1ο ΘΕΜΑ: Συσταδοποίηση σύντομων κειμένων

1η επιλογή)
Κατεβάστε το dataset που υπάρχει εδώ:
https://www.kaggle.com/datasets/smagnan/1-million-reddit-comments-from-40-subreddits

Επιλέξτε 1000 τουλάχιστον σχόλια (ιδανικά από το train dataset) τα οποία ανήκουν σε πολλά και διαφορετικά topics (subreddits)
Ορίστε αναπαραστάσεις των σχολίων.
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://huggingface.co/datasets/olm/gdelt-news-headlines

Ορίστε αναπαραστάσεις των κειμένων.
Ορίστε μέτρα ομοιότητας.
Δημιουργήστε συστάδες τίτλων άρθρων που να αναφέρονται στο ίδιο γεγονός.
Αξιολογήστε τα αποτελέσματα διαφορετικών τεχνικών ως προς την εσωτερική ποιότητα των clusters.
EXTRA: Αποτυπώστε χρονικά την εξέλιξη των γεγονότων.

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

1η επιλογή)
Κατεβάστε το dataset που υπάρχει εδώ:
https://huggingface.co/datasets/ccdv/cnn_dailymail

Επιλέξτε 500 τουλάχιστον άρθρα (από το train dataset)
Ορίστε αναπαραστάσεις των κειμένων.
Υλοποιήστε μια εξαγωγική τεχνική δημιουργίας περιλήψεων (μπορείτε εναλλακτικά να υλοποιήσετε μια αφαιρετική τεχνική)
Δημιουργήστε τίτλους/highlights
Αξιολογήστε το κείμενο που παράξατε σε σχέση με τα highlights που δίνονται για κάθε άρθρο.

2η επιλογή)
Κατεβάστε το dataset που υπάρχει εδώ:
https://huggingface.co/datasets/alexfabbri/multi_news

Επιλέξτε 500 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 παράξατε σε σχέση με την περίληψη που δίνεται στο dataset.

3ο ΘΕΜΑ: Δημιουργία μιας μηχανής ερωταπαντήσεων

1η επιλογή)
Κατεβάστε το dataset που υπάρχει εδώ:
https://www.kaggle.com/datasets/jrobischon/wikipedia-movie-plots
και δημιουργήστε πάνω σε αυτό μια μηχανή που θα απαντά ερωτήσεις του χρήστη, πάνω στην πλοκή (plot) των ταινιών.

Ορίστε αναπαραστάσεις των κειμένων.
Υλοποιήστε έναν μηχανισμό ευρετηρίασης της πλοκής (είτε ανά ταινία, είτε συνολικά)
Υλοποιήστε το μηχανισμό ανάκτησης σχετικών κειμένων που θα λειτουργούν ως απάντηση (ή ως βάση για την παραγωγή της απάντησης).

Αξιολογήστε το χρόνο που χρειάζεται για να σας απαντήσει σε 30 ερωτήματα. Αξιολογήστε τη συνάφεια/ορθότητα του πρώτου και των πρώτων 3 αποτελεσμάτων που σας επιστρέφονται σε κάθε ερώτηση.
Extra: Δοκιμάστε να δίνετε την ερώτηση με 4 πιθανές απαντήσεις (μία μόνο σωστή) και να επιλέγει η μηχανή τη σωστή απάντηση.

2η επιλογή)
Κατεβάστε το dataset που υπάρχει εδώ:
https://www.kaggle.com/datasets/thedevastator/the-stanford-question-answering-dataset
Χρησιμοποιήστε τουλάχιστον 100 διαφορετικές παραγράφους από το dataset.

Ορίστε αναπαραστάσεις των παραγράφων του corpus.
Για κάθε ερώτηση (από τις ερωτήσεις του corpus):
- ανακτήστε τις k πιο σχετικές παραγράφους (μπορεί και k=1)
- εφαρμόστε βαθμονόμηση των αποτελεσμάτων

Υλοποιήστε έναν μηχανισμό εξαγωγής απάντησης από τις k σχετικές παραγράφους, π.χ. επιλογή προτάσεων ή δημιουργία απάντησης

Αξιολογήστε το χρόνο που χρειάζεται για να σας απαντήσει σε 30 ερωτήματα. Αξιολογήστε τη συνάφεια των πρώτων k αποτελεσμάτων (παραγράφων) που σας επιστρέφονται σε κάθε ερώτηση. Αξιολογήστε την ορθότητα των τελικών απαντήσεων.