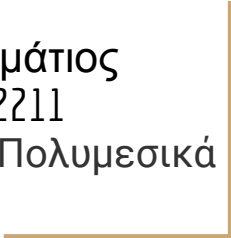


# Music Video Clips Retrieval based on similarity

Όνομ/νυμο: Ορφανός Σταμάτιος  
Αριθμός Μητρώου: mtn2211  
Μάθημα: Μηχανική Μάθηση σε Πολυμεσικά  
Δεδομένα



# Περιεχόμενα

1. Εισαγωγή
2. Δεδομένα και Προεπεξεργασία
3. Image Similarity Analysis
4. Text-Based Similarity Analysis
5. Fusion and Integration of Similarity Metrics
6. Demo
7. Συμπεράσματα

## Εισαγωγή

- Η μουσική και τα μουσικά βίντεο έχουν γίνει αναπόσπαστο μέρος της ζωής μας, με δισεκατομμύρια βίντεο να παρακολουθούνται και να κοινοποιούνται στο διαδίκτυο καθημερινά.
- Ωστόσο, με τόσο μεγάλο όγκο διαθέσιμου περιεχομένου, η εύρεση συγκεκριμένων μουσικών βίντεο ή η ανακάλυψη παρόμοιων μπορεί να είναι μια πρόκληση.

## Εισαγωγή

- Η εργασία μας στοχεύει να αντιμετωπίσει αυτήν την πρόκληση αναπτύσσοντας ένα σύστημα που αξιοποιεί μετρήσεις ομοιότητας
  - Εικόνας,
  - Κειμένου

για την ανάκτηση μουσικών βίντεο με βάση τα οπτικά, ακουστικά και στιχουργικά χαρακτηριστικά τους.

## Δεδομένα

Για την εργασία μας χρησιμοποιήσαμε τα 90 δημοφιλέστερα τραγούδια από την πλατφόρμα Spotify.

Φυσικά για κάθε ένα από τα τραγούδια χρειαζόμαστε και το Video Clip, έτσι δημιουργήσαμε ένα script το οποίο να αποθηκεύει την πληροφορία σε μορφή MP4.

# Δεδομένα και Προεπεξεργασία

## Εικόνα:

Όπως γνωρίζουμε ένα video είναι μια αλληλουχία από frames, τα οποία παρουσιάζονται με ένα συγκεκριμένο frame rate, συνήθως από 24 μέχρι 60 frames per second (fps).

Για να εξάγουμε εικόνες από ένα video χρησιμοποιούμε ένα Frame Sampling.

Frame Sampling είναι η διαδικασία επιλογής ενός frame ανά κάποιο χρονικό διάστημα.

## Δεδομένα και Προεπεξεργασία

### Εικόνα:

Στο date-set μας είχαμε video με διαφορετική διάρκεια και αποφασίσαμε να εφαρμόσουμε διαφορετικό Sampling Rate ανάλογα με τη διάρκεια του video.

Έτσι έχουμε τα εξής:

1. Διάρκεια μεγαλύτερη από **300sec** smampling rate 0.5 samples/sec
2. Διάρκεια **200sec-300sec** smampling rate 1 samples/sec
3. Διάρκεια μικρότερη από **200sec** smampling rate 2 samples/sec

## Δεδομένα και Προεπεξεργασία

### Εικόνα:

Για την εξαγωγή εικόνων από το video clip χρησιμοποιήσαμε τη βιβλιοθήκη [ffmpeg-python](#) με αποτέλεσμα εικόνες διαφορετικής ανάλυσης ανάλογα με το video resolution.



# Δεδομένα και Προεπεξεργασία

## Κείμενο:

Ένα πολύ σημαντικό χαρακτηριστικό ενός Music Video Clip είναι οι στίχοι του τραγουδιού, οι οποίοι προσφέρουν υψηλού επιπέδου πληροφορία.

Η διαδικασία εξαγωγής στίχων από ένα τραγούδι είναι πιο περίπλοκη διαδικασία σε σχέση με την εικόνα, καθώς απαιτεί τη χρήση πιο εξελιγμένων speech-to-text μοντέλων.

# Δεδομένα και Προεπεξεργασία

## Κείμενο:

Για κάθε ένα από τα music video clips χρησιμοποιήσαμε το μοντέλο [OpenAI-Whisper English Base model](#) για να εξάγουμε τους στίχους.

Η διαδικασία αυτή παρουσίασε ένα σύνολο από προβλήματα:

1. Ορισμένα τραγούδια είχαν έντονο voice-tuning
2. Ορισμένα τραγούδια είχαν πολλαπλές γλώσσες
3. Ορισμένα τραγούδια είχαν λίγους και επαναλαμβανόμενους στίχους

# Δεδομένα και Προεπεξεργασία

## Κείμενο:

Για να λύσουμε τα παραπάνω προβλήματα χρησιμοποιήσαμε το English version έτσι ώστε να αποφύγουμε τα παρπάνω προβλήματα και να επιστρέψουμε όσο το δυνατόν πιο ακριβή αποτελέσματα.

# Δεδομένα και Προεπεξεργασία

## Κείμενο:

Τα αποτελέσματα της προεπεξεργασίας κειμένου είναι:

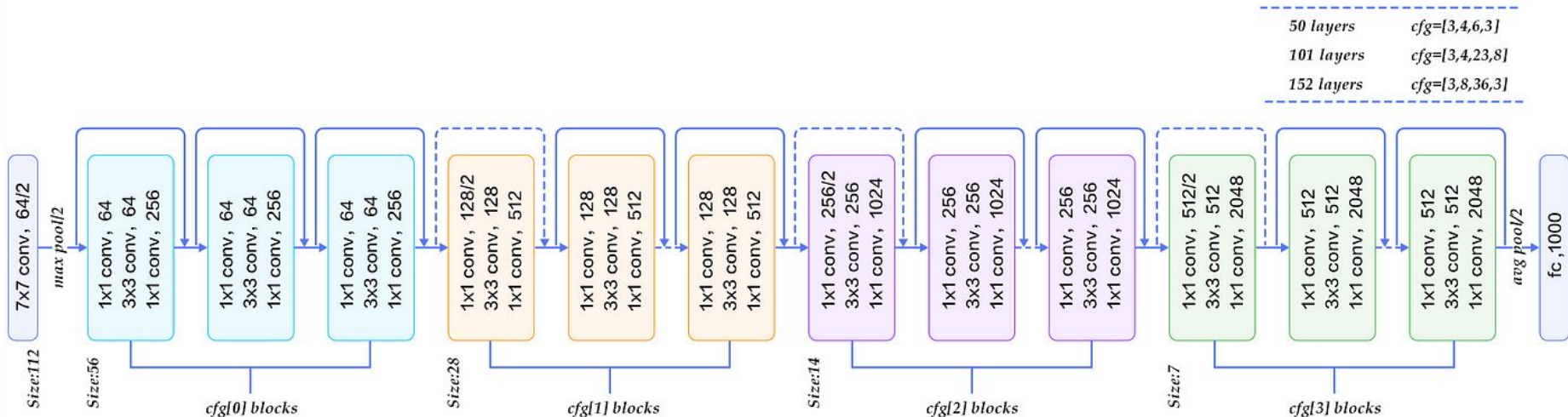
1. Ένα text file με τους στίχους
2. Ένα csv file με ακριβής λεπτομέρειες του τραγουδιού ανά segment όπως:
  - a. Αρχή του segment του τραγουδιού
  - b. Τέλος του segment του τραγουδιού
  - c. Πιθανότητα ομιλία
  - d. Κείμενο
  - e. Tokens από το κείμενο του segment

```
lyrics > 0-Seconds-of-Summer-Youngblood.txt
1 Speaking English
2 We have cake.
3 This is lastream.
4 prices are pretty much Confederate in 2020
5 tactics were
6 You made me believe your mind
7 Yeah, you used to call me baby now
8 You're calling me my name
9 Takes one and one, yeah, you beat me at my own damn game
10 You pushing, you pushing, I'm pulling away, pulling away from you
11 I give it, I give it, I give it, you take it, you take it
12 Young blood, so you want me, so you want me
13 The body is up and I'm just a dead man
14 Walking tonight, but you need it, yeah, you need it
15 All in this time, yeah, you're in this time, yeah
16 Young blood, so you want me, so you want me
17 Back in your life, so I'm just a dead man
18 Gonna call it tonight, cause I need it, yeah, I need it
19 All in this time, yeah, you need it
20 Yeah, conversations and like it's the last goodbye
21 That one of us gets to jump in cause about a hundred times
22 So who even call it baby, nobody could save my place
23 When you looking at those strangers hope to guide you sick of my face
24 Young blood, so you want me, so you want me
25 The body is up and I'm just a dead man
26 Walking tonight, but you need it, yeah, you need it
27 All in this time, yeah
28 Young blood, so you want me, so you want me
29 Back in your life, so I'm just a dead man
30 Gonna call it tonight, cause I need it, yeah, you need it
31 All in this time, yeah
```

```
lyrics > 0-Seconds-of-Summer-Youngblood.csv
1 id,ones,stars,end,tex,tokens,temperature_avg,logprob,compression_ratio,nc_speech_prob
2 0,0,0,0,1,76, Speaking English,"[50803, 21393, 3598, 58751]",1.0,-5.567355531634786,1.285687476635514,0.22467490874386488
3 1,1,0,0,0,6,12,58, We have cake,"[50766, 775, 423, 12187, 13, 50892]",1.0,-5.567355531634786,1.285687476635514,0.22467490874386488
4 2,2,0,10,58,15,58, This is lastream,"[51842, 770, 318, 938, 1476, 13, 51142]",1.0,-5.567355531634786,1.285687476635514,0.22467490874386488
5 3,3,0,10,30,19,46, prices are pretty much Confederate in 2020,"[51813, 4536, 389, 2455, 853, 10386, 515, 287, 12131, 51383]",1.0,-5.567355531634786,1.285687476635514,0.22467490874386488
6 4,4,0,27,72,20,26, tactics were,"[51749, 10815, 547, 28542, 51262]",1.0,-5.567355531634786,1.285687476635514,0.22467490874386488
7 5,5,3000,38,0,33,0, You made me believe your mind,"[50863, 921, 925, 582, 1975, 534, 2080, 50513]",0.0,-0.345385781795816,1.7597482597482598,0.7116345
8 6,6,3000,38,0,38,0, Yeah, you used to call me baby now,"[50613, 9425, 11, 345, 973, 284, 869, 582, 5156, 783, 50763]",0.0,-0.345385781795816,1.7597482597482598,0.7116345
9 7,7,3000,38,0,41,0, You're calling me my name,"[50763, 921, 621, 4585, 582, 616, 1438, 50913]",0.0,-0.345385781795816,1.7597482597482598,0.7116345
10 8,8,3000,41,0,49,0, Takes one and one, yeah, you beat me at my own damn game,"[51813, 33887, 530, 290, 530, 11, 10194, 11, 345, 4485, 582, 379, 616, 8,
11 9,9,3000,51,0,55,0, You pushing, you pushing, I'm pulling away, pulling away from you,"[51413, 921, 7796, 11, 345, 7796, 11, 314, 1181, 18427, 1497, 1,
12 10,10,3000,55,0,59,0, I give it, I give it, I give it, you take it, you take it,"[51513, 314, 1577, 340, 11, 314, 1577, 340, 11, 314, 1577, 340, 11, 3
13 11,11,5900,59,0,62,0, Young blood, so you want me,"[51803, 8969, 2918, 11, 523, 345, 765, 582, 11, 523, 345, 765, 582, 51813]",0.0,-0.
14 12,12,5900,62,0,64,0, The body is up and I'm just a dead man,"[50513, 383, 1767, 318, 510, 290, 314, 1181, 655, 257, 9386, 582, 90713]",0.0,-0.286564882
15 13,13,5900,66,0,78,0, Walking tonight, but you need it, yeah, you need it,"[50713, 21276, 9975, 11, 475, 345, 761, 340, 11, 10194, 11, 345, 761, 340,
16 14,14,5900,78,0,75,0, All in this time, yeah, you're in this time, yeah,"[50913, 1439, 287, 428, 640, 11, 10194, 11, 345, 821, 287, 428, 640, 11, 1019
17 15,15,5900,75,0,78,0, Young blood, so you want me, so you want me,"[51803, 8969, 2918, 11, 523, 345, 765, 582, 11, 523, 345, 765, 582, 51813]",0.0,-0.
18 16,16,5900,78,0,82,0, Back in your life, so I'm just a dead man,"[51813, 5157, 207, 534, 1284, 11, 523, 314, 1181, 655, 257, 9386, 582, 51813]",0.0,-0.
19 17,17,5900,82,0,85,0, Gonna call it tonight, cause I need it, yeah, I need it,"[51513, 482, 6415, 869, 349, 9975, 11, 2728, 314, 761, 340, 11, 10194,
20 18,18,8600,85,0,91,0, All in this time, yeah, you need it,"[50863, 1439, 287, 428, 640, 11, 10194, 11, 345, 761, 340, 50613]",0.0,-0.252707558704723,
21 19,19,8600,91,0,97,0, Yeah, conversations and like it's the last goodbye,"[50613, 9425, 11, 40275, 296, 580, 340, 338, 262, 936, 24829, 50613]",0.0,-0.
22 20,20,8600,99,0,100,0, That one of us gets to jump in cause about a hundred times,"[51813, 1220, 530, 290, 530, 11, 345, 3811, 284, 4391, 217, 2728, 546, 257, 34
23 21,21,8600,107,0,113,0, So who even call it baby, nobody could save my place,"[51413, 1486, 508, 772, 869, 340, 5156, 11, 8168, 714, 3613, 616, 1295,
24 22,22,11300,114,0,121,0, When you looking at those strangers hope to guide you sick of my face,"[50413, 1649, 345, 2845, 379, 883, 18981, 2911, 284, 5698
25 23,23,11300,122,0,126,0, Young blood, so you want me, so you want me,"[50613, 8969, 2918, 11, 523, 345, 765, 582, 11, 523, 345, 765, 582, 51813]",0.0,-0.
26 24,24,11300,126,0,130,0, The body is up and I'm just a dead man,"[51813, 383, 1767, 318, 510, 290, 314, 1181, 655, 257, 9386, 582, 51813]",0.0,-0.102182
27 25,25,11300,130,0,134,0, Walking tonight, but you need it, yeah, you need it,"[51513, 21276, 9975, 11, 475, 345, 761, 340, 11, 10194, 11, 345, 761, 340,
28 26,26,11300,134,0,136,0, All in this time, yeah,"[51413, 1439, 287, 428, 640, 11, 10194, 51513]",0.0,-0.1021629877322527,1.7371794871794872,0.83401898
29 27,27,11300,136,0,142,0, Young blood, so you want me, so you want me,"[51613, 8969, 2918, 11, 523, 345, 765, 582, 11, 523, 345, 765, 582, 51813]",0.0,-0.
30 28,28,14200,142,0,146,0, Back in your life, so I'm just a dead man,"[50403, 5157, 207, 534, 1284, 11, 523, 314, 1181, 655, 257, 9386, 582, 90403]",0.0
31 29,29,14200,146,0,158,0, Gonna call it tonight, cause I need it, yeah, you need it,"[50513, 482, 6415, 869, 340, 9975, 11, 2728, 314, 761, 340, 11, 101
```

# Image Similarity Analysis

Ξεκινώντας με τα δεδομένα εικόνων αποφασίσαμε να χρησιμοποιήσουμε ένα Representation από ένα pre-trained μοντέλο όπως το Resnet-50.



# Image Similarity Analysis

Με τη χρήση του Resnet-50 μοντέλου θέλουμε να πετύχουμε τα εξής:

1. Χρήση transfer learning για να επιβεβαιώσουμε ότι το μοντέλο μπορεί να κατηγοριοποιήσει αποδοτικά τα δεδομένα.
2. Χρήση του τελευταίου Linear Layer για την παραγωγή του representation των δεδομένων.
3. Χρήση του εκπαιδευμένου μοντέλου για την παραγωγή representations για test δεδομένα.

# Image Similarity Analysis

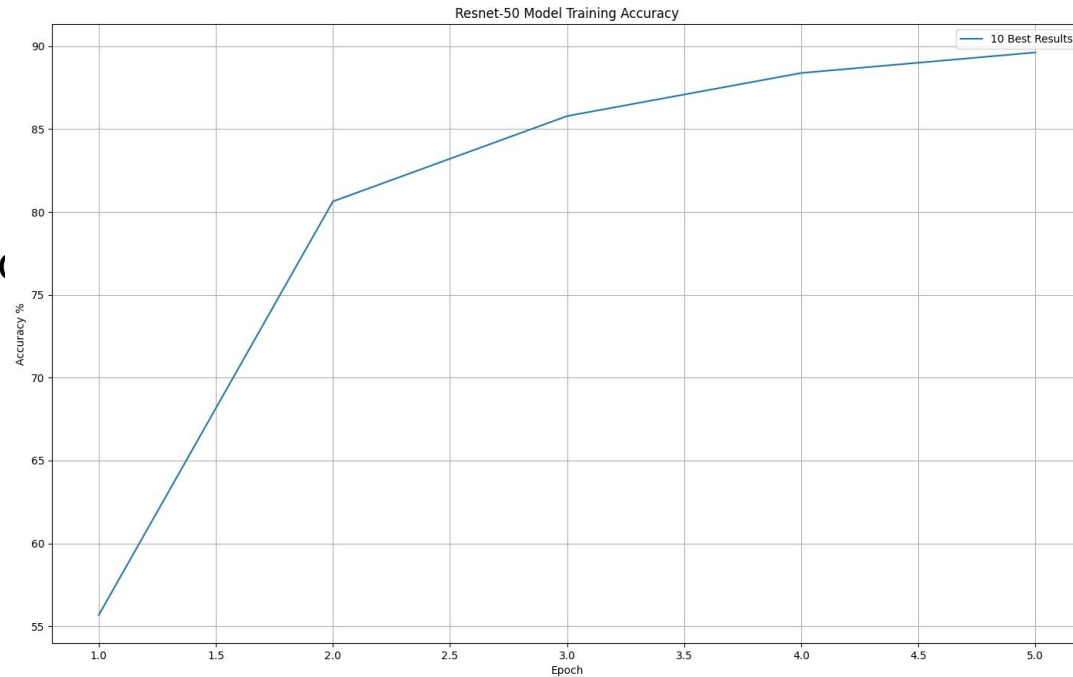
Για την εκπαύδεση του μοντέλου χωρίσαμε τα δεδομένα από την παραγωγή frames σε 90% train και 10% test. Για την εκπαίδευση χρειάστηκε να δημιουργήσουμε μια διαδικασία κανονικοποίησης και augmentation για τα δεδομένα.

Τα βήματα είναι τα εξής:

1. Resize image 224x224.
2. Apply affine transform on an image.
3. Apply Gaussian Blur.
4. Apply random horizontal flip.
5. Normalise the mean and standard deviation.

# Image Similarity Analysis

1. Learning Rate: 0.0001
2. Epochs: 5
3. Batch size: 32
4. Loss Function: Cross Entropy Loss
5. Optimiser Function: Stochastic gradient descent

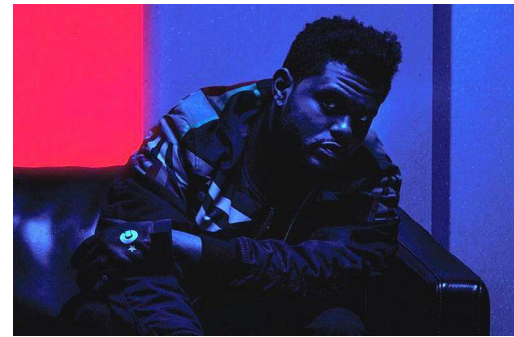




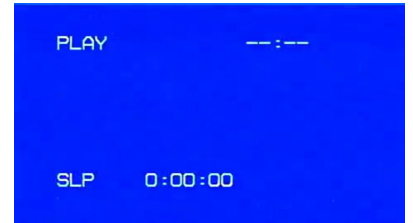
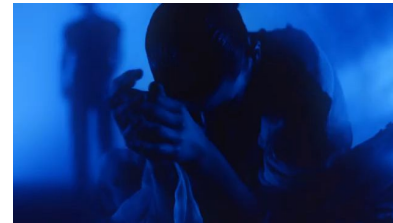
# Image Similarity Analysis

Για να ελέγξουμε την απόδοση του μοντέλου και την ομοιότητα του συνημιτόνου χρησιμοποιούμε ένα δημοφιλές music Video Clip, [StarBoy](#)

Για κάθε music video clip επιλέγουμε 10 τυχαία frames έτσι ώστε να έχουμε μια πιο αντιπροσωπευτική εικόνα για όλο τη διάρκεια του video.



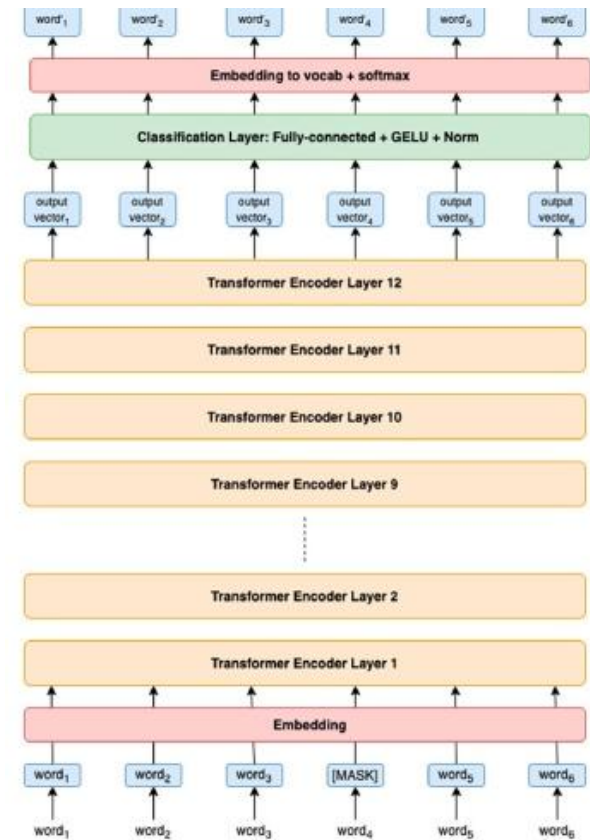
Test Image



Similar Images

# Text Similarity Analysis

Στην ταξινόμηση κειμένου χρησιμοποιούμε το BERT-base-uncased pre-trained μοντέλο. Ομοίως στόχος μας είναι να δημιουργήσουμε ένα Representation των text data στο οποίο να μπορούμε να εφαρμόσουμε Cosine Similarity ως το μέτρο ομοιότητας.



# Text Similarity Analysis

To Representation σε αυτή την περίπτωση προέρχεται από το τελικό Embedding Layer. Το αποτέλεσμα είναι ένας fixed-size πίνακας, όπου σε αυτή την περίπτωση έχει max size 512.

Κάθε token στην input sequence αντιστοιχίζεται με ένα πίνακα στο embedding layer, όπου αυτοί οι πίνακες συλλαμβάνουν contextual information, δηλαδή τις σχέσεις μεταξύ των λέξεων.

# Text Similarity Analysis

Το αποτέλεσμα σε αυτή την περίπτωση είναι μια λίστα με τα πιο κοντινά music video clips και το αντίστοιχο Cosine Similarity Score.

Top 10 similar songs:

Song: Bruno-Mars-That-s-What-I-Like.txt Score: 0.9069370627403259

Song: Conan-Gray-Heather.txt Score: 0.8974332809448242

Song: Post-Malone-Swae-Lee-Sunflower.csv Score: 0.8898136019706726

Song: Coldplay-X-BTS-My-Universe.txt Score: 0.8890013694763184

Song: Khalid-Young-Dumb-Broke.txt Score: 0.88477623462677

Song: Elley-Duh-MIDDLE-OF-THE-NIGHT.csv Score: 0.8831318616867065

Song: Harry-Styles-Golden.txt Score: 0.8817164897918701

Song: Marshmello-x-Jonas-Brothers-Leave-Before-You-Love-Me.csv Score: 0.8806685209274292

Song: POP-SMOKE-WHAT-YOU-KNOW-BOUT-LOVE.txt Score: 0.8787678480148315

Song: Imagine-Dragons-Bad-Liar.csv Score: 0.8710405230522156

# Fusion and Integration of Similarity Metrics

Για κάθε ένα από τα αποτελέσματα ομοιότητας εικόνας και κειμένου χρησιμοποιούμε την παρακάτω συνάρτηση για την δημιουργία ενός συνολικού Similarity Score.

$\text{Combined\_Similarity\_Score} = (w1 * \text{Text\_Similarity\_Score}) + (w2 * \text{Image\_Similarity\_Score}),$

όπου  $w1, w2$  παίρνουν τις τιμές  $\{0.4, 0.6\}$  αντίστοιχα.

Demo

# Συμπεράσματα

- Το σύστημα μας επέδειξε καλή απόδοση, επιτυγχάνοντας εντυπωσιακά αποτελέσματα στα προβλεπόμενα task. Ο συνδυασμός αυτών των modalities επέτρεψε βελτιωμένη κατανόηση και βελτιωμένες δυνατότητες λήψης αποφάσεων για την εύρεση των πιο όμοιων music video clips.
- Η προσθήκη audio έχει μεγάλες δυνατότητες για περαιτέρω βελτίωση της απόδοσης του συστήματος. Με την ενσωμάτωση δεδομένων ήχου, το σύστημα μπορεί να αξιοποιήσει τη δύναμη της ανάλυσης ήχου. Αυτή η συμπερίληψη θα παρέχει μια πιο ολοκληρωμένη και ολιστική κατανόηση των δεδομένων, οδηγώντας σε ακόμα καλύτερα αποτελέσματα.

Σας ευχαριστώ για την *προσοχή* σας