

1ο ΘΕΜΑ: Συσταδοποίηση ειδησεογραφικών άρθρων

1η επιλογή)

Κατεβάστε το dataset που υπάρχει εδώ:

<https://github.com/google-research-datasets/NewSHead>

Επιλέξτε 1000 τουλάχιστον άρθρα (ιδανικά από το train dataset) τα οποία ανήκουν σε πολλές και διαφορετικές μικρές συστάδες ομοειδών άρθρων (200-300 συστάδες)

Κατεβάστε τα άρθρα σε αρχεία.

Ορίστε αναπαραστάσεις των κειμένων.

Ορίστε μέτρα ομοιότητας.

Δημιουργήστε αντίστοιχες συστάδες σε πλήθος με τις αρχικές.

Αξιολογήστε τα αποτελέσματα διαφορετικών τεχνικών.

2η επιλογή)

Κατεβάστε το dataset που υπάρχει εδώ:

<https://www.kaggle.com/datasets/crawford/20-newsgroups>

Ορίστε αναπαραστάσεις των κειμένων.

Ορίστε μέτρα ομοιότητας.

Δημιουργήστε αντίστοιχες συστάδες σε πλήθος με τις αρχικές.

Αξιολογήστε τα αποτελέσματα διαφορετικών τεχνικών.

2ο ΘΕΜΑ: Δημιουργία περιλήψεων (ή τίτλων) για ομάδες άρθρων

1η επιλογή)

Κατεβάστε το dataset που υπάρχει εδώ:

<https://github.com/google-research-datasets/NewSHead>

Επιλέξτε 300 τουλάχιστον συστάδες ομοειδών άρθρων (ιδανικά από το train dataset)

Κατεβάστε τα άρθρα σε αρχεία.

Ορίστε αναπαραστάσεις των κειμένων.

Υλοποιήστε μια εξαγωγική τεχνική δημιουργίας περιλήψεων (μπορείτε εναλλακτικά να υλοποιήσετε μια αφαιρετική τεχνική)

Δημιουργήστε τίτλους.

Αξιολογήστε το κείμενο που παράξατε σε σχέση με τον τίτλο που δίνεται στο dataset για κάθε συστάδα.

2η επιλογή)

Κατεβάστε το dataset που υπάρχει εδώ:

<https://www.kaggle.com/datasets/sbhatti/news-summarization>

Επιλέξτε 300 τουλάχιστον άρθρα.
Ορίστε αναπαραστάσεις των κειμένων.
Υλοποιήστε μια εξαγωγική τεχνική δημιουργίας περιλήψεων (μπορείτε εναλλακτικά να υλοποιήσετε μια αφαιρετική τεχνική)
Δημιουργήστε περιλήψεις αντίστοιχου μήκους με τις περιλήψεις των αρχείων.
Αξιολογήστε το κείμενο που παράζετε σε σχέση με την περίληψη που δίνεται στο dataset.

3ο ΘΕΜΑ: Δημιουργία μιας μηχανής αναζήτησης

1η επιλογή)

Κατεβάστε το dataset που υπάρχει εδώ:

<https://www.kaggle.com/datasets/jrobischon/wikipedia-movie-plots>

και δημιουργήστε πάνω σε αυτό μια μηχανή αναζήτησης.

Ορίστε αναπαραστάσεις των κειμένων.

Υλοποιήστε έναν μηχανισμό ευρετηρίασης.

Υλοποιήστε το μηχανισμό ανάκτησης και βαθμονόμησης των αποτελεσμάτων.

Αξιολογήστε το χρόνο που χρειάζεται για να σας απαντήσει σε 30 ερωτήματα. Αξιολογήστε τη συνάφεια των πρώτων 10 αποτελεσμάτων που σας επιστρέφονται σε κάθε ερώτηση.

2η επιλογή)

Μελετήστε το arXiv API

<https://info.arxiv.org/help/api/index.html>

Κατεβάστε τουλάχιστον 10000 άρθρα (τίτλο, συγγραφείς, abstract) και δημιουργήστε πάνω σε αυτά μια μηχανή αναζήτησης.

Ορίστε αναπαραστάσεις των κειμένων.

Υλοποιήστε έναν μηχανισμό ευρετηρίασης.

Υλοποιήστε το μηχανισμό ανάκτησης και βαθμονόμησης των αποτελεσμάτων.

Αξιολογήστε το χρόνο που χρειάζεται για να σας απαντήσει σε 30 ερωτήματα. Αξιολογήστε τη συνάφεια των πρώτων 10 αποτελεσμάτων που σας επιστρέφονται σε κάθε ερώτηση.