Ιόνιο Πανεπιστήμιο

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΠΛΗΡΟΦΙΚΗΣ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ



Διαχείριση Πολυμεσικής πληροφορίας

Τουλάκης Αλέξανδρος Π2015067

Επιβλέπων : Ιωάννης Καρύδης Επίκουρος καθηγητής , Ιόνιο Πανεπιστήμιο

Πρόβληματα – Ανάγκες Που Παρουσιάζονται

- Ογκος πληροφορίας;
- Ανάγκη Οργάνωσης πληροφοράς θέματα
- υπολογισμός ομοιότητας

Λύσεις Που Προτείνουν οι υπόλοιποι

Ανέρευση Θεμάτων [1, 2]

- LSI (Latent Semantic Indexing)
- PLSI (Probabilistic Latent Semantic Indexing)
- LDA (Latent Dirichlet Allocation)

Μέθοδοι υπολογισμού ομοιότητας [3, 4]

- με χρήση χαρακτηριστικών χαμηλού επιπέδου
- με χρήση ετικετών π.χ Jaccard, Subset, Tanimoto

Λύση Που Προτάθηκε

- Αλγόριθμος LDA (Latent Dirichlet Allocation)
- Παραλλαγή της Μεθόδου της Subset

Πειράματα

1) Σύγκριση υλοποιήσεων LDA σε ελληνικά κείμενα
2) Παραλλαγή Subset μουσικά κομμάτια
3) Παραλλαγή Subset για Εικόνες

Προεπεξεργασία Κειμένου

- Stopwords
- Stemming
- Capitalization
- Punctuations removal

Μέτρα Αξιολόγησης

- Precision
- Recall
- F1-Score
- K-Fold Cross Validation

Subset

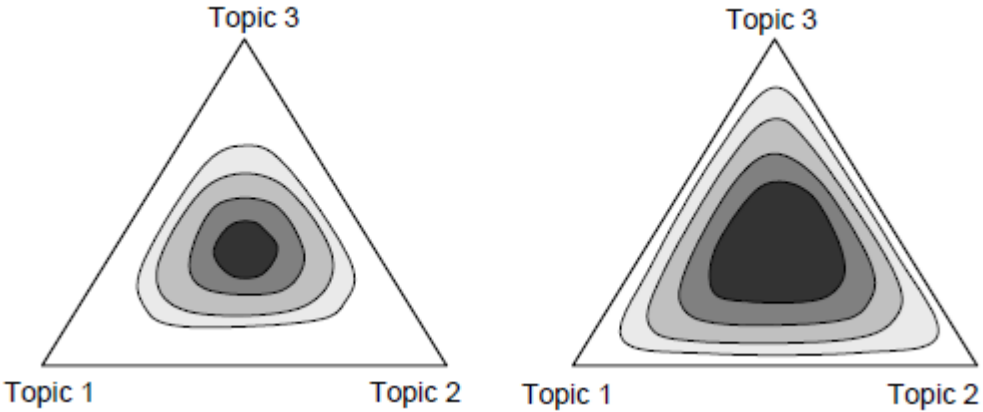
Κλασσική Subset;
Παραλλαγή ;

Latent Dirichlet Allocation

- Αλγόριθμος για την εύρεση θεματολογίας
- Μη επιβλεπόμενης μάθησης
- Bag of words
- έγγραφο ποικίλα θέματα φ(κ)
- θέμα κατανομή λέξεων θ(κ)
- Dirichlet Priors α, β

Dirichlet Prior

[1]



Statistical Inference

1) Variational Gibs Sampling

2) Collapsed Gibs Sampling

$$P(z_i = j \mid \mathbf{z}_{-i}, w_i, d_i, \cdot) \propto \frac{C_{w_i j}^{WT} + \beta}{\sum_{w=1}^{W} C_{wj}^{WT} + W\beta} \frac{C_{d_i j}^{DT} + \alpha}{\sum_{t=1}^{T} C_{d_i t}^{DT} + T\alpha}$$

$$\phi_i'^{(j)} = \frac{C_{ij}^{WT} + \beta}{\sum_{k=1}^{W} C_{kj}^{WT} + W\beta} \qquad \theta_j'^{(d)} = \frac{C_{dj}^{DT} + \alpha}{\sum_{k=1}^{T} C_{dk}^{DT} + T\alpha} \qquad (4)$$

Πείραμα 1

- Δεδομένα
- Προεπεξεργασία
- Εκτέλεση των δύο υλοποιήσεων
- Αξιολόγηση των αποτελεσμάτων

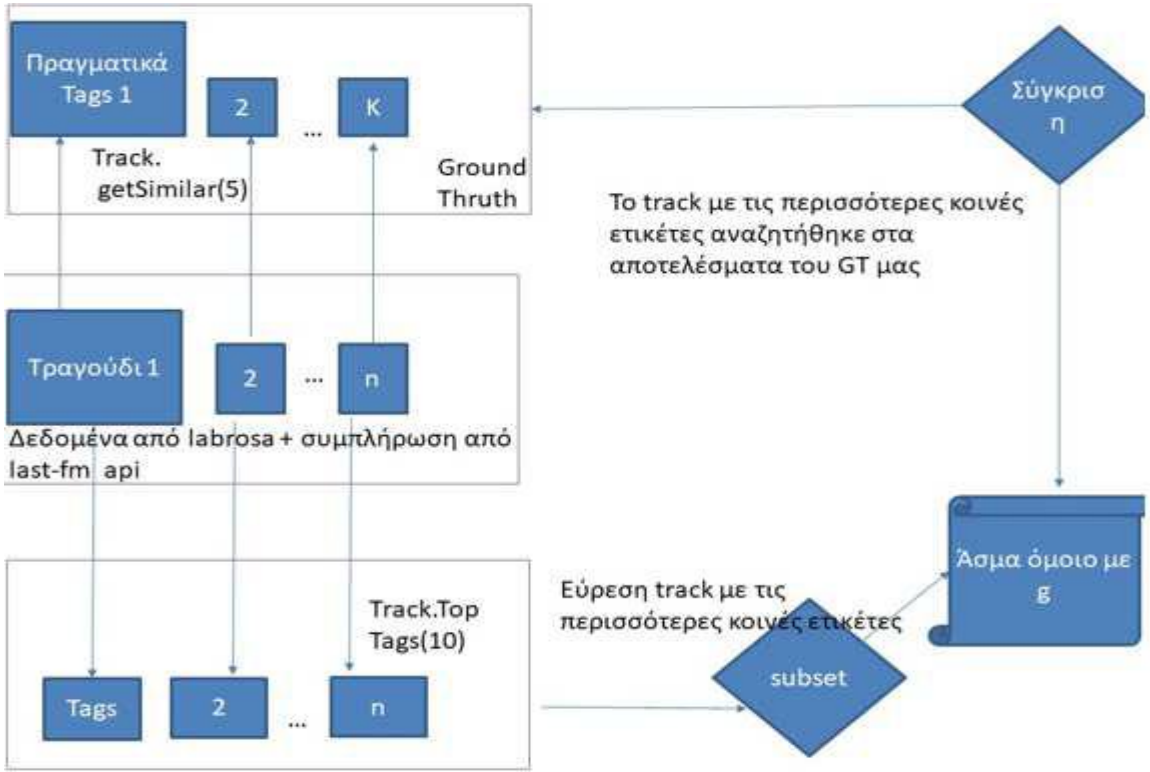
Παράμετροι Εκτέλεσης πειράματος

Όνομα Παραμέτρου	Τιμή
[1] Dirichlet Prior a	50/T
[1] Dirichlet Prior b	0.1
Αριθμός επαναλήψεων	100
Αριθμός θεμάτων	10
K-Fold Cross Validation	10
Seeds	25

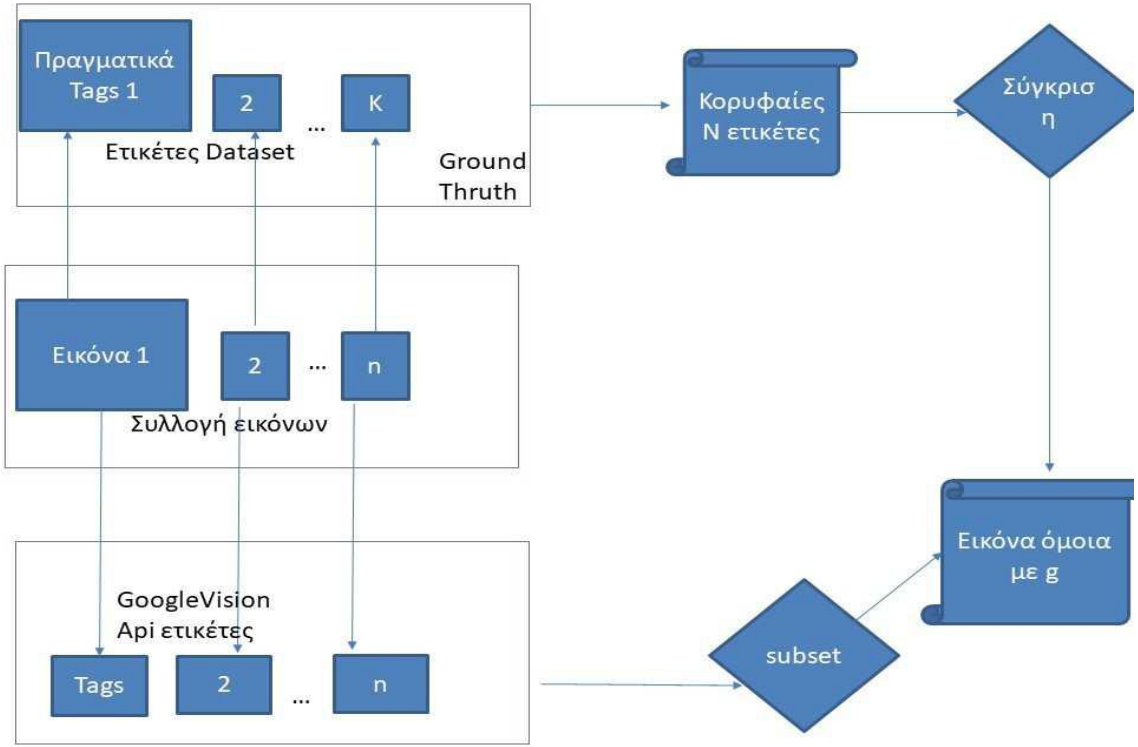
Αξιολόγηση δύο υλοποίησεων



Πείραμα 2


Πραγματικά Tags 1
2
...
K
Track. getSimilar(5)
Ground Thruth
Σύγκριση
Το track με τις περισσότερες κοινές ετικέτες αναζητήθηκε στα αποτελέσματα του GT μας
Τραγούδι 1
2
...
n
Δεδομένα από labrosa + συμπλήρωση από last-fm api
Άσμα όμοιο με g
Εύρεση track με τις περισσότερες κοινές ετικέτες
Track.Top Tags(10)
subset
Tags
2
...
n

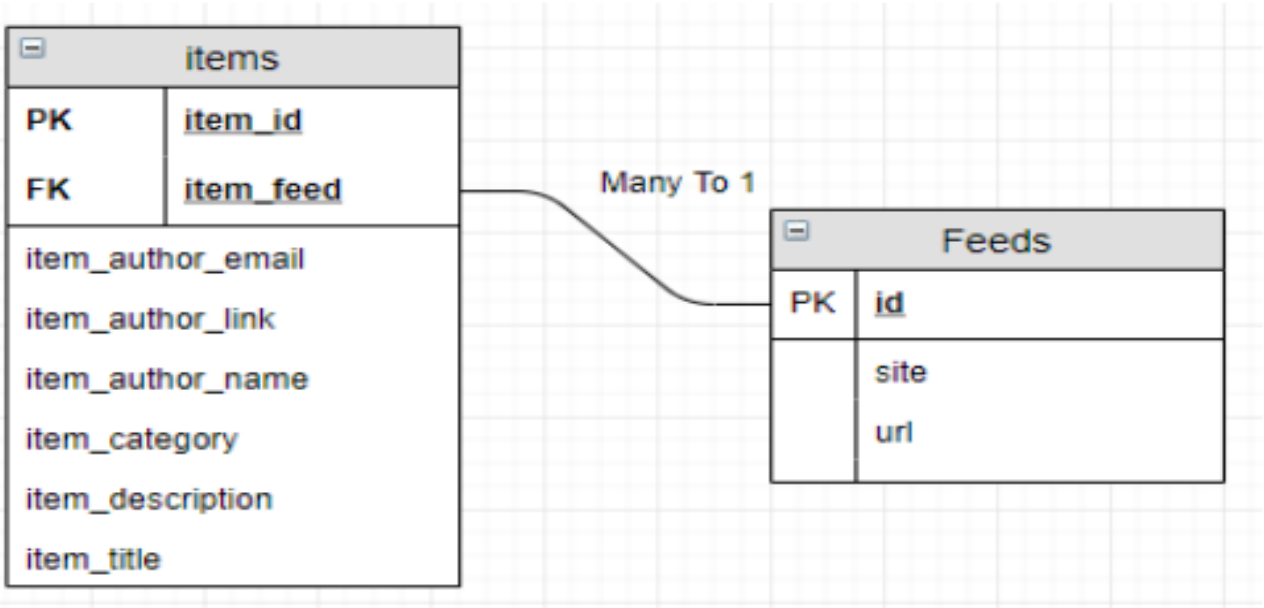
Πείραμα 3


Πραγματικά Tags 1
2
...
K
Ετικέτες Dataset
Ground Thruth
Κορυφαίες Ν ετικέτες
Σύγκριση
Εικόνα 1
2
...
n
Συλλογή εικόνων
Εικόνα όμοια με g
GoogleVision Api ετικέτες
Tags
2
...
n
subset

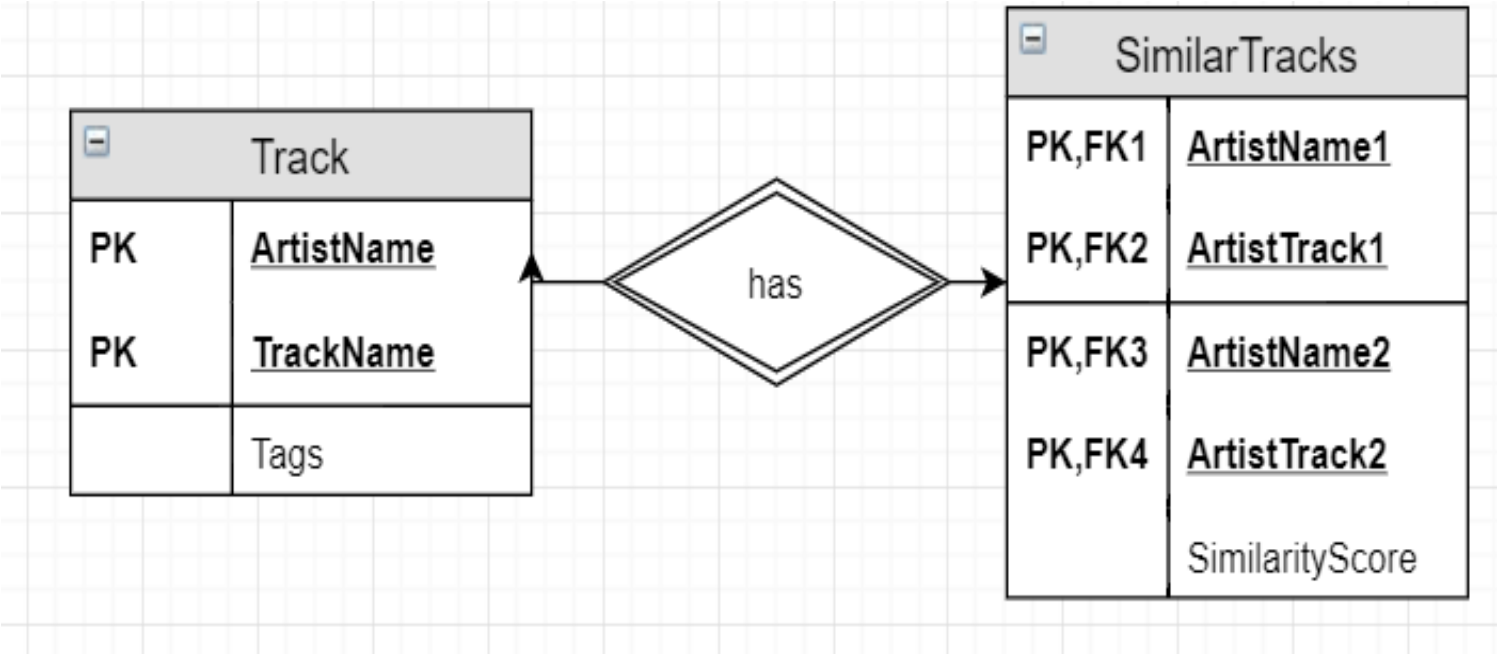
Αποτελέσματα

Πείραμα	F1-Score
Πείραμα 1 LDA Toulaki	0.519125
Πείραμα 1 LDA Gensim	0.560125
Πείραμα 2 Μουσικά Κομμάτια	0.81
πείραμα 3 Εικόνες	0.712

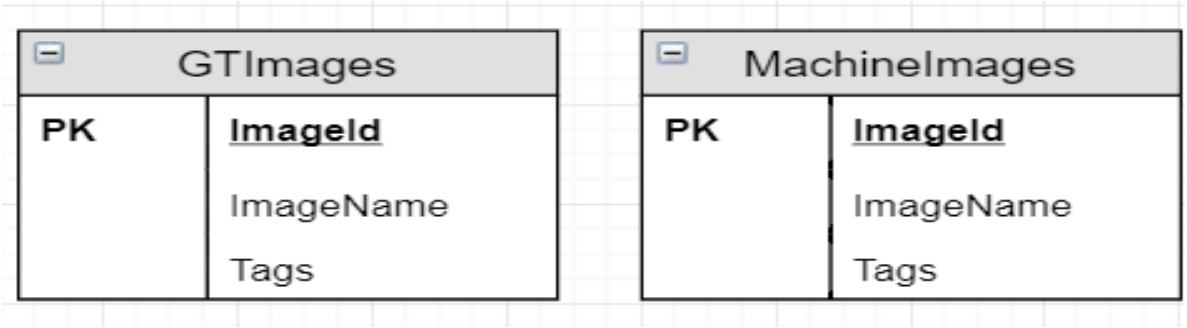
Σχήμα Βάσεων Δεδομένων Πείραμα 1



Σχήμα Βάσεων Δεδομένων Πείραμα 2



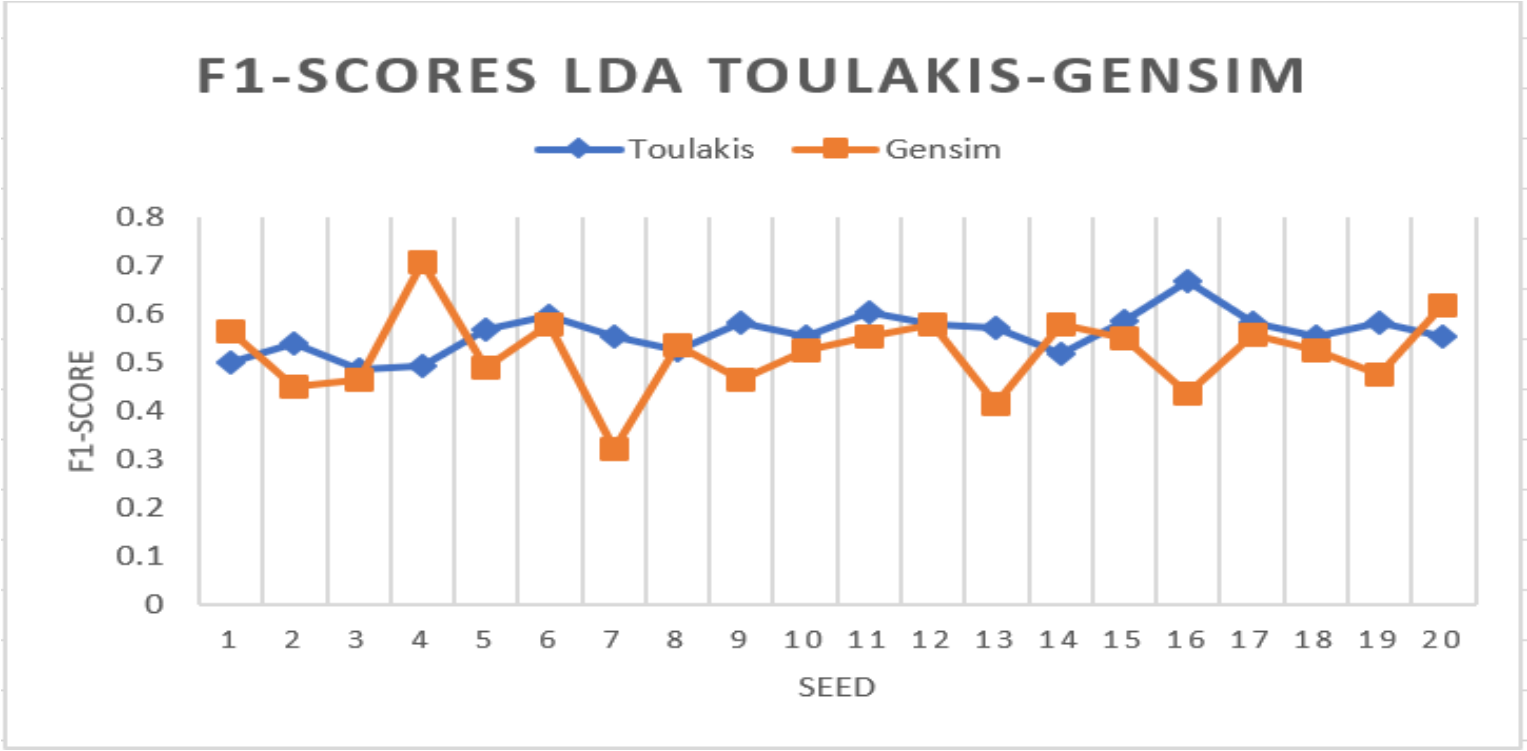
Σχήμα Βάσεων Δεδομένων Πείραμα 3



Πίνακες Βάσεις Δεδομένων

Πείραμα	Όγκος Δεδομένων	Προέλευση Δεδομένων
Κείμενα	200	Επέκταση έτοιμη Βάσης από τον επιβλέποντα
Μουσικά Κομμάτια	5000	Last-fm
Εικόνες	110	LabelMe

LDA αποτελέσματα Γράφημα



Συμπεράσματα

Μεταξύ LDA υλοποιήσεων

Στα πειράματα Εικόνων/Μουσικής

Βιβλιογραφία Παρουσίασης

[1] Griffiths and Mark Steyversx: Finding scientic topics

[2] Diane J. Hu Latent Dirichlet Allocation for Text, Images, and Music

[3] Ioannis Karydis, Spyros Sioutas, Andreas Kanavos, Markos Avlonitis LeSiM:

A Novel Lexical Similarity Measure Technique for Multimedia Information Re-

trieval.

[4] Zheng Hu, Xiangfeng Lu Measuring Semantic Relatedness between Flickr Images: From a social tag based view