Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, λογότυπο, γραφικά

Το περιεχόμενο που δημιουργείται από τεχνολογία AI ενδέχεται να είναι εσφαλμένο.

Απαλλακτική Εργασία:

Επεξεργασία Φυσικής Γλώσσας

Αλέξανδρος Κόλλιας

Π21225

# Εισαγωγή

Η σημασιολογική ανακατασκευή περιλαμβάνει την αναδιαμόρφωση ενός κειμένου έτσι ώστε να διατηρεί το ίδιο νόημα ή επιπρόσθετη πληροφορία, ωστόσο αλλάζοντας την διατύπωση του. Κάτι το οποίο μπορεί να περιλαμβάνει την αναδιατύπωση φράσεων, ή ανακατανομή τους χωρίς να αλλάζει το νόημα. Στην πράξη πρόκειται για την μεταφορά ενός δοσμένου κειμένου σε άλλα λόγια με τρόπο ώστε να εκφράζεται το ίδιο νόημα και να μεταφέρουμε λεκτικά την ίδια πληροφορία εξ’ολοκλήρου. Η συμβολή του NLP είναι τεράστια διότι πλέον χειροκίνητα δεν κάνουμε παραφράσεις. Μέσω του translating γλώσσα σε γλώσσα παραφράζουμε, ενώ χάρη στα μοντέλα T5, Pegasus, και το BART της Google η παραφράση βελτιώνεται. Πρόκειται για τεχνολογία αναδιατύπωσης κειμένου με ένα κλικ, βασισμένη σε τεχνολογία Tensorflow, η οποία εμπεριέχει ένα τεράστιο data set διαφόρων πληροφοριών το οποίο εφοδιάζει την μηχανική εκμάθησης όπου στην ουσία το: Κατανοεί λέξεις με παρεμφερείς σημασίες Κατανοεί σύνταξη κειμένων προς αναδιαμόρφωση Κατανοεί υπονοούμενα και επιπρόσθετο περιεχόμενο προς του το διατυπώνει σενάρια που η πληροφορία μπορεί να λεχθεί με άλλο τρόπο. Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, γραμματοσειρά

Το περιεχόμενο που δημιουργείται από τεχνολογία AI ενδέχεται να είναι εσφαλμένο.

# Μεθοδολογία

Για τη στρατηγική αναδιατύπωσης των Α, Β, C χρησιμοποιήθηκαν τρία διαφορετικά NLP μοντέλα μηχανικής μάθησης σε κάθε περίπτωση το καθένα με τις δικές του παραδοχές για τη γλωσσική μετασχηματιστική λειτουργία το σχήμα T5 βασίζεται στην προσέγγιση του «συντονισμού/συγγείωσης» των γλωσσικών σχέσεων ενώ το BART υποστηρίζει μια αρχιτεκτονική masked sequence-to-sequence και θεωρείται ιδιαίτερα αποτελεσματικό στη δημιουργία ανθρώπινων επιπέδων κατανόησης του κειμένου το Pegasus έχει προσαρμοστεί για το γλωσσολογικό έργο της σύνοψης αλλά χρησιμοποιήθηκε ως εργαλείο παραφράσεων με συνδεδεμένα γλωσσικά σημεία και μορφολογική σημασία των λεκτικών μετασχηματισμών μέσω του transformer-based encoder-decoder μηχανισμού

Οι υπολογιστικές τεχνικές που εξερευνώνται αλληλουχούν κυρίως στους μετασχηματιστές (transformers) στις ενσωματώσεις λέξεων (word/sentence embeddings) και σε έναν δείκτη συγγένειας γνωστό ως συντελεστής συνημιτόνου (cosine similarity) το τελευταίο χρησιμοποιείται για να συγκρίνει το πόσο κοντά είναι σημασιολογικά μια παραφρασμένη πρόταση με την αρχική το εργαλείο SentenceTransformers αποδίδει τις ενσωματώσεις προτάσεων και με συναφής χρησιμοποιούνται ενδεχομένως για να υπολογίσουν μία έκταση ή ομοιότητα μεταξύ των παραλλαγών προς αξιολόγηση η εξαγωγή των ποιοτικών αποτελεσμάτων κάθε παραφράσεως

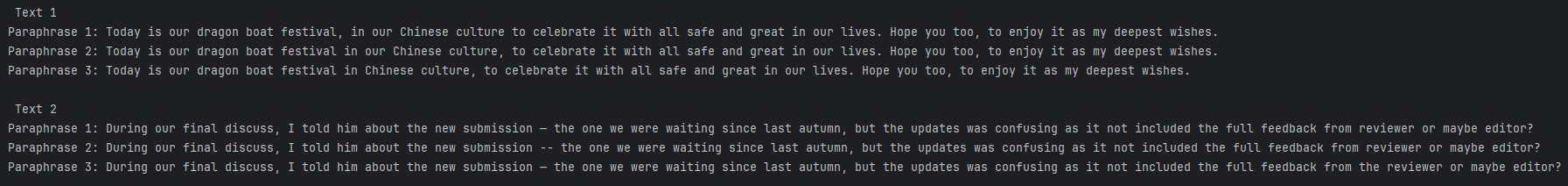
Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, αριθμός

Το περιεχόμενο που δημιουργείται από τεχνολογία AI ενδέχεται να είναι εσφαλμένο.

# Πειράματα & Αποτελέσματα

Για το Παραδοτέο 2 είχαμε 2 κείμενα πρωτότυπα και ζητήθηκε να βγάλουμε παραφράσεις χρησιμοποιώντας NLP πρακτικά περάσαμε τα κείμενα από 3 διαφορετικά μοντέλα που δώσαν παραφράσεις και βγάζαμε το αποτέλεσμα κάθε παραλλαγή την κοιτούσαμε για να δούμε αν έβγαλε κάτι που παραμένει νοηματικά κοινό παραδείγματα στο Dragon Boat Festival η Bart παραλλαγή έβγαλε πολύ κοντινή πρόταση με το ίδιο νόημα αλλά διαφορετική διατύπωση T5 παραπλήσια ενώ το Pegassus το άφηνε να φύγει λίγο από το νόημα του αρχικού

Στηρίξαμε την ανάλυση του Παραδοτέου καθαρά με το πόσο συνάφεια είχε κάθε παραλλαγή και το αποτέλεσμα ήταν ότι το BART είχε την πιο πιστή και ταυτόχρονα πιο φυσική συμπεριφορά σε σύγκριση με τα άλλα μοντέλα και από άποψη ύφους και από άποψη να κρατήσει το νόημα του κειμένου το T5 ήταν αρκετά πιστό αλλά σχετικά τυποποιημένο στη παραλλαγή ενώ το Pegassus ήταν το λιγότερο πιστό στη παραφράση όταν το κείμενο ήταν λίγο πιο σύνθετο και με λίγοι πιο striking έννοιες στο αρχικό



# Συζήτηση

Οι ενσωματώσεις λέξεων και κυρίως οι ενσωματώσεις προτάσεων μέσω του SentenceTransformers μας βοήθησαν πολύ καθώς είχαμε έναν αριθμητικό τρόπο να δούμε πόσο κοντά ήταν μια πρόταση στο αρχικό κείμενο. Όταν, για παράδειγμα, το cosine similarity ήταν πολύ υψηλό (π.χ. 0.9992) σε μοντέλα όπως το BART, αυτό σήμαινε πως τα embeddings μπορούσαν πολύ καλά να αναπαραστήσουν το νόημα μιας πρότασης και να το εκφράσουν με άλλα λόγια. Κάτι που, όμως, δεν ισχύει πάντα για άλλα μοντέλα όπως το Pegasus, καθώς παρατηρήσαμε πως η σημασιολογική τους ακρίβεια ήταν πιο περιορισμένη και συχνά εισήγαγαν νέες έννοιες.

Το μεγαλύτερο πρόβλημα ήταν ότι η μηχανή συνέχιζε να παράγει κείμενα που, παρόλο που ήταν σωστά στη δομή τους, δεν μετέδιδαν πάντα με ακρίβεια το αρχικό μήνυμα, ιδίως όταν το αρχικό κείμενο περιλάμβανε μεταφορές, πολιτισμικά στοιχεία ή ασυνήθιστη δομή των προτάσεων. Επίσης, ήταν πρόκληση να βεβαιωθούμε πως το μοντέλο δεν θα επαναλάμβανε τα ίδια και τα ίδια, αλλά θα παρείχε πραγματικά νέες διατυπώσεις.

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης

Το περιεχόμενο που δημιουργείται από τεχνολογία AI ενδέχεται να είναι εσφαλμένο.

# Συμπέρασμα

Κατά τη διάρκεια του πειράματος γίναμε αντιληπτοί πόσο σημαντικό είναι να διαλέγουμε το κατάλληλο μοντέλο για κάθε γλωσσική εργασία—η παραφραστική ανακατασκευή είναι ένα πολύπλοκο θέμα που σχετίζεται με τη δυσκολία της γλώσσας και τον στόχο που έχει το κάθε μοντέλο. Κάθε μοντέλο έχει τα πλεονεκτήματά του αλλά και τα μειονεκτήματά του, κάτι που ήταν έμφανες από τα ποσοτικά αποτελέσματα και τις συγκρίσεις που κάναμε. Τα βαθιά μοντέλα, όπως τα transformers, και τεχνικές όπως τα embeddings είναι πολύ ισχυρά εργαλεία, ωστόσο η αξιολόγηση από ανθρώπους είναι πάντα απαραίτητη για να βεβαιωθούμε για την ποιότητα του τελικού αποτελέσματος. Επιπλέον, το να συνδυάζουμε την ποικιλία που παράγει ο αλγόριθμος με υπολογιστικά κριτήρια, όπως το cosine similarity, είναι χρήσιμο, αλλά δεν είναι αρκετό από μόνο του για να πάρουμε μια τελική απόφαση.

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, γραμματοσειρά

Το περιεχόμενο που δημιουργείται από τεχνολογία AI ενδέχεται να είναι εσφαλμένο.

# Βιβλιογραφία

Πιο πολύ PyCharm tutorials για NLP μοντέλα