che non condividono i suoi interessi. Tuttavia, nel corso della serie, Sheldon mostra una crescita significativa nel comprendere le emozioni e migliorare le sue abilità sociali, anche se in modo goffo e poco ortodosso sviluppa un affetto profondo per i suoi amici, e mostra la sua vulnerabilità emotiva in modo più evidente. Pur rimanendo un genio della scienza, inizia a concedersi qualche flessibilità rispetto alle sue rigide convinzioni.



Possiamo osservare che il sentimento di Leonard presenta un andamento più positivo rispetto a quello di Sheldon nel corso delle stagioni. Leonard, pur essendo uno scienziato, è meno estremo di Sheldon in termini di razionalità, restando comunque molto incline all'approccio scientifico e logico alla vita, ha maggiori abilità sociali rispetto a Sheldon, ma mostra insicurezze e ansie, soprattutto in relazione alle dinamiche interpersonali. Tuttavia, nel corso della serie Leonard cresce emotivamente e guadagna fiducia in sé stesso, specialmente attraverso la sua relazione con Penny.

Le stagioni che presentano più episodi negativi sono la stagione 3, 4 e 7, giustificato nella terza dalla rottura nella relazione con Penny, che lo lascia emotivamente turbato e gli provoca un senso di perdita; nella quarta invece inizia una relazione con Priya, ma anche questa termina senza successo. Nella settima invece convive con Penny, dimostrando la sua voglia di costruire una vita insieme; tuttavia, occasionali incertezze e differenze lo portano a momenti di tensione.



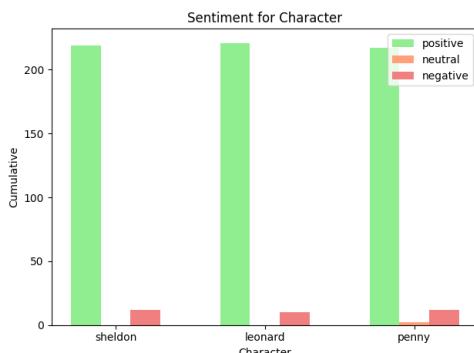
Anche Penny presenta episodi caratterizzati da sentiment negativi in alcune stagioni. Anche lei come Sheldon e Leonard presenta una netta prevalenza di episodi con sentimento positivo.

Penny inizialmente viene presentata come la "ragazza della porta accanto", una cameriera che aspira a diventare un'attrice di successo, con interessi tradizionali e meno orientati alla scienza rispetto a Sheldon e Leonard, tuttavia rappresenta il "collegamento con il mondo esterno" per il gruppo di amici integrandosi nella loro cerchia, anche se inizialmente tende a sentirsi un po' fuori posto. Nelle prime stagioni affronta difficoltà finanziarie e lavorative, spesso lottando per pagare le spese ricercando lavori più stabili e gratificanti.

Nel corso delle stagioni, Penny sviluppa una forte amicizia e successivamente una relazione con Leonard, portandola ad un maggiore coinvolgimento nel mondo scientifico, si evolve emotivamente e prende decisioni più ponderate riguardo alla sua carriera e relazioni. Nelle stagioni centrali, tuttavia, affronta una serie di sfide e delusioni sentimentali, inclusa la fine della sua relazione con Leonard e il continuo insuccesso nella sua carriera di attrice; i continui ostacoli la portano a provare ansie, momenti di incertezza e insoddisfazione.

Infine, per concludere l'analisi stagione per stagione, e per avere una panoramica un po' più globale per ogni personaggio, si utilizza la funzione “plot_combined_sentiment(df)” che costruisce un grafico a barre, in cui nelle ascisse si riportano i nomi dei personaggi e sulle ordinate il conteggio degli episodi. Questo grafico a barre, per ogni personaggio riporta tre barre: una che conteggia gli episodi con sentimento positivo, una che conteggia gli episodi con sentimento negativo e una che conteggia gli episodi con sentimento neutro, ossia il numero di episodi in cui il personaggio non compare.

In questo modo possiamo vedere la situazione dei personaggi, nella totalità della serie senza fare distinzioni tra stagioni.



Possiamo dunque notare che i risultati dei tre personaggi a livello cumulato sono abbastanza identici. Come rilevato dal boxplot Leonard sembra essere leggermente più positivo rispetto a Sheldon e Penny, e quest'ultima presenta qualche episodio catalogato come neutro.

Dopo aver analizzato i sentiment dei personaggi stagione per stagione, concentriamoci ora su un livello più globale per ottenere una panoramica generale delle loro evoluzioni emotive. Per fare ciò, utilizzeremo il file "3_season_results_tbbt.py" contenuto all'interno della directory "df". Questo file riporta per ciascun personaggio il valore del sentiment aggregato per ogni stagione, sia in forma numerica che con etichette, utilizzando sia Vader che TextBlob.

Questo approccio ci consentirà di comprendere ancora meglio, e da un punto di vista diverso, come i sentiment dei personaggi principali si siano evoluti nell'arco delle otto stagioni della serie televisiva "The Big Bang Theory".

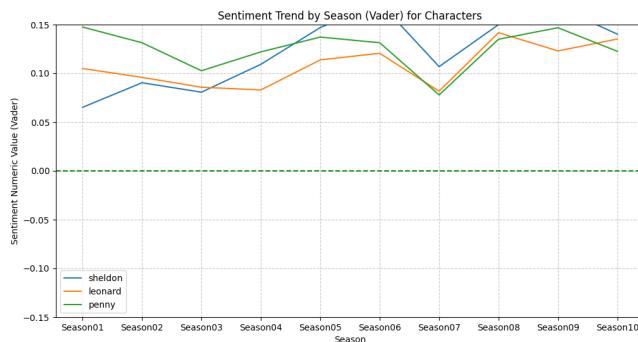
Questo file ha 31 righe e 6 colonne e ne viene di sotto mostrata un'anteprima:

```
Character,Season,Vader_sent_label,Textblob_sent_label,Vader_sent_numeric,Textblob_sent_numeric
sheldon,Season01,positive,positive,0.0381,0.0217
sheldon,Season02,positive,positive,0.0633,0.0512
sheldon,Season03,positive,positive,0.0569,0.0316
sheldon,Season04,positive,positive,0.0771,0.0388
sheldon,Season05,positive,positive,0.0997,0.041
sheldon,Season06,positive,positive,0.1193,0.0751
sheldon,Season07,positive,positive,0.0723,0.0559
sheldon,Season08,positive,positive,0.0988,0.0579
sheldon,Season09,positive,positive,0.1302,0.0843
sheldon,Season10,positive,positive,0.0857,0.0736
leonard,Season01,positive,positive,0.093,0.0688
leonard,Season02,positive,positive,0.0717,0.0323
leonard,Season03,positive,positive,0.0625,0.0361
leonard,Season04,positive,positive,0.0708,0.0606
```

Per condurre questa analisi ci si serve delle funzioni contenute all'interno del file "3_Sentiment_characters_graphs.py".

Per prima cosa, per completezza, anche in questo caso si calcola la correlazione dei valori restituiti dai due metodi per verificare se a livello globale vi è una variazione del risultato rispetto al valore ottenuto precedentemente nell'analisi stagionale. Il valore della correlazione restituito dalla funzione "calculate_correlation(df)" è 0.74223; questo valore riflette una correlazione abbastanza positiva dal valore altamente predittivo. Sceglieremo quindi di utilizzare i valori restituiti da Vader per semplicità.

Per visualizzare il trend del sentiment aggregato per stagione di ogni personaggio, si utilizza la funzione "plot_line_for_seasons(df)" che costruisce un singolo grafico a linee che permette di mostrare simultaneamente l'andamento in analisi per tutti e tre i personaggi. Sulle ascisse si riportano le stagioni, mentre sulle ordinate i valori del sentiment.

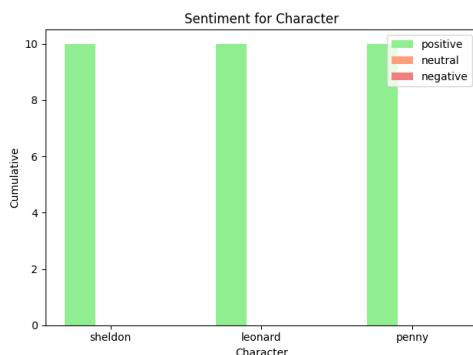


Questo grafico permette di notare che i tre personaggi presentano un'evoluzione del sentiment pressoché simile per Penny e Leonard, le cui linee tendono a presentare lo stesso andamento. Partono infatti con un punteggio più positivo rispetto a quello di Sheldon, e in seguito tendono a diminuire intorno alla terza stagione per poi aumentare di nuovo intorno alla quinta stagione, con una brusca discesa nella settima, che sembra essere la stagione più negativa della serie, e una ripresa dall'ottava. La linea di Penny tende a mantenere un

andamento nettamente più positivo rispetto a quella di Leonard che, come visto dai precedenti grafici, sembra essere il personaggio meno positivo, con i valori più bassi.

La linea di Sheldon invece è quella che presenta picchi di valori maggiormente positivi, parte nelle prime stagioni con una positività molto più bassa rispetto agli altri due amici, aumentando rapidamente fino alla sesta stagione, seguita dal comune decremento nella settima, per poi incrementare in un altro picco di positività nella nona stagione.

Infine, per visualizzare lo stesso risultato ma utilizzando i valori espressi in etichetta, calcolati sempre con Vader, si utilizza la funzione “plot_combined_sentiment_seasons(df)” che crea un grafico a barre, riportando per ogni personaggio il conteggio delle stagioni con sentiment positivo, negativo e neutro. Sulle ascisse vengono riportati i nomi dei personaggi mentre sulle ordinate il conteggio delle stagioni.



Come nei grafici creati nel punto 2 riguardanti le stagioni, anche per i personaggi si può notare che a livello aggregato nessun personaggio ha la barra delle stagioni neutre o negative. La positività del sentiment rilevata dai dialoghi dei tre personaggi in tutti gli episodi della serie è in linea con il genere leggero e comico della sitcom confermandone il trend generalmente positivo ancora una volta.

PUNTO 4: TOPIC MODELING

Dopo aver analizzato il sentimento delle due serie e dei loro personaggi principali, il progetto si sposterà verso l'identificazione dei temi chiave presenti in entrambe le opere attraverso l'uso della tecnica di Topic Modeling. Questo metodo consente di estrarre e visualizzare in modi diversi i principali argomenti trattati all'interno delle serie esaminate.

L'individuazione dei temi trattati permetterà una comprensione più approfondita delle singole serie, ma soprattutto aprirà la strada a ulteriori confronti tra le due. Per condurre questa analisi, saranno impiegate due metodologie: il modello Latent Dirichlet Allocation (LDA) e il modello Non-Negative Matrix Factorization (NMF).

Sono stati sviluppati due script dedicati a questo obiettivo:

- "4_topic_modeling.py": in cui verranno utilizzati i modelli LDA e NMF, sfruttando la libreria Sklearn, per estrarre i cinque principali topic stagionalmente e poi aggregati per l'intera serie;
- "4_html_visualization.py": in cui sarà impiegato il modello LDA, utilizzando la libreria "Gensim", per estrarre i cinque principali topic stagionalmente e poi aggregati per l'intera serie. Successivamente, la " sarà utilizzata libreria "pyLDAvis" per creare file HTML interattivi che consentiranno di visualizzare i risultati relativi ai topic estratti.

4.1: Modello LDA e NMF

4.1.1 Il codice

L'analisi del sentimento all'interno dello script "4_topic_modeling.py" è un processo strutturato in vari passaggi che sfrutta i metodi Latent Dirichlet Allocation (LDA) e Non-Negative Matrix Factorization (NMF) per identificare i temi principali che permeano la serie televisiva in esame. L'utilizzo di due diversi approcci è fondamentale per valutare la similitudine dei risultati ottenuti e determinare quale metodo si adatti meglio ai dati forniti, offrendo risultati più precisi.

L'analisi della Topic Modeling è condotta su due livelli distinti:

1. Identificazione dei principali temi per ogni singola stagione;
2. Identificazione dei principali temi per l'intera serie.

Per effettuare l'analisi verranno utilizzati come fonte di dati, i corpus delle due serie che sono stati implementati nel punto primo di questo progetto.

Tuttavia, prima di addentrarsi nell'analisi vera e propria, è necessario eseguire alcune operazioni preliminari per estrarre e preparare il testo da esaminare di ogni episodio per stagione. Queste operazioni sono gestite dalle seguenti funzioni:

- "clean_text(text)": Questa funzione accetta il testo di un episodio e esegue una serie di operazioni di pre-elaborazione, tra cui la divisione del testo per battute, l'eliminazione dei nomi dei personaggi all'inizio delle battute, la rimozione della punteggiatura, la trasformazione del testo in token, l'eliminazione delle stopwords in lingua inglese, la lemmatizzazione del testo, la rimozione dei token con meno di quattro caratteri e la rimozione dei token inclusi in elenchi di parole poco significative ("names_tvd", "names_tbtt", "verbs" e "other_words"). Infine, converte la lista dei token in una stringa, che rappresenta l'output della funzione;
- "get_episode_scripts(corpus_directory)": Questa funzione scorre il corpus fornito della serie, composto da stagioni ed episodi, ed esegue la funzione "clean_text(text)" per ogni episodio.

Successivamente, i risultati vengono salvati in un dizionario, dove le chiavi rappresentano le stagioni e i valori sono liste di stringhe contenenti gli episodi ripuliti dalla precedente funzione.

Una volta completata la preparazione del testo, e dunque dal corpus della serie aver ottenuto come struttura dati, un dizionario che riporta per ogni stagione della serie, il testo degli episodi corrispondenti alla stessa, è possibile eseguire la Topic Modeling. Sono state create due funzioni, la prima per calcolare i topic utilizzando il modello LDA e la seconda che utilizzi per lo stesso scopo il modello NMF, rispettivamente chiamate "topics_lda(seasons_scripts, top=5)" e "topics_nmf(seasons_scripts, top=5)". Entrambe queste funzioni accettano una lista di testi e il numero desiderato di topic (in questo caso, fissato a cinque). Entrambi i metodi seguono uno schema simile: viene creato un oggetto vettorizzatore, per trasformare gli script in ingresso in una rappresentazione vettoriale, successivamente viene costruito il modello, e viene eseguito il calcolo dei topic. Il risultato viene salvato all'interno di una lista che la quale conterrà tutti i topic estratti. All'interno di questa lista, per ogni topic, viene creato un dizionario che riporta l'indice del topic, le top words rappresentative di quel topic e infine il peso di ognuna di esse. Il peso, calcolato con la tecnica di pesatura delle parole TF-IDF (Term Frequency-Inverse Document Frequency) permette di valutare l'importanza relativa a ogni top word estratta, all'interno del corpus di documenti.

Create le due funzioni per il calcolo dei topic, ora è possibile estrarli per ogni stagione e successivamente per l'intera serie, attraverso rispettivamente due funzioni, che accettano in ingresso l'output restituito dalla funzione "get_episode_scripts" e il metodo di calcolo dei topic (LDA o NMF):

- "print_season_topics(seasons_scripts,method)": itera tra le chiavi del dizionario in ingresso e per ogni chiave, che rappresenta una stagione, esegue, se viene inserito il metodo LDA, la funzione "topics_lda", se viene inserito il metodo "NMF", la funzione "topics_nmf" e infine se viene inserito in ingresso qualsiasi altra stringa che non coincida con questi metodi viene stampato un messaggio di errore. I risultati calcolati per ogni stagione vengono inseriti in un dizionario, in cui ogni chiave rappresenta una stagione e i valori saranno i dizionari restituiti dalle funzioni di calcolo dei topic. Per visualizzare meglio l'output si utilizza un ciclo for per iterare tra le stagioni del dizionario e stampare a video i topic estratti per ogni stagione;
- "print_global_topics(seasons_scripts,method)": crea una lista all'interno della quale sono presenti tante stringhe quante sono le stagioni, e ogni stringa contiene al suo interno tutti gli script degli episodi della specifica stagione. Anche in questo caso se viene inserito il metodo LDA, viene eseguita la funzione "topics_lda", se viene inserito il metodo "NMF", viene eseguita la funzione "topics_nmf" e infine se viene inserito in ingresso qualsiasi altra stringa che non coincida con questi metodi viene stampato un messaggio di errore. I risultati vengono inseriti in una apposita lista, che contiene dunque, i cinque topic globali, con i rispettivi indici e pesature delle parole, rappresentativi della serie. Infine, con un ciclo for vengono stampati a video ordinatamente i topic estratti.

Questo codice, quindi, offre un'analisi approfondita dei dialoghi e delle descrizioni all'interno della serie televisiva, consentendo di scoprire i principali temi trattati nelle stagioni e nella serie nel suo complesso. L'utilizzo delle tecniche LDA e NMF consente di estrarre in modo sistematico e automatizzato gli argomenti chiave, fornendo una visione dettagliata della trama e delle dinamiche dei personaggi nel corso delle diverse stagioni arricchendo la nostra comprensione della narrazione complessiva.

4.1.2 Analisi dei topics: "The Vampire Diaries"

Ora andiamo a confrontare ed interpretare i topic risultati dall'analisi stagione per stagione.

Stagione	Topics LDA	Topics NMF
Stagione 1	Topic0: 'vampire', 'blood', 'party', 'hurt', 'mayor' Topic1: 'screwed', 'drew', 'deserve', 'felt', 'writes' Topic2: 'journal', 'pearl', 'vampire', 'dance', 'grab'	Topic0: 'school', 'home', 'party', 'diary', 'crystal' Topic1: 'journal', 'room', 'tomb', 'smile', 'grab' Topic2: 'blood', 'vampire', 'mayor', 'dance', 'founder'

	Topic3: 'yard', 'cross', 'fill', 'fault', 'shoot' Topic4: 'vampire', 'pearl', 'room', 'bree', 'blood'	Topic3: 'pearl', 'vampire', 'harper', 'house', 'storm' Topic4: 'football', 'team', 'field', 'cheerleader', 'tanner'
Stagione 2	Topic0: 'rose', 'brady', 'friend', 'moonstone', 'join' Topic1: 'stood', 'fails', 'focus', 'handsome', 'tonight' Topic2: 'jonas', 'blood', 'rose', 'greta', 'dagger' Topic3: rose', 'brady', 'slater', 'witch', 'outside' Topic4: 'curse', 'blood', 'werewolf', 'drink', 'moon'	Topic0: 'lockwood', 'moonstone', 'friend', 'blood', 'werewolf' Topic1: 'rose', 'slater', 'blood', 'help', 'tomb' Topic2: 'brady', 'approach', 'place', 'friend', 'werewolf' Topic3: 'jonas', 'dance', 'dagger', 'power', 'witch' Topic4: 'greta', 'maddox', 'blood', 'curse', 'katerina'
Stagione 3	Topic0: 'sage', 'dead', 'stake', 'family', 'necklace' Topic1: 'respond', 'desperate', 'crossbow', 'entrance', 'hard' Topic2: 'coffin', 'hybrid', 'dead', 'brother', 'side' Topic3: 'throat', 'cause', 'disappears', 'destroying', 'read' Topic4: 'polite', 'waist', 'master', 'smarter', 'save'	Topic0: 'coffin', 'hybrid', 'family', 'house', 'daniel' Topic1: sage', 'ring', 'journal', 'woman', 'stake' Topic2: 'necklace', 'gloria', 'dead', 'head', 'witch' Topic3: 'stake', 'dance', 'rose', 'white', 'dead' Topic4: 'mother', 'family', 'father', 'evening', 'dagger'
Stagione 4	Topic0: 'blood', 'attacker', 'nail', 'cure', 'feed' Topic1: 'forth', 'girl', 'paper', 'house', 'apparent' Topic2: 'large', 'obviously', 'nobody', 'unleashing', 'biggest' Topic3: 'cure', 'shane', 'blood', 'vaughn', 'witch' Topic4: 'nearby', 'sadness', 'master', 'shell', 'killed'	Topic0: 'shane', 'cure', 'hunter', 'chris', 'professor' Topic1: 'cure', 'smiling', 'vamp', 'prom', 'girl' Topic2: vaughn', 'cure', 'cave', 'veil', 'graduation' Topic3: 'blood', 'officer', 'pastor', 'council', 'begin' Topic4: 'adrian', 'hybrid', 'bond', 'sire', 'blood'
Stagione 5	Topic0: 'amara', 'anchor', 'observed', 'awakens', 'costume' Topic1: 'hidden', 'anti', 'accident', 'psychopath', 'blacked' Topic2: 'traveler', 'vision', 'witch', 'onto', 'redeemed' Topic3: 'traveler', 'side', 'house', 'help', 'people' Topic4: 'called', 'undo', 'shape', 'round', 'weakly'	Topic0: 'maxfield', 'head', 'brother', 'house', 'friend' Topic1: amara', 'side', 'cure', 'anchor', 'thousand' Topic2: 'traveler', 'witch', 'doppelga', 'side', 'vision' Topic3: 'mother', 'church', 'hallucinating', 'smile', 'venom' Topic4: 'whitmore', 'cell', 'augustine', 'cage', 'year'
Stagione 6	Topic0: 'blood', 'sanguinem', 'magic', 'vampire', 'ascendant' Topic1: 'exhales', 'music', 'gladly', 'certain', 'phesmatos' Topic2: 'makeshift', 'perfect', 'eclipse', 'picture', 'pass' Topic3: 'tripp', 'door', 'grab', 'blood', 'face' Topic4: 'abruptly', 'standing', 'magic', 'plane', 'smashing'	Topic0: 'sigh', 'moment', 'blood', 'humanity', 'human' Topic1: 'magic', 'sanguinem', 'merge', 'blood', 'knife' Topic2: 'tripp', 'face', 'door', 'bourbon', 'vampire' Topic3: 'exhales', 'cell', 'final', 'office', 'click' Topic4: 'sangina', 'mearma', 'world', 'door', 'funeral'
Stagione 7	Topic0: 'hole', 'anxious', 'ring', 'miserable', 'scalpel' Topic1: 'vampire', 'armory', 'vault', 'door', 'smile' Topic2: 'explain', 'sensing', 'unhappy', 'street', 'nodding' Topic3: 'smile', 'beau', 'sword', 'vampire', 'door' Topic4: 'sudden', 'cliff', 'debate', 'awake', 'circling'	Topic0: 'smile', 'beau', 'stone', 'house', 'town' Topic1: 'armory', 'vault', 'door', 'flash', 'vampire' Topic2: 'sword', 'beau', 'baby', 'smile', 'doctor' Topic3: 'hammond', 'officer', 'memphis', 'captain', 'fraternity' Topic4: 'vampire', 'baby', 'smile', 'crowd', 'door'
Stagione 8	Topic0: 'knife', 'spilled', 'belief', 'beer', 'food' Topic1: 'wife', 'begin', 'humor', 'happening', 'pitchfork' Topic2: 'time', 'girl', 'kill', 'siren', 'hell' Topic3: 'piece', 'sentence', 'screwed', 'also', 'sacrifice' Topic4: 'loophole', 'video', 'gave', 'closed', 'miss'	Topic0: 'siren', 'girl', 'georgie', 'fight', 'help' Topic1: 'beer', 'knife', 'taste', 'hand', 'grill' Topic2: 'grill', 'message', 'lifetime', 'closed', 'stabbing' Topic3: 'christmas', 'time', 'people', 'kill', 'student' Topic4: 'incendia', 'dagger', 'hell', 'hellfire', 'girl'

Analizzando i topic estratti per la prima stagione, è evidente che sia il metodo LDA che il metodo NMF hanno rivelato alcuni temi chiave. Entrambi i modelli hanno identificato temi legati ai vampiri, situazioni problematiche e conflitti. Inoltre, sono emersi anche temi legati alla vita quotidiana e all'ambiente domestico.

I termini chiave come "vampire" e "blood" riflettono chiaramente i temi centrali della prima stagione, e dell'intera serie. Altri termini come "party" e "dance" riflettono le dinamiche sociali presenti nella trama, mentre parole come "tomb" richiamano l'atmosfera oscura e misteriosa che pervade la stagione.

Tuttavia, è interessante notare che con il modello LDA, i topic identificati sembrano avere una connessione meno diretta con la trama principale. Molti dei termini sono generici e ciò potrebbe suggerire che LDA ha difficoltà nell'estrazione di informazioni specifiche e rilevanti dalla trama. Al contrario, il modello NMF sembra identificare topic più strettamente correlati alla trama, come evidenziato dai termini "school", "home", "party" e "diary", che riflettono luoghi e situazioni chiave nella storia. Questo suggerisce che NMF sia più efficace nel catturare gli elementi centrali della trama, concentrando su eventi, luoghi e oggetti significativi. Nella seconda stagione, sia il modello LDA che il modello NMF hanno estratto topic direttamente collegati alla trama della serie. In particolare, sono emersi chiaramente temi come le maledizioni, i licantropi e la luna. Questi temi sono stati individuati dai topic 0 e 4 per LDA e principalmente dal topic 0 per NMF. Le parole chiave come "werewolf," "moonstone," e "curse" richiamano l'importanza centrale dei licantropi nella storia e la loro connessione con il mondo magico. Il richiamo al tema del sangue, che è un elemento ricorrente in tutte le stagioni, è anch'esso presente. Nei topic 3 di entrambi i metodi, compare la parola "witch," evidenziando il tema delle streghe che svolgono un ruolo rilevante durante la stagione. Nella terza stagione, sia il modello NMF che il modello LDA hanno estratto temi simili, condividendo numerosi termini chiave come "dead," "coffin," "stake," e "family," che richiamano fortemente temi legati alla morte, alle battaglie e alle dinamiche familiari. Un momento cruciale di questa stagione è rappresentato dalla creazione di un esercito di ibridi, una nuova figura fantastica introdotta nella serie. Questo evento è riflesso nel topic 2 identificato dal modello LDA e nel topic 0 del modello NMF. Nella quarta stagione, emerge chiaramente il tema della cura per il vampirismo e del sangue, che costituiscono un elemento centrale nella storia. Tuttavia, sembra che i topic identificati dal modello NMF siano più direttamente correlati agli elementi chiave della trama. Ad esempio, i termini "hunter," "vamp," "graduation," "hybrid," e "cave" sembrano essere più rilevanti per la storia principale, evidenziando il coinvolgimento dei cacciatori di vampiri e il ruolo degli ibridi nella trama. Nella quinta stagione, i topic estratti dai due modelli presentano numerosi elementi di somiglianza, inclusi i temi come "anchor," "traveler," "witch," e "Whitmore", "Amara" che sono strettamente connessi alla trama. Nella sesta stagione, i topic estratti dai due modelli presentano notevoli somiglianze. Entrambi affrontano tematiche legate al mondo magico, al sangue e ai vampiri, come indicato chiaramente dai termini chiave come 'blood,' 'sanguinem,' 'magic,' 'vampire,' e "face." Tuttavia, nel topic 0 generato dal modello NMF emerge un tema particolarmente rilevante, ossia quello dell'umanità, evidenziato dalle parole "human" e "humanity". Nella settima stagione, i topic estratti dai due modelli presentano alcune similitudini, concentrando su tematiche come "vampire," "vault," e "Armory," quest'ultima essendo un'organizzazione misteriosa di centrale importanza nella trama. Si può notare come gli LDA Topics sembrano mettere maggiormente in evidenza emozioni intense e conflitti, come dimostrato dai termini "miserable," "anxious," e "unhappy." D'altra parte, i NMF Topics sembrano coprire una gamma più ampia di elementi, inclusi quelli legati all'azione e alle relazioni romantiche, come indicato dai termini 'beau,' 'baby,' e 'smile.' L'ultima stagione affronta una serie di temi fondamentali, tra cui la lotta tra umanità e immortalità, il senso di colpa e il sacrificio per il bene comune, questi temi emergono chiaramente dai topic generati da entrambi i modelli e sono in parte legati anche alla decisione cruciale di rimanere immortali o tornare umani. Gli altri topic riflettono tutti gli altri elementi fantastici e le epiche battaglie che caratterizzano la stagione finale.

In sintesi, dall'analisi dei topic estratti dai modelli LDA e NMF per le diverse stagioni di "The Vampire Diaries" si può notare come entrambi i modelli sono stati in grado di identificare temi chiave delle singole stagioni, tra cui la presenza dei vampiri, le sfide e i conflitti, così come elementi legati alla vita quotidiana e all'ambiente domestico. Tuttavia, ci sono state alcune differenze nelle prestazioni dei due modelli. Il modello NMF sembra aver catturato in modo più efficace gli elementi centrali della trama, concentrando su eventi, luoghi e oggetti significativi. Questo è particolarmente evidente nelle stagioni in cui sono emersi temi specifici come le maledizioni e i licantropi nella seconda stagione, la ricerca della cura per il vampirismo nella quarta stagione, e il ritorno all'umanità nella sesta stagione. D'altra parte, il modello LDA ha talvolta prodotto topic con

connessioni meno dirette alla trama principale, con termini più generici. Tuttavia, entrambi i modelli hanno contribuito a mettere in evidenza elementi distintivi e atmosfere delle diverse stagioni.

Topics LDA:	Topics NMF:
Topic0: 'feared', 'altar', 'sicker', 'devil', 'hellscape'	Topic0: 'armory', 'siren', 'stone', 'heretic', 'frown'
Topic1: 'traveler', 'hybrid', 'whitmore', 'necklace', 'sanguinem'	Topic1: 'shane', 'hybrid', 'cure', 'hunter', 'dagger'
Topic2: 'pearl', 'armory', 'brotherhood', 'founder', 'disappoint'	Topic2: 'traveler', 'whitmore', 'doppelga', 'augustine', 'nger'
Topic3: 'ruining', 'constantly', 'despair', 'tense', 'bayou'	Topic3: 'pearl', 'johnathan', 'founder', 'necklace', 'mayor'
Topic4: 'shane', 'werewolf', 'moonstone', 'cure', 'dagger'	Topic4: 'moonstone', 'rose', 'werewolf', 'brady', 'dagger'

Globalmente, si può dire che entrambi i modelli rilevano chiaramente temi come la ricerca della cura per il vampirismo e le varie figure che hanno giocato un ruolo significativo nella trama, tra cui viaggiatori, sirene, eretici, ibridi, licantropi. Inoltre, tra i topic generati da NMF, emerge anche il tema dei doppelganger, un elemento fondamentale nella storia. Concludendo, si può dire che entrambi i metodi hanno identificato i principali temi centrali della serie, anche se potrebbero differire leggermente nell'organizzazione delle parole chiave all'interno dei topic. Tuttavia, entrambi i modelli si sono dimostrati validi strumenti per analizzare in profondità i temi predominanti nella serie.

Di seguito, vengono mostrate le word cloud relative a ciascun topic, costruite con la funzione "create_wordclouds(topics)" contenuta nel file "4_topic_modeling.py", le quali forniscono una chiara rappresentazione visuale delle parole chiave e dell'importanza relativa di ciascun termine all'interno del topic corrispondente. Questo strumento visuale sarà un valido supporto per comprendere meglio le tematiche trattate all'interno delle stagioni di "The Vampire Diaries" e per individuare i punti salienti di ciascun topic.



4.1.3 Analisi dei topic: "The Big Bang Theory"

Ora andiamo a confrontare ed interpretare i topic risultati dall'analisi stagione per stagione.

Stagione	Topics LDA	Topics NMF
Stagione 1	Topic0: 'wave', 'bitch', 'spill', 'panic', 'colonoscopy' Topic1: 'party', 'sister', 'costume', 'sick', 'everybody' Topic2: 'machine', 'knock', 'door', 'hallway', 'work' Topic3: 'team', 'question', 'halo', 'answer', 'head' Topic4: 'parent', 'building', 'papa', 'sperm', 'funny'	Topic0: 'sister', 'year', 'work', 'thirty', 'love' Topic1: 'party', 'costume', 'knock', 'kurt', 'peanut' Topic2: 'team', 'answer', 'question', 'bowl', 'physic' Topic3: 'machine', 'hundred', 'travel', 'sell', 'dollar' Topic4: 'soup', 'knock', 'sick', 'night', 'happy'
Stagione 2	Topic0: 'knock', 'friend', 'living', 'night', 'three' Topic1: 'comic', 'secret', 'thursday', 'book', 'shoot' Topic2: 'mother', 'money', 'course', 'drunk', 'people' Topic3: 'train', 'gift', 'dave', 'online', 'level' Topic4: 'robot', 'hearty', 'pathetic', 'creepy', 'underwear'	Topic0: 'money', 'people', 'living', 'door', 'great' Topic1: 'knock', 'north', 'strike', 'model', 'house' Topic2: 'comic', 'book', 'thursday', 'store', 'space' Topic3: 'secret', 'friend', 'college', 'spock', 'date' Topic4: 'robot', 'train', 'mother', 'summer', 'drive'
Stagione 3	Topic0: 'ring', 'knock', 'work', 'kitty', 'bowling' Topic1: 'wheaton', 'football', 'date', 'friend', 'pole' Topic2: 'knock', 'friend', 'mother', 'moon', 'favour'	Topic0: 'friend', 'night', 'date', 'morning', 'moon' Topic1: 'knock', 'kitty', 'sauce', 'door', 'friend' Topic2: 'wheaton', 'wesley', 'crusher', 'bowling', 'ball'

	Topic3: 'friend', 'excellent', 'operating', 'roll', 'salmon' Topic4: 'knock', 'tattoo', 'girlfriend', 'chocolate', 'agree'	Topic3: 'ring', 'mother', 'bean', 'experiment', 'electron' Topic4: 'football', 'friend', 'shot', 'game', 'toast'
Stagione 4	Topic0: 'movie', 'love', 'line', 'secret', 'great' Topic1: 'starfleet', 'family', 'enough', 'goodness', 'laundry' Topic2: 'knock', 'girlfriend', 'night', 'robot', 'mother' Topic3: 'knock', 'room', 'sleep', 'bernie', 'hotel' Topic4: 'troll', 'night', 'knock', 'world', 'sister'	Topic0: 'sister', 'mother', 'friend', 'door', 'night' Topic1: 'knock', 'troll', 'shower', 'kitty', 'night' Topic2: 'party', 'knock', 'movie', 'line', 'girl' Topic3: 'girlfriend', 'shamy', 'fowler', 'farrah', 'robot' Topic4: 'card', 'trick', 'magic', 'security', 'agent'
Stagione 5	Topic0: 'time', 'school', 'space', 'hair', 'kind' Topic1: 'wheaton', 'loop', 'wedding', 'space', 'launch' Topic2: 'knock', 'woman', 'cake', 'pocket', 'indian' Topic3: 'knock', 'bird', 'office', 'date', 'lizard' Topic4: 'office', 'time', 'dear', 'please', 'wrong'	Topic0: 'time', 'mother', 'wedding', 'friend', 'woman' Topic1: 'knock', 'chair', 'pack', 'playing', 'billy' Topic2: 'office', 'bird', 'spock', 'rock', 'lizard' Topic3: 'hair', 'roommate', 'haircut', 'agreement', 'astronaut' Topic4: 'space', 'wheaton', 'married', 'loop', 'soyuz'
Stagione 6	Topic0: 'forty', 'professor', 'flag', 'woman', 'knock' Topic1: 'party', 'closet', 'letter', 'space', 'costume' Topic2: 'knock', 'space', 'hawking', 'stephen', 'paper' Topic3: 'spot', 'live', 'parking', 'knock', 'flag' Topic4: 'dungeon', 'dragon', 'vega', 'happy', 'knock'	Topic0: 'space', 'comic', 'party', 'love', 'spot' Topic1: 'knock', 'flag', 'live', 'davis', 'spider' Topic2: 'dungeon', 'christmas', 'ogre', 'dragon', 'vega' Topic3: 'work', 'office', 'science', 'professor', 'woman' Topic4: 'hawking', 'stephen', 'forty', 'friend', 'paper'
Stagione 7	Topic0: 'funeral', 'puzzle', 'team', 'hunt', 'bitch' Topic1: 'real', 'advanced', 'married', 'excellent', 'shoulder' Topic2: 'comic', 'time', 'ticket', 'darth', 'jones' Topic3: 'taught', 'store', 'project', 'miss', 'drink' Topic4: 'knock', 'time', 'great', 'night', 'friend'	Topic0: 'knock', 'time', 'mother', 'work', 'kind' Topic1: 'element', 'time', 'train', 'romantic', 'without' Topic2: 'movie', 'woman', 'gorilla', 'night', 'seeing' Topic3: 'comic', 'rock', 'time', 'puzzle', 'book' Topic4: 'funeral', 'annoying', 'proton', 'star', 'professor'
Stagione 8	Topic0: 'tonight', 'bear', 'professor', 'show', 'probably' Topic1: 'time', 'knock', 'house', 'money', 'mother' Topic2: 'knock', 'flag', 'episode', 'clown', 'date' Topic3: 'perfect', 'yellow', 'people', 'fact', 'clean' Topic4: 'wave', 'tech', 'planet', 'speech', 'store'	Topic0: 'time', 'money', 'future', 'movie', 'mine' Topic1: 'knock', 'prom', 'flag', 'anxiety', 'mother' Topic2: 'date', 'relationship', 'married', 'pitch', 'year' Topic3: 'store', 'comic', 'puzzle', 'book', 'barry' Topic4: 'surgery', 'christmas', 'gift', 'room', 'parent'
Stagione 9	Topic0: 'zone', 'baby', 'bernie', 'daddy', 'tall' Topic1: 'ever', 'song', 'secret', 'game', 'baby' Topic2: 'line', 'wine', 'helium', 'people', 'apology' Topic3: 'patent', 'contract', 'work', 'barry', 'university' Topic4: 'knock', 'birthday', 'made', 'wedding', 'batman'	Topic0: 'knock', 'wedding', 'line', 'marriage', 'woman' Topic1: 'birthday', 'batman', 'party', 'knock', 'meeting' Topic2: 'baby', 'helium', 'money', 'ever', 'contract' Topic3: 'jones', 'house', 'room', 'spock', 'feel' Topic4: 'flag', 'valentine', 'rabbit', 'wine', 'woman'
Stagione 10	Topic0: 'work', 'comic', 'rule', 'knock', 'notebook' Topic1: 'accuse', 'engineer', 'roof', 'questioned', 'sort' Topic2: 'project', 'gone', 'machine', 'bedroom', 'boyfriend' Topic3: 'baby', 'flag', 'room', 'help', 'together' Topic4: 'halloween', 'read', 'model', 'hawking', 'mention'	Topic0: 'brother', 'mother', 'machine', 'baby', 'room' Topic1: 'flag', 'living', 'bedroom', 'wall', 'room' Topic2: 'work', 'project', 'pregnant', 'feel', 'four' Topic3: 'comic', 'room', 'baby', 'bedroom', 'five' Topic4: 'knock', 'rock', 'woman', 'cafeteria', 'angry'

Analizzando i topic estratti per la prima stagione, è possibile notare come sia il metodo LDA che il metodo NMF abbiano rilevato il tema principale della serie legato alla vita quotidiana e all'ambiente scientifico. Nella prima stagione termini chiave come "work", "physic" e "machine" riflettono chiaramente situazioni legate all'ambiente di lavoro, così come "door", "hallway" riferiscono temi di vita quotidiana in appartamento, luogo in cui si svolgono la quasi totalità delle vicende nella serie. In questa stagione vengono rilevati anche i termini "party", "costume", "funny" e "night" che indicano feste o attività sociali che percorrono la vita ricreativa dei

protagonisti. Nella seconda stagione compare insieme a parole riguardanti situazioni di vita quotidiana, anche il topic dei fumetti, infatti parole come "comic", "book" e "store" sono legate agli interessi ricreativi dei protagonisti, come riportato dal topic 2 di NMF. Inoltre entrambi i modelli rilevano la parola "money", sicuramente legata al tema del lavoro, ma probabilmente in riferimento maggiore alla situazione economica precaria di Penny all'inizio della serie televisiva. Anche nella terza stagione compaiono altre parole che sono prevalentemente legate a momenti di socializzazione e appuntamenti, come "date", "friend", "bowling" e "game", ma anche parole che riguardano interazioni quotidiane come "knock", riferimento all'iconica routine di Sheldon nel bussare tre volte e chiamando tre volte la persona al di là della porta; questa parola appare nello stesso topic insieme a "door", definendo la precisione del modello nella rilevazione dell'argomento. Insieme a "knock" viene anche rilevato "kitty" sempre in riferimento a Sheldon nella sua canzone "soft kitty" che canta per sé stesso o per gli amici quando sono malati o hanno bisogno di conforto. Si può finora notare come i topic risultanti da NMF sembrano essere più precisi, sempre nella terza stagione il secondo topic mette insieme "Wheaton" un personaggio "rivale" di Sheldon con "bowling", l'attore Wil Wheaton infatti è noto per essere un grande appassionato di giochi da tavolo e in particolare del bowling. Sempre in questa stagione appaiono i termini "experiment" e "electron", riportando ancora lo sguardo sui temi scientifici della serie. Nella quarta stagione invece oltre ai temi precedentemente rilevati anche nelle altre stagioni NMF rivela "girlfriend", "farrah" e "fowler", questi ultimi due termini compongono il nome completo di Amy che in seguito diventerà la fidanzata di Sheldon; si possono notare anche le parole "card", "trick" e "magic" nel topic 4 del modello NMF, che potrebbero riguardare un momento ricreativo di trucchi con le carte. Nella quinta stagione appaiono per la prima volta parole come "astronaut", "launch" e "space", rilevate dal topic1 del modello LDA, probabilmente si riferiscono a Howard che proprio nella quinta stagione ha l'opportunità di andare nello spazio. Nella sesta stagione entrambi i modelli hanno estratto temi simili, individuando i termini "space", "stephen" e "hawking" che fanno riferimento a conversazioni scientifiche legate all'astronomia, uno dei temi più trattati nella serie; inoltre viene rilevato anche il tema del gioco di ruolo preferito dai protagonisti, infatti le parole "dungeon", "dragon", "ogre" e "vega" fanno riferimento al gioco Dungeons&Dragons. NMF in questa stagione raggruppa anche i termini "work", "office", "science" e "professor", che fanno riferimento al tema accademico-lavorativo dei protagonisti. Nella settima stagione entrambi i modelli rilevano di nuovo il tema dell'intrattenimento con "comic", "puzzle", "book" e i personaggi "darth" e "jones". In particolare in questa stagione, il modello NMF nel topic 4 associa le parole "professor", "proton", "funeral" e "star" in un unico topic, proprio nella settima stagione infatti muore il personaggio del professor Proton, questo avvenimento crea un impatto emotivo soprattutto su Sheldon in quanto lo considerava un modello da seguire. Nell'ottava stagione i modelli oltre ai temi già rilevati sulla quotidianità e sull'intrattenimento, compare nuovamente il tema sulle relazioni con le parole "date", "relationship" e "married", ma anche temi riguardanti il Natale con parole come "christmas" e "gift". Nelle ultime due stagioni compare il tema della genitorialità con le parole "baby", "daddy", "house", "pregnant" e familiare con le parole "mother" e "brother", oltre ai soliti temi legati alla vita quotidiana e sociale nell'appartamento, e al tema riguardante l'interesse per i fumetti, con le parole "comic", "jones" e "spock".

Nell'analisi stagione per stagione, sia il modello LDA che il modello NMF hanno estratto topic significativi che affrontano i temi principali presenti nella serie TV "The Big Bang Theory", dimostrando di essere entrambi piuttosto efficaci nell'estrazione di topic rilevanti e significativi per la serie, pur differendo leggermente nell'organizzazione delle parole chiave all'interno dei topic. Dall'analisi dei topic si può notare come quelli restituiti dal modello NMF siano più precisi rispetto a quelli restituiti dal modello LDA, che invece risultano essere più dispersivi e generici; ciò potrebbe suggerire che LDA ha difficoltà nell'estrazione di informazioni specifiche. Al contrario, il modello NMF sembra identificare topic più strettamente correlati, come evidenziato dai termini "dungeon", "dragon", "ogre" e "vega" specifici per il gioco di ruolo Dungeons&Dragons. Questo suggerisce che NMF sia più efficace nel definire topic con parole più legate nel contesto tra loro. Tuttavia si può concludere che entrambi i metodi sono stati in grado di estrarre i temi rilevanti della serie, catturando efficacemente l'atmosfera leggera e quotidiana che la caratterizza.

Volgendo lo sguardo su una panoramica dei topics sull'intera serie nella tabella sottostante, si può notare come i topic restituiti da NMF siano molto più in relazione tra loro rispetto a quelli più generici restituiti dal modello LDA. Entrambi i topic hanno restituito le parole "cheesecake" e "factory" che riferiscono al luogo di lavoro di Penny, emozioni come "worried", "excited", "lost" e "misunderstanding" che riflettono il tema umano dei protagonisti, come rilevato dai grafici sul sentiment nei precedenti punti 2 e 3; compare anche il tema scientifico con "heisenberg", "hawking", "robot" e "astronaut" e quello ricreativo con "xbox", "dungeon" e "flag" (in riferimento al gioco immaginario "Divertiamoci con le bandiere" creato da Sheldon durante la serie).

Topics LDA:	Topics NMF:
Topic0: 'faeces', 'anti', 'scrape', 'heisenberg', 'misunderstanding'	Topic0: 'element', 'xbox', 'cheesecake', 'howie', 'factory'
Topic1: 'flag', 'lost', 'cheesecake', 'worried', 'rock'	Topic1: 'halo', 'robot', 'winkle', 'cheesecake', 'building'
Topic2: 'candidate', 'display', 'stream', 'organism', 'nipple'	Topic2: 'flag', 'prom', 'worried', 'excited', 'howie'
Topic3: 'court', 'repeat', 'diagnosis', 'tackle', 'bending'	Topic3: 'flag', 'helium', 'valentine', 'meemaw', 'thor'
Topic4: 'wheelchair', 'sushi', 'cloud', 'skip', 'gathering'	Topic4: 'flag', 'hawking', 'astronaut', 'howie', 'dungeon'

Di seguito, vengono mostrate le word cloud relative a ciascun topic, costruite con la funzione “create_wordclouds(topics)” contenuta nel file “4_topic_modeling.py”, le quali forniscono una chiara rappresentazione visuale delle parole chiave e dell’importanza relativa di ciascun termine all’interno del topic corrispondente. Questo strumento visuale sarà un valido supporto per comprendere meglio le tematiche trattate all’interno delle stagioni di “The Big Bang Theory” e per individuare i punti salienti di ciascun topic.



4.2 Visualizzazioni HTML

Per ottenere un'idea visuale e intuitiva dei principali temi trattati all'interno delle serie TV, emersi dalla modellazione dei topic nel dettaglio, è stato sviluppato un codice che consente la creazione di file HTML contenenti tali informazioni per entrambe le serie.

Nella visualizzazione interattiva, infatti, si possono notare a sinistra dei cerchi, le “bolle di testo” rappresentanti i topic. La grandezza di questi cerchi indica l’importanza del topic all’interno dell’intero corpus di testo. Un cerchio più grande indica un topic più rilevante e viceversa. Ogni cerchio contiene le parole chiave (presenti a sinistra) associate a quel topic, offrendo un’idea immediata degli argomenti principali rappresentati da quel determinato topic. Queste parole istantaneamente vengono evidenziate in rosso nelle barre delle parole presenti a sinistra. Le barre di fatto rappresentano l’importanza relativa delle parole chiave all’interno di ciascun topic. Una barra più lunga indica che quella parola è più significativa per quello stesso topic e viceversa se più corta.

4.2.1 Codice

Il codice relativo alla visualizzazione dei risultati della Topic Modeling è presente nel file '4_visualizzazioni_html.py'

Dopo aver importato le librerie necessarie di nltk per la pulizia dei testi e pyLDAvis per creare i modelli LDA, sono state create delle liste con le parole da escludere durante la pulizia del testo: “excluded_words_tbbt”, “excluded_words_tvd” e una lista di parole extra “excluded_words_words” contenente numeri e altre parole non utili che vengono aggiunte alle parole da eliminare delle due serie.

Il preprocessing del testo è stato assegnato alla funzione ‘preprocess_text’ che accetta un testo come input e le parole da escludere, restituendo una lista di token del testo ripulito dopo aver applicato alcune operazioni di pre-elaborazione. È stato quindi inizializzato un oggetto lemmatizer basato su WordNet, per eseguire la lemmatizzazione e sono state ottenute le stopwords in inglese utilizzando “stopwords.words('english')”. Il testo è stato quindi suddiviso in token, ossia parole o sequenze di caratteri significativi, utilizzando la funzione “word_tokenize” di nltk, ottenendo singole parole. Nella pulizia dei token sono state eliminate le parole che contengono caratteri diversi da lettere con ‘re.sub’, e sono state convertite in carattere minuscolo. Poi si verifica la parola pulita, se questa non è presente in ‘excluded_words’ viene lemmatizzata; si verifica se non sia una stopword, se è troppo corta (meno di 3 caratteri) o troppo lunga (oltre 15 caratteri), in questi casi le parole “approvate” vengono aggiunte alla lista ‘clean_tokens’. Infine è stata restituita la lista dei token puliti.

Per poter visualizzare il modello LDA per ogni stagione, si è iterato su ognuna di queste all'interno del percorso del corpus generale. Nella funzione ‘perform_topic_modeling_(corpus_path, visualization_path, excluded_words)’ infatti si è iterato attraverso le directory dei corpus principali delle serie, in cui è stato creato il percorso di ogni stagione usando ‘os.path.join’, salvato nella variabile ‘season_path’. Viene poi creata una lista ‘documents’ vuota, che conterrà i documenti testuali estratti dai file della stagione corrente, e per ognuno di questi, viene creato il percorso completo del file sempre con ‘os.path.join’ verificando se è un file utilizzando ‘os.path.isfile’. Se il file è valido viene quindi aperto in modalità lettura e viene aggiunto il testo precedentemente elaborato dalla funzione ‘preprocess_test’ alla lista vuota ‘documents’.

A questo punto viene creato il percorso per la directory ‘lda_model’ all'interno della stagione con ‘os.path.join’, creandola qualora non esista con la funzione ‘os.makedirs’.

Viene poi creato un dizionario ‘dictionary’ utilizzando ‘corpora.Dictionary(documents)’ di Gensim basato sui documenti pre-elaborati che vengono passati come argomento alla funzione. È stato così creato il corpus rappresentato come lista di tuple (bag of words) in cui ognuna contiene l'identificatore della parola nel dizionario e il conteggio delle occorrenze della parola nel documento usando ‘doc2bow’.

È stato poi creato e addestrato il modello LDA (Latent Dirichlet Allocation) utilizzando ‘models.LdaModel’ di Gensim sul corpus, stabilendo 5 topic e 20 iterazioni (passes). Il parametro passes, è un'operazione che indica quante volte il modello passa attraverso il corpus durante l'addestramento. Durante ogni passaggio il modello cerca di migliorare la distribuzione dei topic rispetto ai documenti del corpus. Infine il modello LDA addestrato viene salvato in un file lda all'interno del percorso specificato in precedenza ‘lda_model_path’.

Per la visualizzazione interattiva dei risultati del modello è stato preparato un oggetto ‘vis_data’ utilizzando ‘gensimvis.prepare(lda_model, corpus, dictionary)’.

Per poter visualizzare i temi di tutte le stagioni, è stata creata una struttura di directory. Quindi, una volta creata una cartella specifica per ogni stagione, al suo interno è stata creata una directory ‘lda_visualization’ e all'interno di questa è stata salvata la visualizzazione LDA in un file html chiamato ‘index.html’.

Il codice finora descritto ha creato i file html per ogni singola stagione della serie televisiva.

Per avere una visione generale dei topic sull'intero corpus di documenti combinati dalle diverse stagioni rappresentanti l'intera serie TV, è stata definita la funzione ‘perform_topic_modeling_global(documents, visualization_path)’, che prende la lista di documenti preprocessati, e il percorso in cui verrà salvata la visualizzazione globale. Questa funzione quindi crea un dizionario e un corpus basati sui documenti passati

come argomento (documents), poi crea e addestra il modello LDA sempre con 5 topic utilizzando ‘models.LdaModel()’.

Per la visualizzazione viene preparato l’oggetto ‘vis_data’, come fatto in precedenza per le stagioni, utilizzando ‘gensimvis.prepare()’, salvandola in un file HTML all’interno della directory ‘visualization_path’ creata in precedenza.

La funzione che raccoglie i documenti dai file del corpus è definita con ‘collect_documents(corpus_path, excluded_words)’, questa funzione restituisce una lista di documenti preprocessati che, iterando attraverso le directory nel percorso del corpus, legge ogni file di testo preprocessandolo grazie alla funzione “preprocess_txt” e escludendo le parole in ‘excluded_words’, e infine aggiungendo il documento preprocessato alla lista dei documenti.

Il risultato del codice ‘4_visualizzazioni_html.py’ è quindi la creazione di due nuove directory “topic_modeling_tbbt” e “topic_modeling_tvd” al cui interno sono contenute le sottodirectory per ogni stagione contenenti la rispettiva visualizzazione html dei 5 topic e la visualizzazione html globale.

4.2.2 Analisi delle visualizzazioni: “The Big Bang Theory”

Come detto in precedenza le visualizzazioni HTML per “The Big Bang Theory” sono state create dal codice nel file ‘4_visualizzazioni_html.py’. L’analisi dei risultati ottenuti verrà condotta prima in modo specifico stagione per stagione, mentre successivamente a livello globale per individuare i temi principali e riassunti della serie, che permettano di ottenere una panoramica generale e globale della serie stessa.

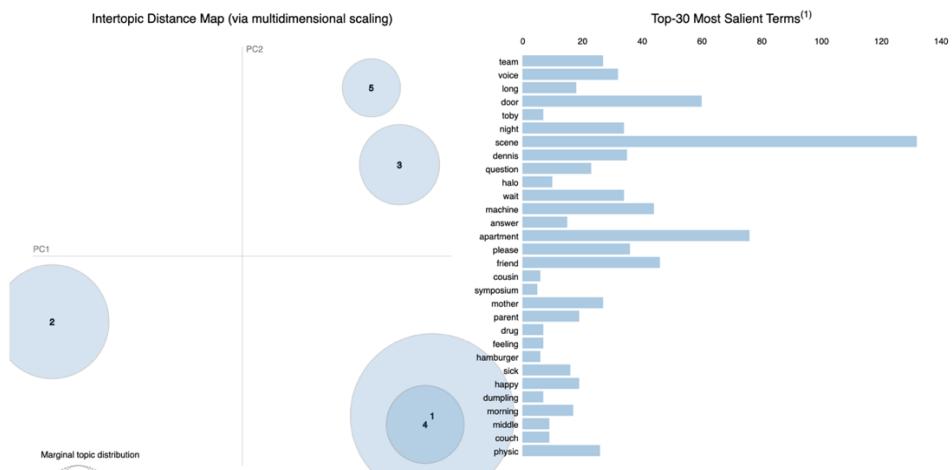
Analisi dei risultati per stagione

All’interno della directory “topic_modeling_tbbt” sono contenute delle sottodirectory relative a ciascuna stagione della serie. All’interno di ciascuna di esse è presente il file ‘index.html’ che permette di visualizzare ed esplorare i topic risultanti per ogni stagione.

Verranno di seguito presentate, per ogni stagione delle anteprime di questi file.

Dalla visualizzazione HTML per ogni stagione si può subito notare la distanza tra i cerchi dei topic tra loro, questo in quanto presumibilmente lavorando su una singola stagione alla volta, il modello riesce a cogliere meglio varie parole e a distinguere quindi più argomenti.

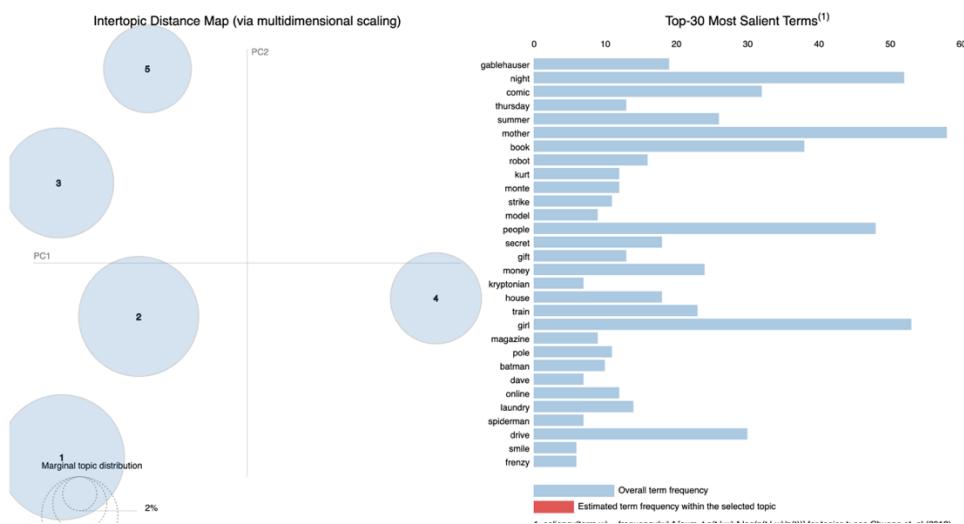
STAGIONE 1



I topic nella visualizzazione sono ben distanziati tra loro, ma si può notare come il quarto topic si sovrapponga al primo; questi due topic sembrano concentrarsi prevalentemente su temi quotidiani riguardanti

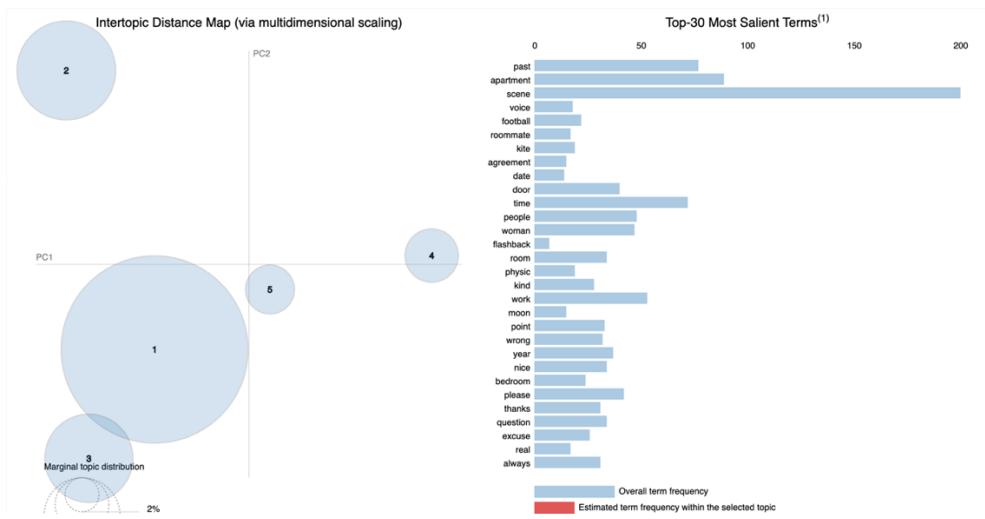
l'appartamento, riportando come particolarmente rilevanti le parole "apartment", "door", "machine", "room", ossia parole centrali nel tema della quotidianità della serie. eventi come feste e compleanni, riflettendo l'ambiente sociale e festoso della serie. Il secondo topic sembra concentrarsi su questioni relative al lavoro e alla vita quotidiana con riferimento a temi come "physic" e "work" altro tema centrale della serie. Il terzo sembra coinvolgere più parole legate al cibo, con parole come "order" e "waiter", "chicken" e "dumpling" e menzioni ai "tangerine". Il quinto topic sembra coinvolgere elementi di riabilitazione (forse da un intervento) con parole come "rehab", "feeling", "intervention", "drug". Questi topic in generale sembrano catturare diversi aspetti della vita quotidiana, delle interazioni sociali e conviviali, ossia i temi centrali della serie.

STAGIONE 2



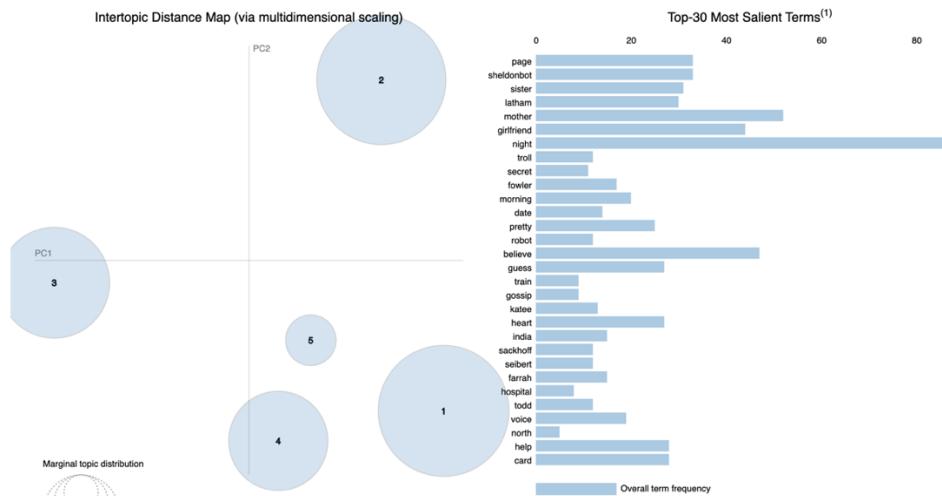
Anche in questa visualizzazione i topic sono ben distanziati tra loro, il primo topic sembra ruotare ancora attorno a scene quotidiane e momenti di vita degli amici, che spesso si svolgono nell'appartamento. Sembra esserci un'attenzione particolare su come trascorrono il tempo insieme con le parole "together", "time" e sulla loro routine lavorativa. Anche il secondo topic sembra trattare temi di relazioni e amicizia con parole come "help", "please", "thanks" e "question" che introducono i temi. Il terzo topic sembra toccare aspetti più complessi, tra cui temi come segreti e questioni finanziarie, con le parole "money", "work" e "problem" che possono indicare qualche tensione in questa serie, come già predetto dai precedenti grafici. Nel quarto topic si fa riferimento anche a festività come il Natale, "christmas" indicando che alcune scene potrebbero essere ambientate durante questo periodo, insieme anche a parole come "gift" e "night" e "year". Nel quinto topic invece emerge il tema dei fumetti, con le parole "book", "store" e supereroi con menzione di "Batman" e "Spiderman", indicando l'interesse dei personaggi per i fumetti.

STAGIONE 3



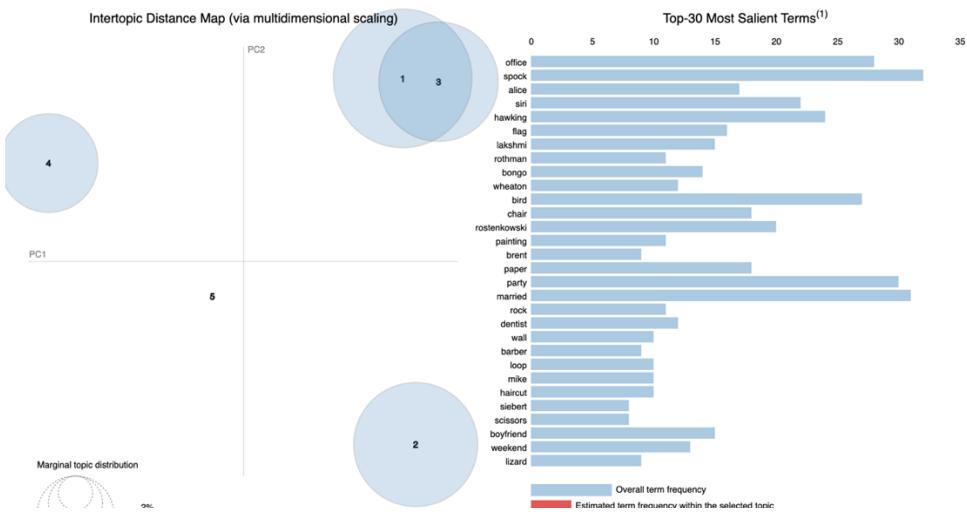
Anche qui i topic sembrano essere ben distanziati, il primo è quello più grande che racchiude presumibilmente i temi più rilevanti della stagione. Qui si trovano ancora parole che rimandano con una forte enfasi alle relazioni, dinamiche di convivenza e le interazioni quotidiane tra i personaggi principali all'interno degli appartamenti, con le parole "apartment", "night", "work", "girlfriend" e "love". Il secondo topic sembra concentrarsi maggiormente su attività ricreative e momenti di svago, con parole quali "football", "kite", "night", "game". Nel terzo invece parole "moon", "laser", "judge" e "comic" sembrano suggerire un interesse per la scienza e la fantascienza, che sono uno dei centri degli interessi dei personaggi nella serie televisiva. Nel quarto topic parole "flashback", "roommate", "elevator", "agreement" ma soprattutto la grande rilevanza che ha la parola "past" sembrano indicare una riflessione su eventi passati o cambiamenti nella dinamica della convivenza tra i personaggi all'interno dell'appartamento. Il quinto topic sembra concentrarsi ancora sulle relazioni dei personaggi, compaiono infatti "date" e "relationship" che suggeriscono sviluppi romantici o appuntamenti tra i personaggi. Ancora una volta i topic sembrano catturare i vari aspetti della serie riguardo le dinamiche del gruppo e dei loro interessi scientifici.

STAGIONE 4



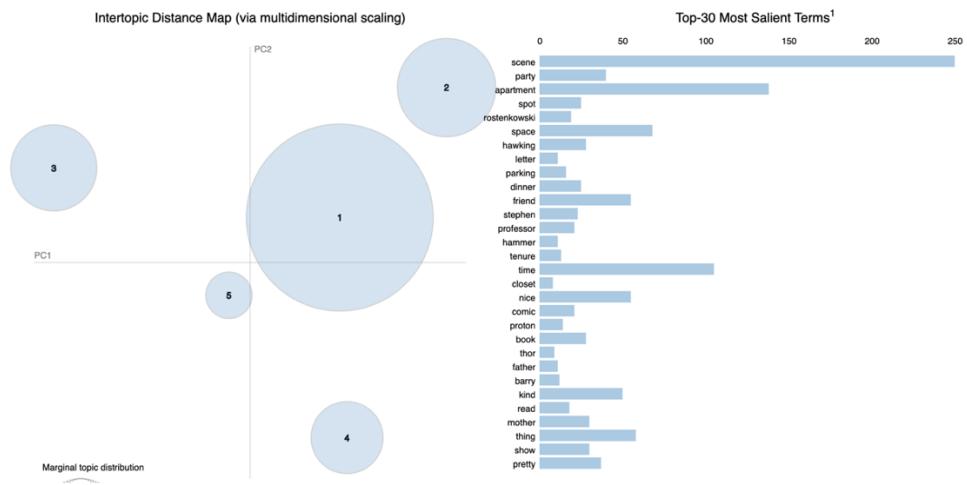
Anche nella quarta stagione i topic sono abbastanza distanziati tra loro. Come nelle stagioni precedenti il primo topic sembra ruotare ancora attorno a scene quotidiane, come fa più o meno anche il secondo, che sembra concentrarsi sulle attività degli amici, come guardare film nell'appartamento, con le parole "night", "movie", "tonight", "party", ciò evidenzia l'importanza dell'amicizia e delle attività di gruppo. Anche nel terzo sembrano esserci temi riguardanti le interazioni e argomenti scientifici con la parola "science" e "question". Il quarto topic e il quinto si concentrano ancora prevalentemente sulle relazioni con parole già presenti nelle altre stagioni.

STAGIONE 5



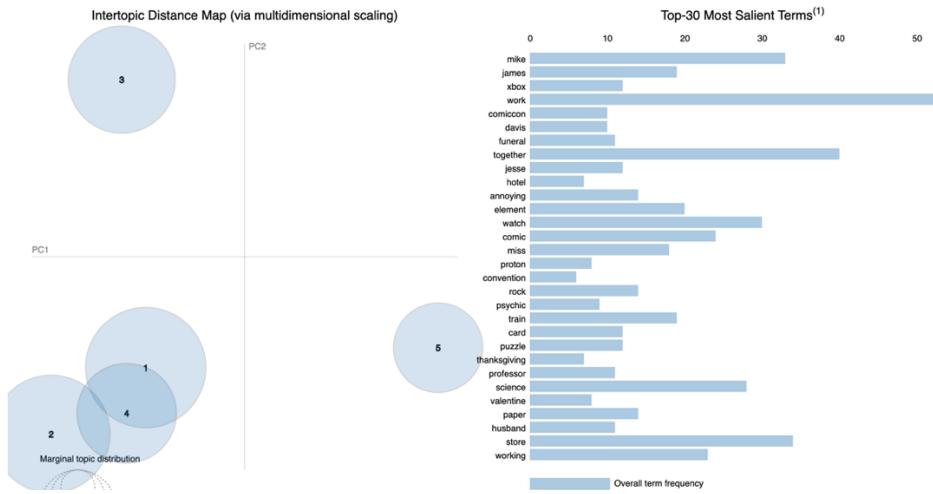
In questa visualizzazione il primo e il terzo topic si sovrappongono: il primo topic sembra ruotare ancora attorno al tema delle relazioni introducendo anche la parola "wedding". Nel terzo invece sembra esserci un'attenzione particolare ai riferimenti dei fumetti con parole come "Spock", "space" e "bongo". Il secondo topic sembra focalizzarsi maggiormente sugli svaghi e le attività ricreative degli amici nell'appartamento, con parole come "game", "playing" e "book". Come nei topic precedenti anche il quarto sembra ruotare attorno a scene e momenti professionali, interazioni con amici e feste, suggerendo un mix di vita lavorativa e sociale. Nel quinto topic il modello non ha fornito parole abbastanza rilevanti.

STAGIONE 6



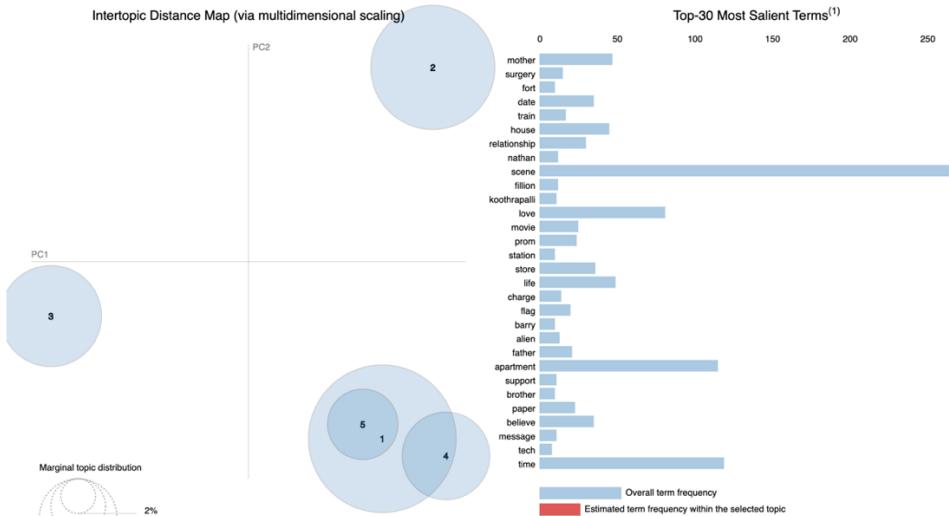
I topic sono ben distanziati tra loro, in particolare il primo risulta essere il topic più grande che ancora una volta sembra concentrarsi su scene nell'appartamento e sulle interazioni tra i personaggi, evidenziando eventi positivi e negativi denotati dalle parole "problem" e "help". Il secondo invece essere più definito, sembra fare riferimenti scientifici a personaggi come Stephen Hawking e Professor Proton, rilevando il tema scientifico della serie con la presenza delle parole "space", "stephen", "professor", "hawking" e "proton". Il terzo topic invece sembra coinvolgere maggiormente argomenti accademici e professionali, come la ricerca di lavoro "work", "office", "committee" o l'ottenimento di una promozione "tenure". Il quarto si concentra maggiormente sui fumetti, con riferimenti a personaggi come "Thor" e "martello", menzionando eventi come il "comic-con", evidenziando l'interesse dei protagonisti per l'immaginario. Il quinto topic ruota attorno a oggetti e situazioni riguardanti le interazioni sociali con "party", "dressed", "birthday" e "happy" menzionando il cibo con "dinner" e "food".

STAGIONE 7



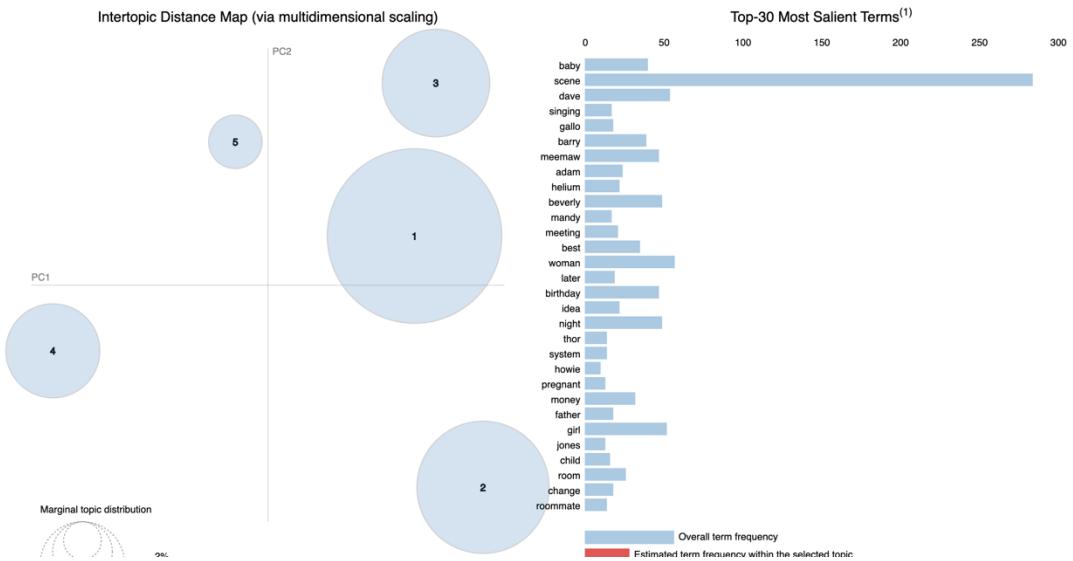
Qui il primo, il secondo e il quarto topic sembrano sovrapporsi, il primo topic continua ancora a trattare elementi che rimandano a scene di vita quotidiana e sulle interazioni tra amici come nelle stagioni precedenti. Il secondo, sembra concentrarsi sui giochi compaiono le parole “puzzle”, “comic” e “store”. Il terzo topic sembra coinvolgere le relazioni e elementi negativi come “annoying” e “funeral”, è in questa stagione infatti che avviene la morte del professor Proton. Anche il quarto topic si concentra sulla vita quotidiana e relazionale come nei topic precedenti. Il quinto topic ruota ancora attorno a eventi come “comic-con”, “convention” e “ticket”.

STAGIONE 8



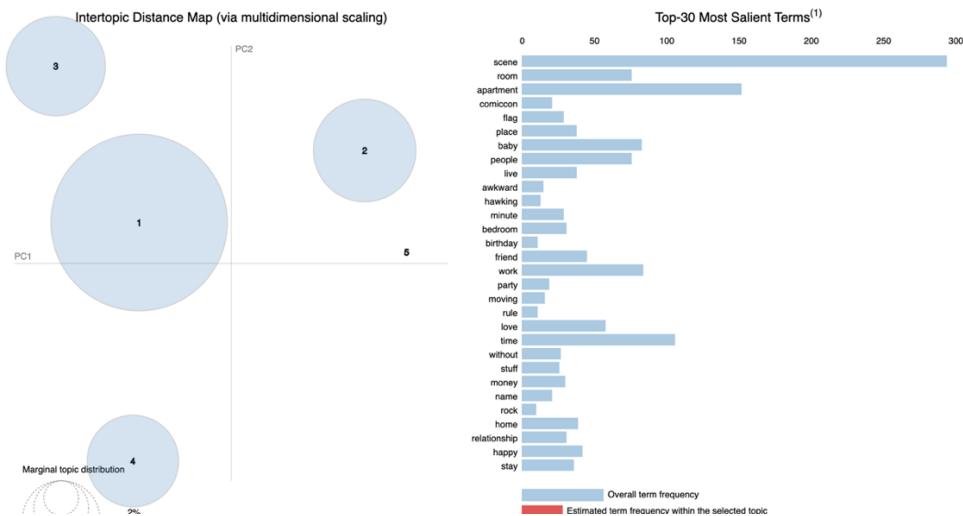
In questa visualizzazione il primo e il quinto topic sono sovrapposti: il primo topic sembra concentrarsi ancora sulle interazioni sociali, come appuntamenti, amicizia e relazioni, mentre il quinto invece sembra ruotare attorno a temi familiari e relazionali, con parole come “mother”, “father”, “brother” e “home”. Il secondo topic sembra concentrarsi su festività con parole come “christmas”, “year”, “santa” e “food”. Il terzo topic sembra far riferimento a situazioni legate alla scienza e al lavoro con le parole “scientist”, “theory”, “speech”, “money”, “tech”, “school” e “paper”. Nel quarto sono menzionate temi medici, come un intervento chirurgico “surgery”, “cause” e “worried”, questo topic probabilmente riguarda l’intervento chirurgico all’appendice di Raj.

STAGIONE 9



Anche nella nona stagione i topic sono ben distanziati, il primo sembra concentrarsi ancora su varie situazioni legate all'appartamento e alle relazioni interpersonali. Il secondo si focalizza sui fumetti, riportando personaggi immaginari come "spock", "jones" e "batman". Il terzo topic sembra coinvolgere ancora le interazioni sociali, ma soprattutto parole come "attorney", "contract" e "agreement" riguardano la routine di Sheldon nel definire dei contratti. Il quarto topic sembra concentrarsi ancora su temi legati alla quotidianità. Il quinto topic sembra ruotare attorno a cambiamenti nella vita dei personaggi, come l'arrivo di un bambino con le parole "baby", "pregnant", "father" e "child".

STAGIONE 10

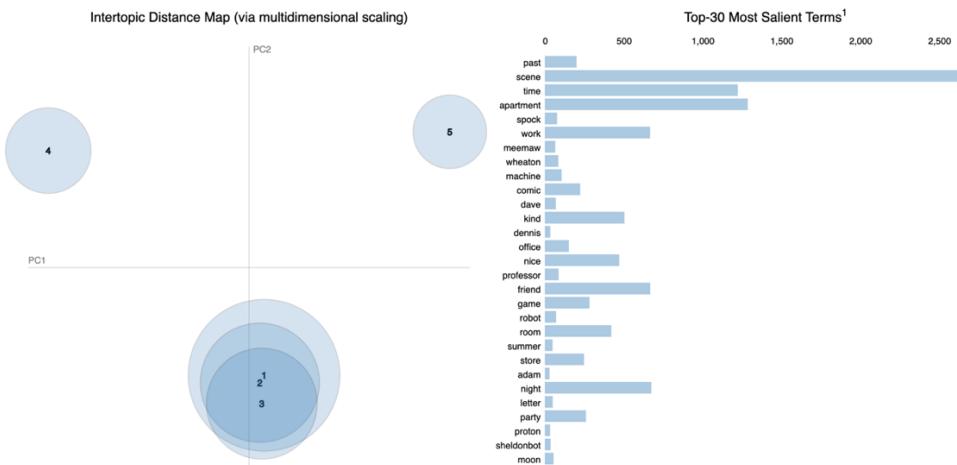


Anche nella decima stagione i topic sono ben distanziati tra loro. Il primo sembra concentrarsi, come il quinto della stagione precedente, sulle dinamiche relative all'arrivo di una nuova vita, rappresentate dai termini "baby", "together", "woman", "help", "mother" e "life". Il secondo e il terzo topic si focalizzano di nuovo su attività quotidiane, il quarto topic invece si concentra su traslochi e cambiamenti nella vita dei personaggi legati all'appartamento con riferimento a trasferimenti "moving", "living", "home", "stay", "place" e "room" e i relativi problemi "problem". Il quinto topic sembra non essere stato rilevato dal modello.

Analisi visualizzazione globale

Dopo aver analizzato i topic per ogni stagione, per concludere concentriamo l'attenzione sui cinque topic riassuntivi che vengono estratti per tutta la serie. Eseguendo il file 'global_da_visualization.html' presente nel corpus "topic_modeling_tbtt", viene aperta la visualizzazione interattiva. Si può subito notare come i primi

tre cerchi sono quasi sovrapposti, ciò significa che i topic corrispondenti hanno un'alta somiglianza nelle parole chiave, ossia i documenti associati a questi topic contengono termini correlati tra loro. Al contrario i topic 4 e 5 sono più distanti, suggerendo che i documenti associati a questi topic trattano argomenti diversi o hanno una rappresentazione lessicale distintiva. Si può notare come la grandezza dei cerchi decresce dal primo topic, che risulta essere quello che contiene parole più rilevanti, fino al quinto che risulta essere il topic che ne contiene parole meno rilevanti rispetto al corpus complessivo.



Si può notare in generale come i primi tre topic presentino termini che riguardano la vita quotidiana dei personaggi e l'ambientazione in cui si svolgono le vicende e le relazioni, vi è infatti pochissima differenza tra i temi di questi primi tre topic che, coerentemente con il genere della serie comica, riguardano la vita quotidiana, relazionale e l'amicizia. Gli ultimi due topic invece si focalizzano su riferimenti culturali presenti nella serie. Il quarto topic fa riferimento al tema dei fumetti e degli interessi dei protagonisti con parole come "macchina", "book", "comic" e "store". Il quinto topic invece fa riferimento a personaggi di Star Trek "Spock" e a Wil Weaton (attore presente nella serie) "weaton" e in generale parole legate all'intrattenimento come "game", "comic" e "store". Nella serie, infatti, i personaggi principali sono appassionati di fumetti e supereroi, spesso partecipano a convention e fanno riferimenti a questi personaggi nelle loro conversazioni quotidiane. Anche la presenza di "comic" e "store" vicine può far riferimento al personaggio Stuart Bloom, amico dei protagonisti e proprietario di un negozio di fumetti in cui i protagonisti comprano fumetti e giochi da tavolo.

Dalla visualizzazione globale dei topic si può notare come questa suggerisca che la serie "The Big Bang Theory" affronti una gamma diversificata di temi, in generale il tema della vita quotidiana è predominante in questa serie, insieme a quello delle relazioni e quello accademico-lavorativo e quello scientifico, fino a riferimenti culturali di intrattenimento quali fumetti e giochi di ruolo.

Tuttavia ultime stagioni, si nota una maggiore varietà di argomenti, con un'attenzione crescente alle relazioni familiari, matrimoni, gravidanze e cambiamenti nella vita dei personaggi principali. La serie sembra evolversi attraverso le sfide e i successi personali dei protagonisti, mantenendo comunque intatto l'elemento umoristico.

4.2.3 Analisi delle visualizzazioni: "The Vampire Diaries"

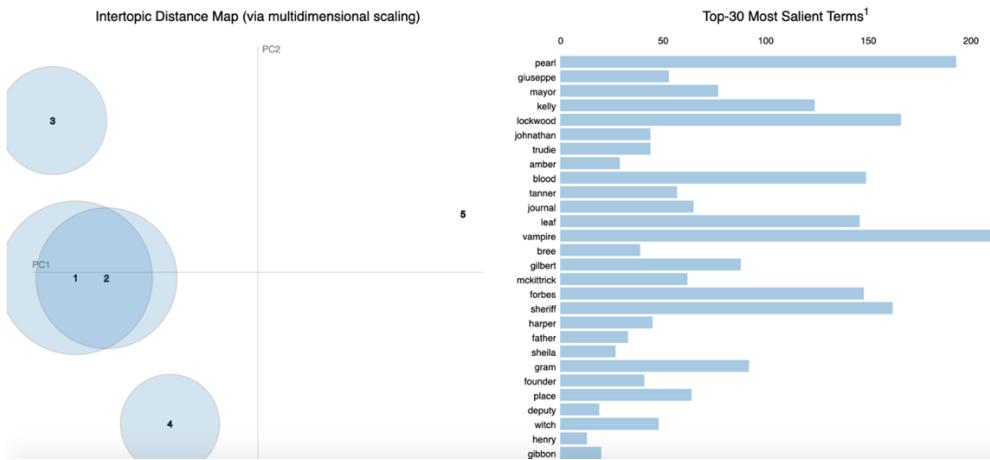
Utilizzando il codice contenuto all'interno del file "4_html_visualization.py", come detto precedentemente, vengono create, utilizzando il modello LDA della libreria "Gensim" per l'estrazione dei topics e la libreria "pyLDAvis", delle visualizzazioni html interattive, che permettono di esplorare approfonditamente i topics e le top words estratte per ognuno di essi.

L'analisi dei risultati ottenuti verrà condotta, come per "The Big Bang Theory" prima in modo specifico stagione per stagione, mentre successivamente a livello globale per individuare i temi principali e riassunti della serie, che permettano di ottenere una panoramica generale e globale della serie stessa.

Analisi dei risultati per stagione

All'interno della directory "topic_modeling_tvd" sono contenute delle sottodirectory relative a ciascuna stagione della serie. All'interno di ciascuna di esse è presente un file HTML che permette di visualizzare ed esplorare i topic risultanti per ogni stagione.

Verranno di seguito presentate, per ogni stagione delle anteprime di questi file.

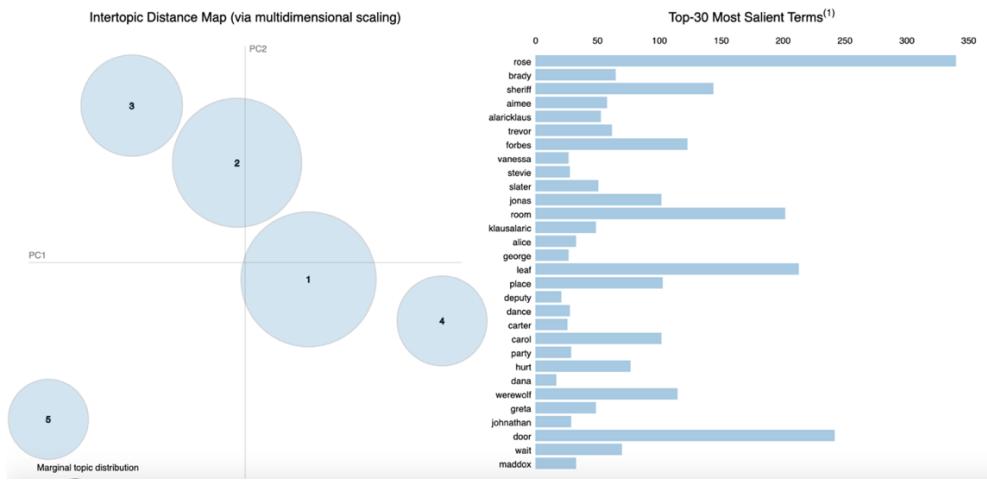


STAGIONE 1

Dall'analisi dello spazio bidimensionale in cui sono disposte le bolle, emerge chiaramente la vicinanza di molte di esse tra loro, suggerendo una notevole somiglianza nei temi trattati. Un'eccezione notevole è rappresentata dal quinto topic, che si distingue per la sua posizione più periferica rispetto agli altri. Le parole collocate lateralmente rivestono un ruolo di particolare rilevanza, in quanto rappresentano le parole chiave comuni a tutti i topics e offrono un riassunto delle principali tematiche nella stagione.

Il terzo e il primo topic, caratterizzato dalla bolla più grande e significativa, contengono le parole chiave più rappresentative del tema generale della serie, con termini come "vampire," "blood," "kill," e "hurt"; nel primo in particolare vengono inclusi nomi di luoghi e personaggi fondamentali della stagione, tra cui "house," "room," "sheriff," "forbes," e "lockwood." Il secondo topic, evidenzia una notevole intersezione con il primo nei temi trattati, contiene parole come "party," "school," e "family," che suggeriscono una maggiore focalizzazione su situazioni sociali e momenti di convivialità tra i personaggi. Infine, il quarto e il quinto topic, sebbene meno rilevanti in termini di dimensione, mettono maggiormente in evidenza tematiche relazionali, affettive e di vita quotidiana. Le parole chiave in questo comprendono "friend," "life," "kiss," e "school." In generale, nella prima stagione emergono i principali temi della serie: il sovrannaturale, la vita quotidiana e relazionale.

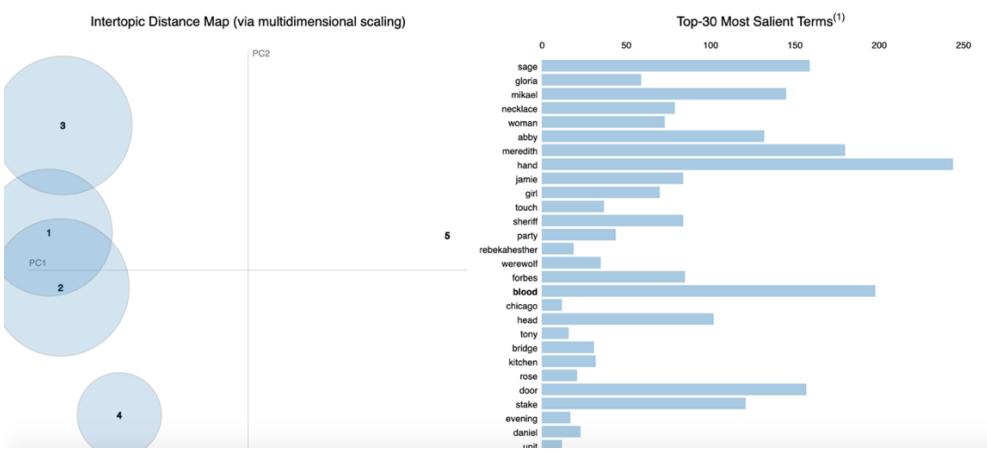
STAGIONE 2



Nella seconda stagione, notiamo un disegno delle bubbles più distribuito rispetto alla prima. Le parole chiave estratte per questa stagione si concentrano principalmente su nomi propri dei personaggi e introducono una nuova creatura fantastica, i licantropi.

Osservando il primo topic, emerge immediatamente il termine più rilevante, ossia "Rose", che è il nome di un personaggio di notevole importanza in questa stagione. In un conflitto tra vampiri e licantropi, Rose viene morsa da uno di questi ultimi e, per evitare ulteriori sofferenze, verrà successivamente uccisa da Damon, uno dei protagonisti che abbiamo analizzato in modo approfondito nel corso di questo progetto. Inoltre, il topic presenta ulteriori riferimenti alla suddetta lotta, con parole come "place", "blood", "kill", "dead", e "moonstone". Il secondo topic presenta molte affinità con il primo, introducendo inoltre le streghe come figure fantastiche e includendo parole come "love", che fanno riferimento alle relazioni romantiche. Anche il terzo topic tratta tematiche simili ai primi due ma individua, oltre a Rose, altri personaggi di rilievo, come lo sceriffo, "Alaric" e "Klaus". Il quarto topic, invece, si concentra principalmente sui luoghi in cui si svolgono gli episodi della stagione, con parole come "room", "house", "door", e "floor", anche se presenta numerosi termini in comune con i topic precedenti. Infine, il quinto topic presenta termini simili al primo e al secondo topic.

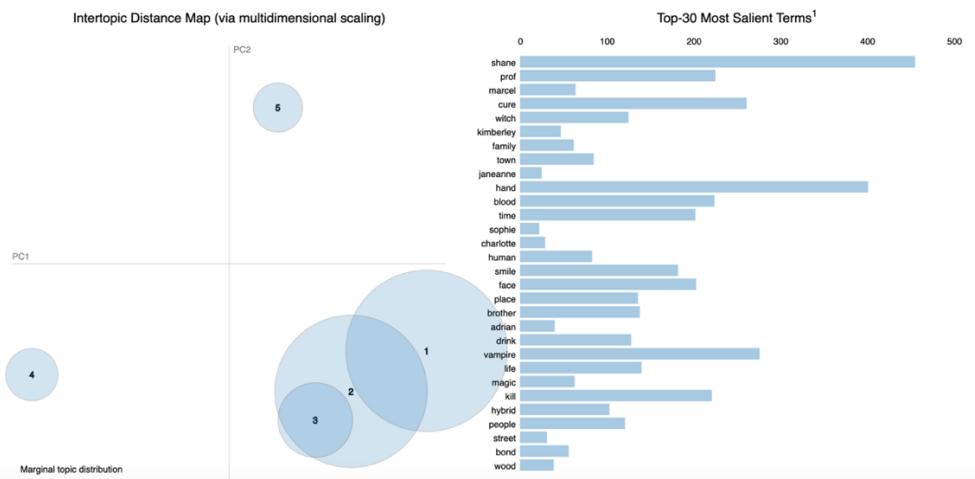
STAGIONE 3



Analizzando la terza stagione, notiamo una disposizione delle bubbles molto ravvicinata, con l'eccezione del quinto topic, che appare notevolmente più piccolo e quindi di minore rilevanza. Nel primo topic rilevano ancora le parole "vampire," "kill," "life," e "blood" legate alle vicende dei vampiri e alle loro sfide. Nel secondo topic invece compaiono anche termini legati agli scontri e alle battaglie, oltre alle streghe e agli ibridi. Il terzo topic oltre al tema sui vampiri, si concentra anche su temi legati alla famiglia e alle relazioni sentimentali. Nel quarto, il termine "gloria" assume rilevanza, e notiamo anche la presenza della parola "necklace," che

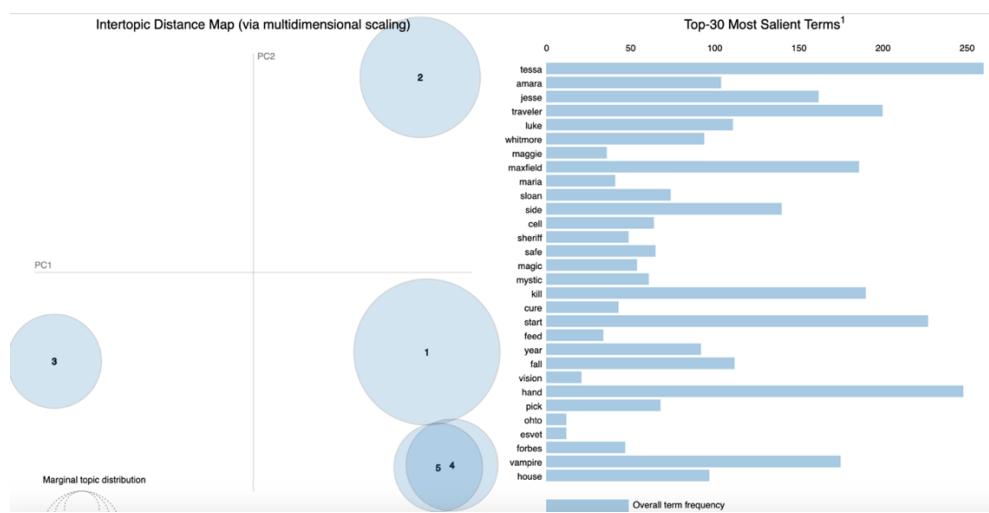
sicuramente fa riferimento a un oggetto di particolare importanza all'interno della stagione. Il quinto topic appare molto più piccolo e meno rilevante rispetto agli altri.

STAGIONE 4



Analizzando le bubbles dei topic della quarta stagione, osserviamo che i primi tre topic appaiono molto vicini, se non addirittura sovrapposti, mentre i topic quattro e cinque sono più piccoli e distanti. Nel primo topic emergono anche altri temi importanti, come la ricerca della cura per il vampirismo e la presenza degli ibridi, con le parole "cure". Il secondo comprende ancora una volta termini che rimandano alle dinamiche familiari e relazionali. Il terzo topic è più piccolo e quasi completamente sovrapposto al secondo. Il quarto topic, anch'esso di minore rilevanza fa maggior riferimento al tema delle streghe con la parola "witch." Il quinto e ultimo topic riprende il tema della ricerca della cura del vampirismo, ma è importante anche il termine "human," poiché il tema della perdita dell'umanità a causa del vampirismo è fondamentale. In questo topic, continuano a emergere le tematiche legate ai vampiri, al magico e alle relazioni interpersonali.

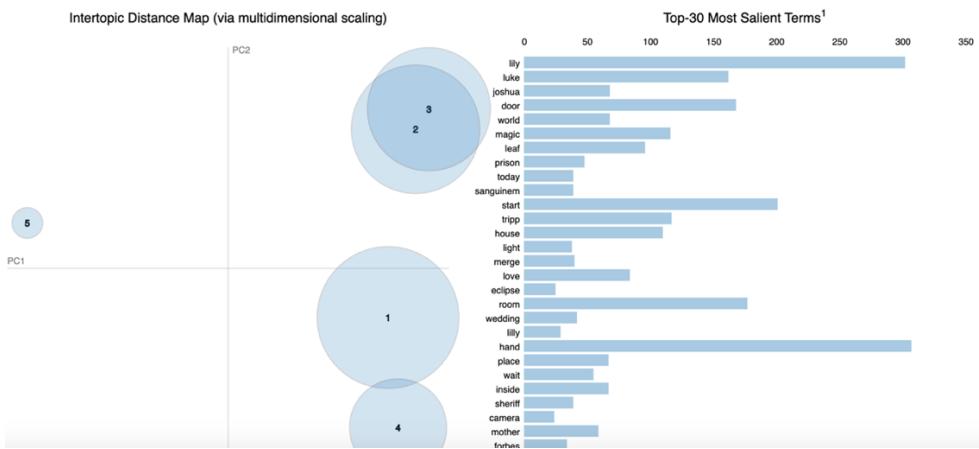
STAGIONE 5



Nel primo topic emergono ancora una volta le parole chiave "blood," "vampire," "kill," e "dead," relativo al tema fondamentale della serie. Il secondo topic, al contrario, esplora un tema diverso trattato in questa stagione, ovvero l'"ancora". La parola chiave principale in questo topic è "Tessa," una strega che ha creato la cura per il vampirismo e l'"Altra parte," una dimensione in cui le anime delle creature magiche finiscono quando muoiono. Il terzo topic affronta il tema dei "viaggiatori," un'altra tipologia di personaggi rilevanti nella serie. In questo topic, emergono ancora temi legati al mondo magico, alle streghe, al sangue e alle uccisioni, tutti collegati alle lotte e alle battaglie che caratterizzano la serie. Le bubbles relative al quarto e al quinto topic

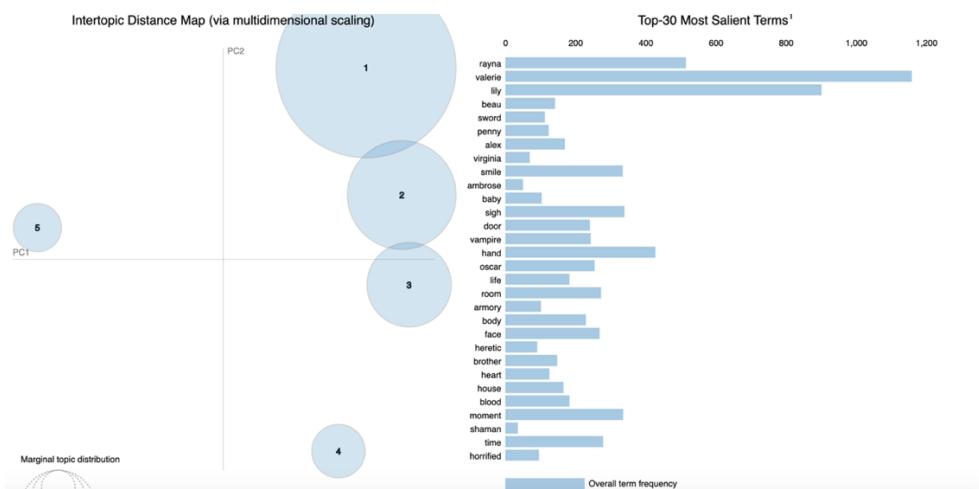
sembrano sovrapposte quasi completamente. Qui ritroviamo ancora temi legati alle relazioni, con la presenza della parola "love."

STAGIONE 6



Nel primo topic, il termine chiave principale è associato a un nuovo personaggio, "Luke." Luke è uno stregone che in questa stagione si unisce a Elena per affrontare la minaccia dei viaggiatori. Questo topic si concentra principalmente su questa tematica specifica. Nel secondo topic emergono anche termini legati allo stato confusionale dei personaggi, al sangue, alla famiglia e all'amore. Il terzo topic appare quasi completamente sovrapposto al secondo, evidenziando la similitudine delle tematiche trattate. Tuttavia, compare anche la parola "humanity," che come nelle stagioni precedenti, riveste un ruolo significativo. Il quarto topic tratta principalmente gli scontri sanguinosi tra i vari personaggi, enfatizzando la tensione e gli eventi drammatici. Infine, l'ultimo topic, sebbene meno rilevante rispetto agli altri, affronta principalmente tematiche cupo, come suggerito dalla presenza della parola "funeral." Inoltre, tocca il tema dell'eclissi e dell'imprigionamento, poiché in questa stagione Damon si trova imprigionato in una dimensione.

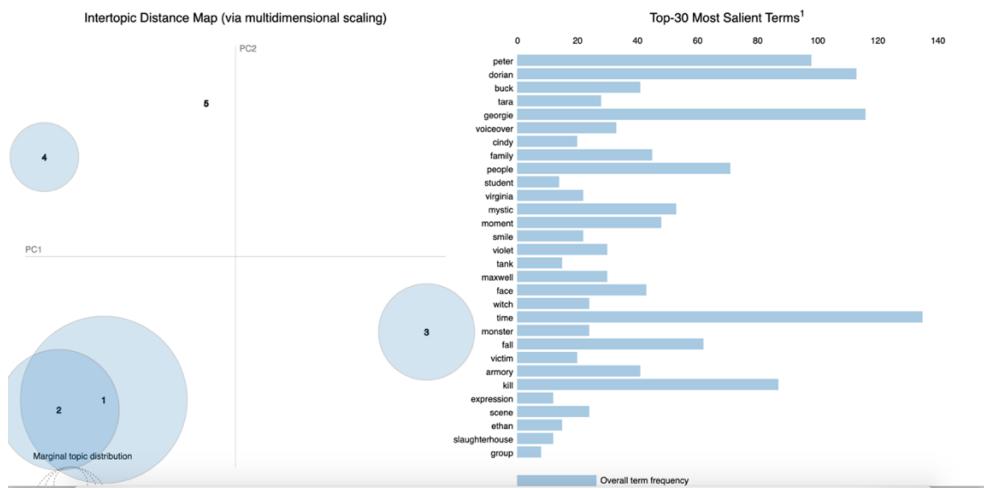
STAGIONE 7



Nel primo topic, così come nel quarto e nel quinto, alcune parole suggeriscono una particolare attenzione alle emozioni e alle reazioni dei personaggi. Termini come "smile," "moment," "sigh," e "confused" indicano che la trama potrebbe esplorare i sentimenti e le risposte emotive dei personaggi in varie situazioni. L'ambientazione fisica è un elemento importante, con menzioni di "room," "house," "town," e "Mystic," che forniscono un contesto fisico per la storia. Continuano ad apparire termini come "blood," "stone," e "dead" che rimandano ad elementi. Nel secondo topic, termini rilevanti richiamano il tema della cripta dell'Organizzazione Armory. Nel terzo topic termini come "sigh," "smile," "frown," e "voice" indicano una riflessione sulle emozioni e le

espressioni dei personaggi, suggerendo una certa profondità nei loro sentimenti e nei loro dialoghi. La presenza di "start," "time," "year," e "finally" suggerisce anche un elemento di cronologia o il passare del tempo all'interno della narrazione.

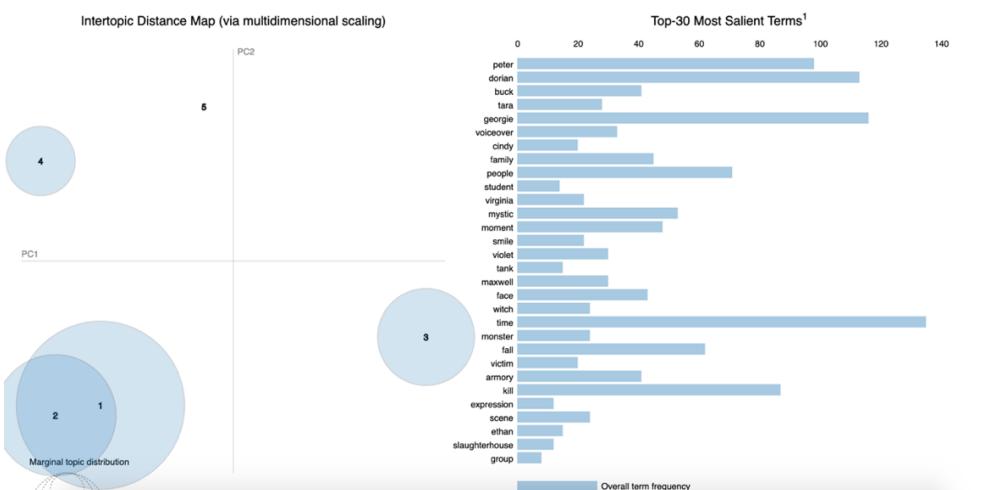
STAGIONE 8



Nell'ottava stagione, possiamo osservare che il primo topic assume un'importanza notevolmente maggiore rispetto agli altri, mentre il quinto topic è molto più piccolo e contiene quasi interamente il secondo topic. Analizzando i termini estratti per questa stagione finale, emergono nuovi personaggi, mentre altri termini rimangono strettamente legati al mondo del magico e delle creature mostruose. Il primo topic è particolarmente significativo e contiene termini legati alle emozioni e alle relazioni affettive, con parole come "love," "life," "mind," e "better." Questo suggerisce una profonda esplorazione delle emozioni dei personaggi e delle loro vite. Allo stesso tempo, emergono parole che evocano situazioni di pericolo e dramma, come "hell," "kill," "dead," "leave," e "fall." Il secondo topic sembra principalmente incentrato sul tema dell'umanità, che in questa stagione assume un'importanza estrema. Il terzo topic è principalmente concentrato sul tema della magia, mentre il quarto rileva principalmente nomi di personaggi già citati precedentemente, consolidando la continuità delle relazioni e delle storie. Il quinto topic, come accennato, appare piuttosto piccolo e meno rilevante rispetto agli altri.

Analisi dei risultati globali

Dopo aver analizzato i topic per ogni stagione, per concludere concentriamo l'attenzione sui cinque topic riassuntivi che vengono estratti per tutta la serie.



Dopo un'analisi approfondita dei temi presenti in ogni stagione, emerge chiaramente che i termini selezionati per l'intera serie risultano altamente informativi, rappresentativi e coerenti. Il primo tema predominante è senza dubbio quello legato al mondo dei vampiri, che costituisce l'essenza stessa della serie e costituisce il fulcro di tutti gli eventi che si susseguono nel corso della narrazione.

Il secondo tema evidenzia una serie di parole chiave correlate agli avvenimenti cruciali che si sviluppano nel corso della trama, come "hell" (inferno), "kill" (uccidere) e "blood" (sangue). Parallelamente, termini come "sigh", "smile", "confused" e "love" suggeriscono una forte enfasi sulle emozioni e sulle relazioni affettive tra i personaggi. Il terzo tema ruota principalmente attorno al mondo magico e alle creature straordinarie presenti (streghe e viaggiatori). In questa tematica emerge anche un aspetto significativo riguardante le dinamiche familiari e le relazioni affettive. Nel quarto tema emerge il concetto di cura per il vampirismo, un elemento ricorrente e di rilevanza nell'arco della narrazione. Infine, nel quinto tema, le emozioni e le relazioni interpersonali continuano a giocare un ruolo centrale, come evidenziato dai termini "smile", "confused" e "love". Inoltre, la parola "mother" (madre) suggerisce che le dinamiche familiari svolgono un ruolo significativo nella trama. Riguardo all'ambientazione fisica, i termini come "room" (stanza), "door" (porta) e "house" (casa) indicano che il luogo in cui si svolge la storia potrebbe avere un'influenza rilevante sulla trama. La presenza di parole come "stone" (pietra) e "blood" (sangue) suggerisce l'eventuale presenza di elementi misteriosi o sovrannaturali all'interno della storia.

In conclusione, l'analisi dei temi emersi in ciascuna stagione della serie ha rivelato un quadro affascinante e ricco di sfaccettature. I termini selezionati per ciascun tema sono risultati essere accurati, informativi e coerenti con l'evoluzione della trama. Il tema centrale legato al mondo dei vampiri ha costantemente permeato l'intera serie, come anche le dinamiche familiari e le relazioni affettive. Nel complesso, l'analisi dei cinque temi principali ha contribuito a sottolineare la complessità e la profondità della narrazione, restituendo effettivamente le tematiche più caratteristiche della serie.

PUNTO 5: CONFRONTO TOPICS TRA LE SERIE IN ANALISI

5.1: Analisi con WordCloud

Dopo aver esaminato i topic estratti dai vari modelli utilizzati, ed avendo esaminato approfonditamente uno ad uno con le visualizzazioni HTML, è possibile ora effettuare un confronto degli stessi, alla luce dei risultati emersi, per verificare ed esaminare le differenze e le similitudini di tematiche trattate. Nel corso dell'analisi delle due serie, è stato sicuramente possibile notare numerose differenze tra le stesse. Le differenze dipendono dalla natura intrinseca delle due serie, una cupa e misteriosa, che ha l'obiettivo di tenere in tensione lo spettatore e l'altra che intrattiene lo spettatore in modo leggero.

Infatti, "The Vampire Diaries" come è stato confermato, è una serie che ruota principalmente attorno al paranormale e al vampirismo. I suoi temi principali, come abbiamo potuto constatare precedentemente, includono il conflitto tra il bene e il male, le relazioni, il vampirismo e l'umanità, esplorando anche temi legati all'amicizia, alla famiglia e alla magia. Dall'altro lato, "The Big Bang Theory" è una commedia incentrata sulla vita quotidiana di un gruppo di scienziati. I temi principali includono l'amicizia, la crescita personale e le relazioni. Nella serie compare spesso il tema dei fumetti e della cultura nerd con riferimenti a film, videogiochi, giochi di ruolo e scienza. Ovviamente al contrario di "The Vampire Diaries" non contiene elementi sovrannaturali, non essendo una serie fantastica. "The Big Bang Theory" presenta quindi come temi generali le sfide della vita quotidiana, le dinamiche di gruppo e le aspirazioni personali dei personaggi.

Sebbene queste due serie appartengano a generi molto diversi, abbiamo potuto notare, nel corso dell'analisi stagionale dei topic, che ci sono alcune somiglianze nei loro temi, di fatto entrambe affrontano le dinamiche delle relazioni. Sia "The Vampire Diaries" che "The Big Bang Theory" esplorano il tema dell'amicizia, con i personaggi principali che si sostengono a vicenda attraverso le sfide della vita. "The Vampire Diaries" si concentra su relazioni romantiche complicate, mentre "The Big Bang Theory" esplora le relazioni di coppia in modo più leggero. Inoltre, entrambe le serie includono momenti di crescita personale per i loro personaggi. In "The Vampire Diaries," i personaggi devono affrontare le loro nature sovrannaturali e cercare di mantenere la loro umanità. In "The Big Bang Theory," i personaggi crescono nel corso della serie, affrontando sfide nella loro vita personale e professionale. Perciò, sebbene "The Vampire Diaries" e "The Big Bang Theory" appartengano a generi diametralmente diversi con trame molto diverse, condividono temi comuni legati alle relazioni umane, all'amicizia e alla crescita personale.

Per rendere ancora più chiare le distinzioni e le similitudini tra le due serie vengono di seguito riportate le parole chiave estratte dal metodo NMF per ciascun tema, affiancando i topic delle due serie attraverso l'utilizzo di Wordcloud. Questo permetterà di effettuare un confronto più chiaro ed evidente tra le parole chiave estratte a livello globale di "The Vampire Diaries" e "The Big Bang Theory". Inoltre, visivamente, la dimensione delle parole, riflettendo l'importanza delle stesse all'interno dei documenti utilizzati per effettuare l'analisi, permette di comprendere quelle che hanno una maggiore rilevanza per ogni tematica estratta dal modello.

Per fare ciò viene utilizzata la funzione "create_combined_wordclouds(topics_tvt,topics_tbtt)" contenuta nel file "4_topic_modeling.py".





Dalle Wordclouds e dalle parole chiave estratte dai topic delle due serie, emergono evidenti differenze.

In "The Big Bang Theory," alcune parole chiave nei topic risultano essere particolarmente rilevanti. Tra queste spiccano "astronaut", "hawking", "robot", "dungeon" e "cheesecake", elementi strettamente connessi al tema scientifico, ai fumetti e ai giochi di ruolo e in generale agli aspetti della vita quotidiana dei personaggi. Mentre in "The Vampire Diaries" al contrario, notiamo parole chiave quali "doppelganger", "hybrid", "cure," "traveler" e "moonstone" tutte legate al mondo soprannaturale, agli antagonisti e agli oggetti magici fondamentali per la trama della serie. Benché questi cinque temi estratti per ogni serie non permettano effettivamente di evidenziare tutti i temi, l'analisi dettagliata dei topic per ciascuna serie ha permesso invece di individuare temi simili, anche se sviluppati in maniera unica. Tra questi, rientrano gli argomenti precedentemente esaminati, legati alla vita quotidiana, alla convivenza domestica e alle relazioni, i quali riflettono una sfera tematica comune nelle due serie, per quanto in modo diverso.

CONCLUSIONE

In questo progetto, abbiamo esplorato approfonditamente le serie televisive "The Vampire Diaries" e "The Big Bang Theory" attraverso l'analisi del contenuto dei dialoghi, la valutazione del sentimento e la modellazione dei temi. Il nostro obiettivo era comprendere meglio le emozioni e i temi dominanti all'interno di queste serie, concentrandoci su specifici personaggi chiave in ciascuna.

Abbiamo iniziato raccogliendo dati testuali da episodi di entrambe le serie mediante scraping dei siti web, consentendoci di creare un corpus di testi dettagliato. Successivamente, abbiamo calcolato il sentimento dei dialoghi utilizzando metodologie di analisi testuale, ottenendo una panoramica delle emozioni predominanti in ogni episodio e in seguito scendendo nei dettagli per ciascun personaggio.

Attraverso l'approccio di Topic Modeling, tramite i modelli LDA (Latent Dirichlet Allocation) e NMF (Non-Negative Matrix Factorization), abbiamo identificato e interpretato i temi chiave presenti nei dialoghi. Questa analisi ci ha permesso di scoprire quali argomenti dominanti e concetti ricorrevano più frequentemente all'interno delle serie.

Le differenze nel sentimento e nei temi dominanti tra "The Vampire Diaries" e "The Big Bang Theory" rispecchiano la natura distintiva di ciascuna serie. "The Vampire Diaries" è caratterizzata da un'atmosfera più intensa e drammatica, ha confermato la presenza di un sentimento maggiormente negativo. Al contrario, "The Big Bang Theory" offre un tono più leggero e comico confermando un sentimento nettamente positivo. Tutto ciò si è riflesso sull'analisi dei personaggi chiave nelle due serie, focalizzazione che ci ha consentito di comprendere le sfumature emotive e tematiche legate a ognuno di essi. L'analisi del sentimento e la Topic Modeling hanno evidenziato le differenze intrinseche tra queste serie, fornendo un'ulteriore comprensione dei mondi narrativi distintivi creati da "The Vampire Diaries" e "The Big Bang Theory".