Εργασία Ανάλυσης Φυσικής γλώσσας 2025

Επισκόπηση

Αυτή η εργασία απαιτεί από τους φοιτητές να εφαρμόσουν τεχνικές σημασιολογικής ομοιότητας, ενσωμάτωσης λέξεων (word embeddings), και γλωσσικής ανακατασκευής. Ο στόχος είναι να μετασχηματιστούν μη δομημένα ή σημασιολογικά αμφίβολα κείμενα σε σαφείς, ορθές/ ορθολογικές και καλά δομημένες εκδοχές.

Η ανάλυση αυτών των ανακατασκευών θα βασιστεί στη συνάφεια μέσω συνημιτόνου (cosine similarity), στις ενσωματώσεις λέξεων και σε τεχνικές NLP. Οι φοιτητές πρέπει να τεκμηριώσουν τα ευρήματά τους σε δομημένη αναφορά συνοδευόμενη από εκτελέσιμο και αναπαράξιμο κώδικα με διατήρηση ίδιων αποτελεσμάτων ανα εκτέλεση (παραδοτέο 3).

———

Παραδοτέα εργασίας - Υποχρεωτική - Απαλλακτική

Κείμενο 1:

“Today is our dragon boat festival, in our Chinese culture, to celebrate it with all safe and great in our lives. Hope you too, to enjoy it as my deepest wishes.
Thank your message to show our words to the doctor, as his next contract checking, to all of us.
I got this message to see the approved message. In fact, I have received the message from the professor, to show me, this, a couple of days ago. I am very appreciated the full support of the professor, for our Springer proceedings publication”

Κείμενο 2:

“During our final discuss, I told him about the new submission — the one we were waiting since last autumn, but the updates was confusing as it not included the full feedback from reviewer or maybe editor?
Anyway, I believe the team, although bit delay and less communication at recent days, they really tried best for paper and cooperation. We should be grateful, I mean all of us, for the acceptance and efforts until the Springer link came finally last week, I think.
Also, kindly remind me please, if the doctor still plan for the acknowledgments section edit before he sending again. Because I didn’t see that part final yet, or maybe I missed, I apologize if so.
Overall, let us make sure all are safe and celebrate the outcome with strong coffee and future targets”

Παραδοτέο 1: Ανακατασκευή Κειμένου

Απο τα παραπάνω κείμενα σας ζητείται να υλοποιήσετε τα εξής:

A. Ανακατασκευή 2 προτάσεων της επιλογής σας με αυτόματο που θα διαμορφώσετε εσείς *__(1)__
B. Ανακατασκευή του συνόλου των 2 κειμένων με χρήση 3 διαφορετικών αυτόματων βιβλιοθηκών python pipelines __*(2)__
C. Συγκρίνετε τα αποτελέσματα της κάθε προσέγγισης με τις κατάλληλες τεχνικές

Ο στόχος σας είναι να ανακατασκευάσετε κάθε κείμενο σε μια σαφή, καλά δομημένη και σημασιολογικά ακριβή εκδοχή. Πρέπει να βεβαιωθείτε ότι το κείμενο διατηρεί το αρχικό του νόημα, βελτιώνοντας τη σαφήνεια, τη συνοχή και τον σχετικό τόνο.

———

Παραδοτέο 2: Υπολογιστική Ανάλυση

Χρησιμοποιήστε ενσωματώσεις λέξεων (Word2Vec, GloVe, FastText, BERT(embeddings), κ.λπ.)*__(2)__ και δικές σας -custom- αυτόματες ροές εργασίας NLP (προεπεξεργασία, λεξιλόγιο, ενσωμάτωση λέξεων, εννοιολογικά δέντρα κλπ)*__(1)__ για να αναλύσετε την ομοιότητα των λέξεων πριν και μετά την ανακατασκευή. Υπολογίστε βαθμολογίες συνημιτόνου (cosine similarity) μεταξύ των αρχικών

και των ανακατασκευασμένων εκδοχών. Συγκρίνετε τις μεθόδους ως προς τα A, B, C του παραδοτέου 1.

Οπτικοποιήστε τις ενσωματώσεις λέξεων για τα A,B,C χρησιμοποιώντας PCA/t-SNE για να αποδείξετε τις μετατοπίσεις στον σημασιολογικό χώρο.

———

Παραδοτέο 3: Δομημένη Αναφορά
Η αναφορά πρέπει να περιλαμβάνει:
Εισαγωγή:
Εξηγήστε τη σημασία της σημασιολογικής ανακατασκευής και την εφαρμογή του NLP στη διαδικασία.

Μεθοδολογία:
Περιγράψτε τις στρατηγικές ανακατασκευής (γραμματική, αξιώματα, γλωσσικοί κανόνες κλπ) για τα A,B,C.
Αναλύστε τις υπολογιστικές τεχνικές που χρησιμοποιήσατε (συνάφεια συνημιτόνου, ενσωματώσεις λέξεων κλπ) για τα A,B,C.

Πειράματα & Αποτελέσματα:
Παρουσιάστε παραδείγματα πριν/μετά την ανακατασκευή και πλήρη αναφορά και ανάλυση του Παραδοτέου 2.

Συζήτηση:
Πόσο καλά αποτύπωσαν οι ενσωματώσεις λέξεων το νόημα;
Ποιες ήταν οι μεγαλύτερες προκλήσεις στην ανακατασκευή;
Πώς μπορεί να αυτοματοποιηθεί αυτή η διαδικασία χρησιμοποιώντας μοντέλα NLP;
Υπήρξαν διαφορές στην ποιότητα ανακατασκευής μεταξύ τεχνικών, μεθόδων, βιβλιοθηκών κλπ; Συζητήστε τα ευρήματά σας.

Συμπέρασμα:
Αναστοχασμός επί των ευρημάτων και των προκλήσεων της μελέτης.

Βιβλιογραφία:
Παραθέστε σχετικές δημοσιεύσεις και πηγές που χρησιμοποιήσατε στην έρευνά σας.

———————

GitHub Repository:
Παρέχετε ένα αποθετήριο στο GitHub με αρχείο README.md (χρήση .env και .gitignore για απόκρυψη μυστικών, κωδικών, συνθημάτων) που να εξηγεί την υλοποίηση του έργου.
Χρησιμοποιήστε Python για την ανάπτυξη των πειραμάτων σας.

Εργαλεία:

✅ Version: Python >=3.10

✅ Dependency Management: Poetry

✅ Libraries: Numpy, pandas, scikit-learn, pytorch etc

✅ *Environment Considerations: Conda*