Giacomo De Gobbi: 860913
Davide Vercesi: 852483

ANALISI DELLA SERIE ANIMATA SOUTH PARK

ABSTRACT

In questo progetto si è analizzato attraverso gli strumenti del Text Mining come è evoluta la serie animata South Park, dagli esordi ai giorni nostri. Nello specifico sono state utilizzate come oggetto dello studio la prima (1997) e l'ultima stagione (2015) verificando se nel corso degli anni ci sia stata una evoluzione della serie. Tale evoluzione sembrerebbe necessaria data la pungente satira e lo sprezzante black-humor che hanno sempre contraddistinto South Park, che poco si sposa con il crescente politically-correct e cancel-culture degli ultimi anni. In particolare si è analizzata l'evoluzione del linguaggio e contenuti attraverso la sentiment analysis e della struttura attraverso l'utilizzo della BoW,Tf-Idf e Topic Modeling.

SOUTH PARK

South Park è una serie televisiva animata statunitense, creata da Matt Stone e Trey Parker nel 1997 per Comedy Central.

South Park, attraverso la satira, tratta temi di politica e attualità statunitensi e cerca di sfatare i tabù e le demonizzazioni della società, spesso usando la parodia e la satira. Negli Stati Uniti la serie è stata aspramente criticata da gruppi religiosi che la giudicano moralmente offensiva e anti-statunitense a causa delle parolacce presenti nel cartone.

La satira di South Park prende di mira molti aspetti della cultura statunitense e dell'attualità, e sfida molti dei tabù e delle convinzioni radicate nella società, di solito attraverso la parodia e il grottesco. L'abuso di minori in ogni sua forma è tema ricorrente nella serie animata, insieme all'omosessualità e al tema della morte. La serie è inoltre famosa per la trattazione di temi d'attualità in maniera pressoché estemporanea, come gli attentati dell'11 settembre 2001. Tale fatto è dovuto alla sua tecnica di realizzazione (animazione computerizzata), che permette la realizzazione di un episodio in soli quattro giorni, il 75% in meno rispetto ad una puntata dei Simpson.

I temi, i popoli, le religioni e le idee che il cartone tocca e che irride sono molteplici. I personaggi non statunitensi sono spesso stereotipati: i messicani, ad esempio, vengono rappresentati con delle difficoltà a capire l'inglese e con scarsa voglia di lavorare fino a dormire sul posto di lavoro. Oltre ai messicani sono generalmente vittime della satira del cartone animato gli afro-americani, i canadesi, i cinesi, gli italiani, gli ebrei, gli appartenenti a Scientology, i mormoni e i musulmani. Dalla diciannovesima stagione sono stati introdotti i "politicamente corretti", raffigurati come dei ragazzi palestrati e prepotenti che vanno su tutte le furie quando leggono o sentono dire qualcosa di politicamente scorretto.

DATASET

I dati utilizzati sono stati scaricati da GitHub[1], i file .csv contengono informazioni sullo script, tra cui: stagione, episodi e personaggi. I dati sono stati originariamente scaricati prendendo le trascrizioni in html dal sito ufficiale di South Park.

Il dataset è stato diviso formando una lista dove ogni elemento è una singola battuta. All'interno della lista si sono create altre liste per identificare meglio le informazioni. Si è ricavato per ogni elemento il numero della stagione, il numero dell'episodio, il numero della battuta, il nome del personaggio che la pronuncia e la trascrizione della battuta. Successivamente è stato svolto il pre-processing.

```
[500,
 '\r\n1',
 '2',
 'Cartman',
 'yeah i only weigh pounds',
 ['yeah', 'i', 'only', 'weigh', 'pounds'],
 ['yeah', 'weigh', 'pounds']]
```

PREPROCESSING

1. Tokenizzazione: Questo passaggio è necessario per suddividere ogni battuta di ciascun personaggio, che viene archiviata come stringa, in più token costituiti da singole parole e caratteri di punteggiatura. Questa operazione è stata implementata usando la funzione *word_tokenize* all'interno della libreria **nltk**.

2. Minuscole e maiuscole: trasformare in minuscola ogni lettera delle parole analizzate, altrimenti un computer tratterà due parole identiche come diverse solo perché presentano una stessa lettera nelle due versioni, minuscola e maiuscola;

3. Slang e abbreviazioni: nelle puntate analizzate sono presenti svariati slang e abbreviazioni. Un sottoinsieme di essi è stato sostituito con il loro equivalente comune (ad es. "you" anziché "ya");

4. Forma contratta: le parole che presentano una forma contratta (ad es. "won't") vengono espanse nel loro modulo "standard" (ad es. "will not").

5. Caratteri speciali e punteggiatura: questi caratteri vengono rimossi, poiché non sono utili a per il nostro studio. Questa operazione aiuta anche a ridurre la dimensionalità del dataset, portando a un aumento dell'efficienza. I caratteri che potrebbero influire sulla tokenizzazione non vengono eliminati in questo passaggio.

6. Lemmatizzazione e POS: al fine di rimuovere il suffisso dalle parole e ricondurle alla loro forma base, si effettua la lemmatizzazione: in questo modo le forme singolari e plurali della stessa parola saranno sostituite con lo stesso termine. Le implementazioni degli algoritmi utilizzati in questa fase sono della libreria **nltk** di Python. Per rendere più efficace la lemmatizzazione si è effettuato part of speech tagging (POS), per evitare che la libreria commetta errori sulla categoria lessicale e trasformi parole che non ne necessitano (es. "was" deve rimanere "was" e non diventare "wa").

[1] https://github.com/BobAdamsEE/SouthParkData/tree/master/by-season

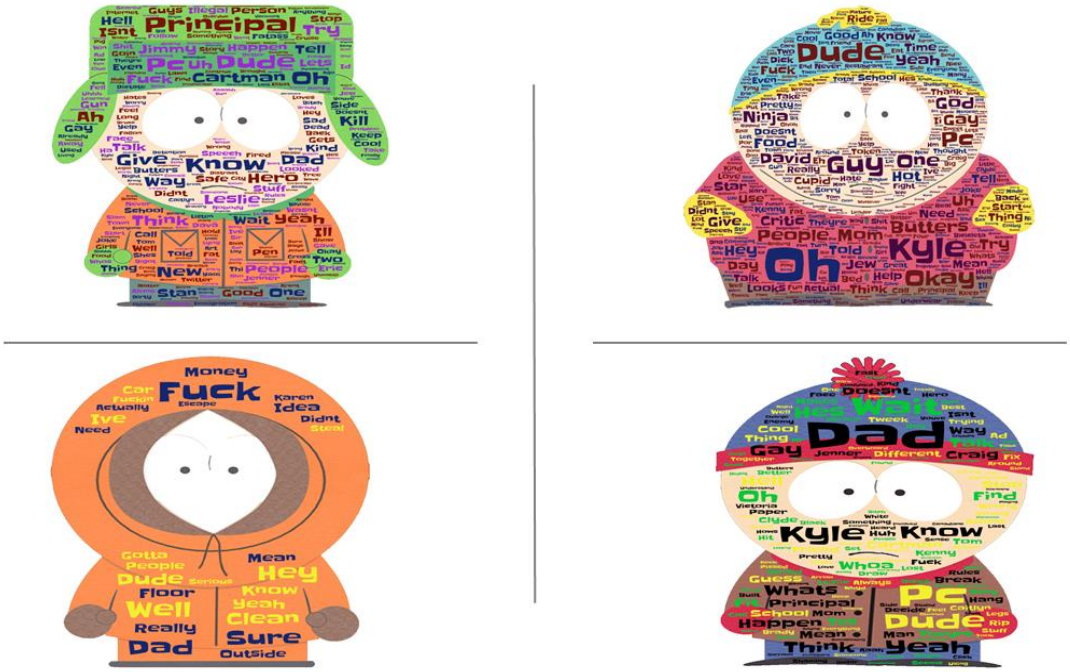
7. Stop words: il passo successivo ha comportato la rimozione delle stop-words. Il vantaggio di questo compito è di ridurre ampiamente la dimensionalità del dataset, pur mantenendo intatta la sua informatività: parole come articoli e pronomi sono davvero frequenti nelle frasi e la loro rimozione contribuisce notevolmente a ridurre la dimensionalità del dataset senza alterarne il significato.

WORD-CLOUD

Nel corso delle stagioni compaiono molteplici personaggi sia fittizi sia caricature di persone realmente esistenti. Nonostante ciò, i protagonisti principali sono quattro ragazzini che si è deciso di descrivere brevemente e di utilizzare come sagome per rappresentare le word-cloud per le loro rispettive battute nelle due stagioni in questione.

- Stanley "Stan" Marsh: È apparentemente il più normale dei quattro e viene presentato come onesto, sensibile e bene intenzionato.
- Eric Theodore “Cartman”: Si caratterizza come il personaggio più cattivo, antagonista e antieroe della serie. È il personaggio che rappresenta l’anima dissacrante e satirica del programma.
- Kyle Broflovski: insieme con Stanley Marsh è il leader del gruppo dei quattro protagonisti. È il più intelligente dei quattro protagonisti: in più occasioni viene fatto presente come ottenga degli ottimi risultati a scuola.
- Kenneth "Kenny" McCormick: appare quasi sempre con addosso un giubbotto arancione col pelo con un cappuccio, che nasconde il suo viso e rende incomprensibili le sue battute. La caratteristica principale di Kenny è che proferisce pochissime battute nel corso della serie ed è spesso vittima di scherno e bullismo da parte degli altri tre protagonisti.

(stagione 1)



(stagione 19)

STATISTICHE DESCRITTIVE

Per prima cosa si è deciso di mostrare le distribuzioni riguardanti le battute e il numero di parole contenute in esse per i primi 10 personaggi della serie nelle due stagioni in esame.

Stagione 1:

- Numero di personaggi per stagione: 240
- Numero di battute in tutta la stagione: 4169

character	Somma parole/battuta	Media parole/battuta	Conta battute	Pct di battute su totale
Stan	4949	7.498485	660	15.8
Cartman	6075	9.720000	625	15.0
Kyle	4179	7.422735	563	13.5
Chef	2643	10.529880	251	6.0
Mr. Garrison	2902	13.688679	212	5.1
Wendy	1046	8.171875	128	3.1
Jimbo	1413	11.487805	123	3.0
Jesus	592	9.548387	62	1.5
Liane	561	9.048387	62	1.5
Kenny	374	6.800000	55	1.3

Stagione 19:

- Numero di personaggi per stagione: 230
- Numero di battute in tutta la stagione: 2259

character	Somma parole/battuta	Media parole/battuta	Conta battute	Pct di battute su totale
Cartman	3783	15.315789	247	10.9
Randy	2821	14.392857	196	8.7
PC Principal	3023	21.439716	141	6.2
Kyle	1450	10.820896	134	5.9
Butters	972	8.452174	115	5.1
Stan	758	7.895833	96	4.2
Jimmy	1225	14.411765	85	3.8
Mr. Garrison	1245	15.370370	81	3.6
Leslie	565	14.487179	39	1.7
Gerald	412	11.444444	36	1.6

Si può osservare che sia nella prima che nell'ultima stagione presa in analisi sono presenti più o meno lo stesso numero di personaggi, però il numero di battute presenti nella prima è nettamente superiore rispetto alla diciannovesima.

Nella prima stagione tra i 10 personaggi con più battute solo un personaggio oltre ai protagonisti compare anche tra i personaggi principali della diciannovesima stagione.

Altro dato molto interessante è relativo al numero medio di parole presenti per ogni battuta. Si nota che nella prima stagione le battute sono numericamente superiori con personaggi che fanno discorsi brevi, mentre nella stagione diciannove le battute sono numericamente inferiori però i personaggi fanno discorsi molto più lunghi. Emblematico il caso di Cartman, uno dei personaggi più scurrili e volgari, che passa da 9,7 a 15,31 parole medie pronunciate per battuta.

TEXT REPRESENTATION

La rappresentazione del testo è uno dei problemi fondamentali nel mondo del Text Mining e dell'Information Retrieval (IR). Si mira a rappresentare numericamente i documenti di testo non strutturati per renderli adatti al lavoro degli algoritmi. Le rappresentazioni scelte sono:

- BoW (Bag of Word): il metodo della Bag-of-Words è utilizzato nell'Information Retrieval e nell'elaborazione del linguaggio naturale per rappresentare documenti ignorando l'ordine delle parole e la loro grammatica. In questo modello, ogni documento è considerato in quanto contiene parole, analogamente a una borsa; ciò consente una gestione di queste basata su liste, dove ogni borsa contiene determinate parole di una lista.
- TF-IDF (term frequency–inverse document frequency): questo tipo di rappresentazione tiene conto dell'informazione legata alla frequenza con cui compaiono i termini, questa frequenza è normalizzata rispetto alla lunghezza variabile delle battute dei personaggi e permette di dare meno peso alle parole più comuni nel corpus, le quali non hanno un buon potere di discriminazione. In particolare, la funzione utilizzata per il calcolo della matrice permette di eliminare le parole troppo ricorrenti e troppo rare, ovvero quelle che si trovano ai lati della curva di Zipf's, fissando il cut-off inferiore e superiore.

La BoW è stata utilizzata per poter scoprire le parole più importanti e caratterizzanti per i personaggi principali della serie. Inoltre, è stata la base fondamentale per poter costruire il modello LDA.

Nel metodo TF-IDF si è considerato un documento come l'insieme delle battute pronunciate da un personaggio, per esempio le parole utilizzate da Cartman sono un singolo documento, le parole utilizzate da Randy sono un altro documento, ragion per cui ci saranno tanti documenti quanti il numero dei personaggi. Nello specifico il TF-IDF è stato invece utilizzato per trovare le parole con maggior frequenza nelle due stagioni in rapporto ai personaggi principali.

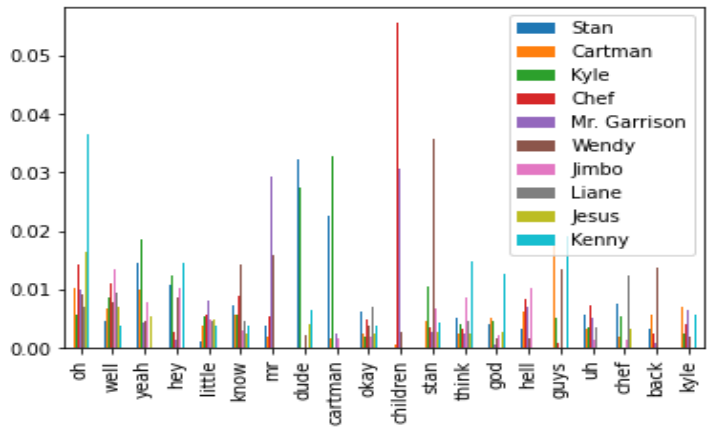
Stagione 1

	Stan	Cartman	Kyle	Chef	Mr. Garrison	Wendy	Jimbo	Liane	Jesus
0	(dude, 69)	(hey, 58)	(cartman, 56)	(children, 53)	(mr, 35)	(stan, 24)	(ned, 24)	(hon, 12)	(son, 8)
1	(cartman, 48)	(guys, 48)	(yeah, 48)	(oh, 27)	(hat, 34)	(know, 11)	(well, 14)	(eric, 8)	(oh, 7)
2	(yeah, 47)	(oh, 45)	(dude, 47)	(well, 21)	(children, 32)	(hi, 10)	(us, 11)	(sure, 5)	(caller, 6)
3	(oh, 34)	(mom, 39)	(mr, 30)	(know, 17)	(okay, 23)	(eww, 7)	(think, 9)	(cheesy, 4)	(satan, 5)
4	(hey, 30)	(yeah, 38)	(hey, 28)	(hello, 13)	(oh, 21)	(back, 7)	(lets, 8)	(poofs, 4)	(way, 4)
5	(know, 27)	(ass, 34)	(stan, 27)	(hell, 12)	(well, 17)	(oh, 7)	(hey, 8)	(well, 4)	(thou, 4)
6	(kenny, 24)	(well, 29)	(well, 26)	(little, 11)	(little, 17)	(mr, 7)	(hell, 8)	(man, 4)	(one, 4)
7	(okay, 23)	(know, 25)	(kenny, 19)	(damn, 9)	(eric, 15)	(ms, 7)	(shoot, 7)	(chef, 4)	(yea, 3)
8	(sparky, 23)	(man, 24)	(hankey, 19)	(uh, 9)	(know, 12)	(ellen, 7)	(boys, 7)	(hello, 3)	(children, 3)
9	(grampa, 23)	(eh, 22)	(know, 17)	(okay, 9)	(wait, 12)	(guys, 6)	(coming, 7)	(okay, 3)	(air, 3)

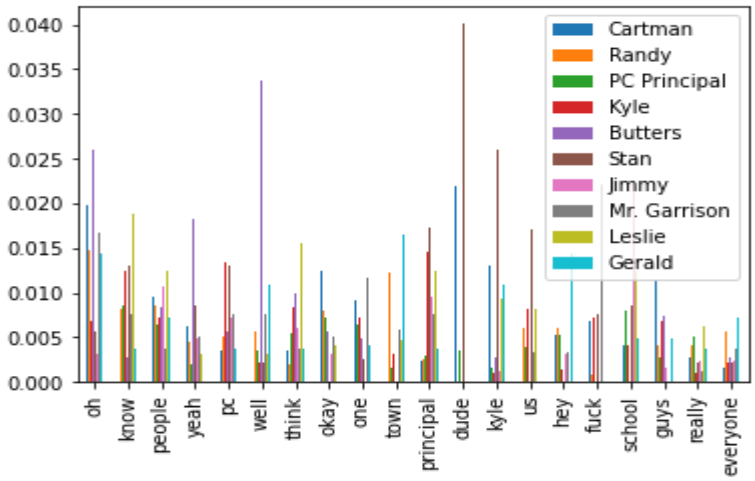
Stagione 19

	Cartman	Randy	PC Principal	Kyle	Butters	Stan	Jimmy	Mr. Garrison	Leslie
0	(oh, 38)	(oh, 22)	(bro, 23)	(principal, 14)	(well, 24)	(kyle, 12)	(ads, 24)	(oh, 10)	(jimmy, 7)
1	(kyle, 33)	(people, 17)	(alright, 19)	(dude, 14)	(eric, 14)	(dad, 11)	(ad, 20)	(fuck, 10)	(know, 6)
2	(dude, 24)	(know, 16)	(pc, 18)	(pc, 13)	(oh, 14)	(wait, 10)	(news, 18)	(one, 8)	(think, 5)
3	(okay, 24)	(town, 16)	(know, 17)	(know, 12)	(yeah, 13)	(dude, 8)	(school, 14)	(country, 8)	(trying, 5)
4	(people, 24)	(gay, 13)	(people, 13)	(think, 8)	(man, 12)	(principal, 8)	(know, 11)	(hell, 7)	(help, 5)
5	(butters, 23)	(okay, 12)	(school, 12)	(cartman, 7)	(think, 7)	(pc, 6)	(people, 9)	(death, 7)	(principal, 4)
6	(guys, 22)	(kids, 11)	(think, 11)	(people, 7)	(people, 6)	(know, 6)	(leslie, 9)	(wall, 6)	(people, 4)
7	(one, 20)	(everyone, 11)	(one, 11)	(jimmy, 7)	(uh, 6)	(us, 6)	(principal, 8)	(youve, 6)	(school, 3)
8	(david, 18)	(foods, 11)	(okay, 11)	(yeah, 6)	(really, 6)	(whats, 5)	(way, 7)	(know, 6)	(whats, 3)
9	(yeah, 16)	(well, 11)	(two, 10)	(hero, 6)	(guys, 4)	(yeah, 4)	(pc, 6)	(well, 6)	(kyle, 3)

Stagione 1



Stagione 19



.

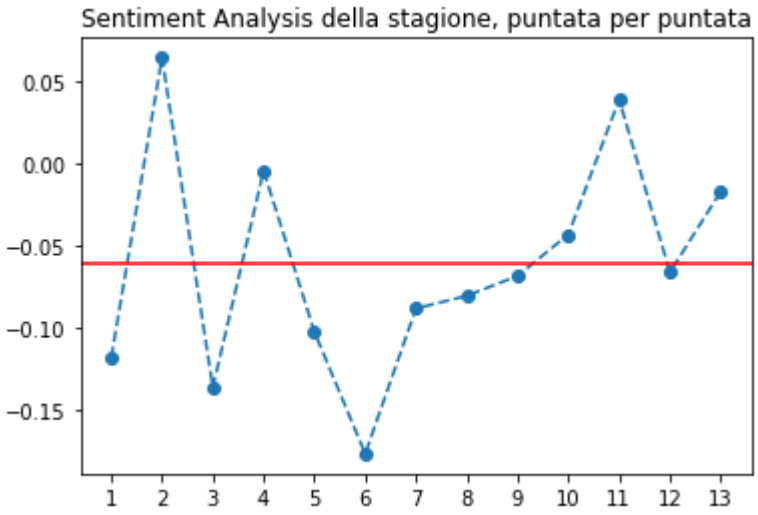
SENTIMENT ANALYSIS

La Sentiment-Analysis, è un'analisi procedurale di calcolo dei sentimenti e delle opinioni espresse nei testi. Questo tipo di analisi ti consente di comprendere la natura delle interazioni svolte nel testo, in un preciso contesto e in un determinato arco temporale.

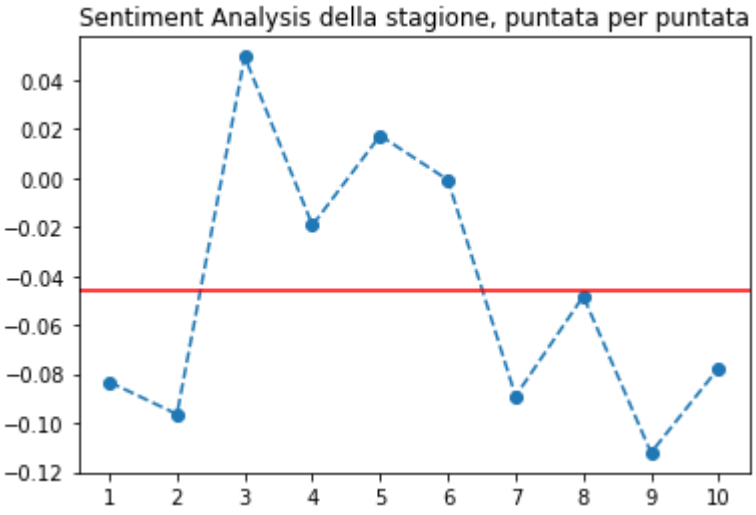
La Sentiment Analysis ai fini del progetto è stata calcolata per mostrare se il linguaggio ha subito delle variazioni nell'arco temporale delle stagioni prese in analisi. Il lessico scelto è stato **AFINN**, il più semplice ma anche il più popolare, contenente più di 3300 parole con un punteggio di polarità associato ad ogni parola. Questo lessico è basato su "unigrammi" (o parole singole) e assegna alle parole un punteggio che va da -5 a 5, con punteggi negativi che indicano sentimenti negativi e punteggi positivi che indicano sentimenti positivi.

La Sentiment è stata calcolata per ogni episodio delle stagioni analizzate. I risultati ottenuti sono i seguenti:

Stagione 1



Stagione 19



Come si può vedere dai grafici i valori oscillano da 0,06 a -0,17, in entrambe le stagioni la media è quindi attorno allo zero. Possiamo perciò interpretare tale risultato come rappresentativo di una certa continuità linguistica e contenutistica delle due stagioni. Probabilmente però tali risultati sono alquanto bias, dato che la serie cita personaggi, luoghi ed eventi legati alle vicende americane in maniera spesso ironica e satirica e perciò molto probabile che molte espressioni non siano state ben interpretate da Afinn e perciò considerate come parole neutra.

TOPIC MODELING

Il Topic Modeling è una tecnica di machine learning unsupervised utilizzata per fare clustering di gruppi di parole e espressioni simili che meglio caratterizzano un set di documenti. In particolare, la Latent Dirichlet Allocation è un modello di analisi del linguaggio naturale che categorizza i documenti, trattati come BoW, per topic mediante generative probabilistic models assumendo che siano prodotti da un insieme di topics distribuiti al loro interno seguendo la distribuzione di Dirichlet.

L'input dell'LDA è un corpus di documenti abbinati a un numero di k topics che si vuole estrarre. L'output è il set di documenti espressi come combinazione dei k topic, grazie all'algoritmo che cerca i weights delle connessioni tra doc e topics e tra topics e words. Una volta fornito il numero ottimale di topic che si vuole ottenere, tutto ciò che fa LDA è riorganizzare la distribuzione degli argomenti all'interno dei documenti e la distribuzione delle parole chiave all'interno degli argomenti per ottenere una buona composizione della distribuzione argomento-parola chiave.

Si è utilizzato LDA dal pacchetto **Gensim**. Per generare un modello di LDA, è necessario capire la frequenza con cui ogni termine si ripete all'interno di ogni documento. Per far ciò è stata costruita una matrice dei termini del documento ("document-term matrix") con il pacchetto **Gensim**. Viene infatti creato un ID univoco per ogni parola all'interno del documento. Il corpus prodotto è una mappa di "word-id" e "word-frequency".

Sono state utilizzate due metriche di valutazione intrinseca:

- Perplessità: misura statistica di quanto bene un modello probabilistico predice un sample, essendo una misura di incertezza sarà necessario che sia il più bassa possibile;
- Coerenza: misura lo score di un singolo topic calcolando il grado di semantic similarity tra parole con alto score, al contrario della precedente metrica deve essere più alta possibile affinchè le performance siano migliori; Queste misurazioni aiutano a distinguere tra argomenti che sono interpretabili semanticamente e argomenti che sono artefatti di inferenza statistica.

Stagione 1: Perplexity= -7.51 Coherence= 0.52

Stagione 19: Perplexity= -7.38 Coherence= 0.42

Oltre a queste due misure è stato adottato un approccio knowledge-based. Infatti, il numero dei topic per ciascun modello LDA è stato settato uguale al numero di episodi di ciascuna stagione. Si è fatto ciò per poter capire se ogni episodio presenta una trama a sé stante o invece gli episodi sono collegati e si sviluppa una trama o temi comuni.

Per la prima stagione, composta da 13 episodi, sono infatti stati utilizzati 13 topics e tutti e 13 i topics corrispondono univocamente ad un episodio specifco, denotando una struttura con episodi che non creano una trama unica e che possono essere visti senza ordine. Per esempio:

'0.031"costum" + 0.019*"candi" + 0.018*"zombi" + 0.014*"wendi" + ' '0.014*"eye" + 0.009*"aaah" + 0.009*"pink" + 0.009*"dress" + '0.009*"trickortreat" + 0.008*"sauc"'*

Tale topic perfettamente riconducibile all'episodio 7 (*Pinkeye*) che ha come tema un'invasione zombie durante la festività di Hallowen.

'0.040*"triangl" + 0.023*"streisand" + 0.022*"barbra" + 0.014*"smith" +' '0.014*"robert" + 0.009*"barbara" + 0.008*"zinthar" +' '0.007*"leonard" + 0.007*"maltin"'

In questo caso si tratta dell'episodio 12 (*Mecha-Streisand*), dove i quattro protagonisti sono coinvolti nello scontro tra l'attrice Barbara Streisand e il cantante dei Cure Robert Smith, entrambi alla ricerca di un artefatto magico di Zinthar.

Per quanto riguarda la stagione 19 si è riusciti ad identificare univocamente alcuni dei 10 episodi dai topics restituiti da LDA, ma la maggior parte di tali topics presentavano parole chiave o espressioni riconducibili ad episodi diversi e mischiati tra loro. La spiegazione si può capire leggendo l'intervista dei produttori della serie in cui affermano che proprio la stagione 19 è stata un punto di svolta per la serie dato che si è cercato di collegare gli episodi e di creare sottotrame complesse che si sviluppano in più episodi.[2]

CONCLUSIONI

Il tempo sembrerebbe aver apportato modifiche alla serie. Il numero dei personaggi è diminuito ma al contempo il numero di battute e il numero di parole per battuta dei personaggi presenti nella stagione 19 è incrementato notevolmente. Anche la struttura della stagione è evoluta. Come spiegato precedentemente nel topic modeling, la prima stagione era composta da episodi con trame a sé stante tutti differenziati e senza continuità, mentre la diciannovesima ha episodi collegati e sottotrame che vengono sviluppate nel corso della stagione. Il linguaggio dai risultati della sentiment non sembra essere cambiato, anche se probabilmente ciò è dovuto all'uso di strumenti non particolarmente adatti allo studio di una serie dal linguaggio pieno di neologismi e di riferimenti alla cultura pop statunitense.

[2] https://www.cinemablend.com/television/1524470/why-south-park-changed-so-much-in-season-19