ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ

ΕΡΓΑΣΙΑ 2025

ΠΟΥΛΑΔΟΣ ΚΩΝΣΤΑΝΤΙΝΟΣ Π22146

ΕΙΣΑΓΩΓΗ

Η σημασιολογική ανακατασκευή αφορά την επανέκφραση των κειμένων ώστε να είναι σωστά δομημένα γραμματικά και συντακτικά χωρίς ορθογραφικά λάθη, διατηρόντας το αρχικό νόημα, με την χρήση αυτομάτων και ημι-αυτομάτων.

Στη συγκεκριμένη διαδικασία η εφαρμογή του NLP μας δίνει την δυνατότητα να αυτοματοποιήσουμε την ανακατσκευή κειμένων με σύγχρονα γλωσσικά μοντέλα όπως το T5, Pegasus, BART που δημιουργούν παραφρασμένες και βελτιωμένες εκδοχές των αρχικών προτάσεων και κειμένων. Με τις ενσωματώσεις λέξεων και τις βαθμολογίες συνιμητόνου αξιολογούμε και στη συνέχεια με άλλα εργαλεία οπτικοποιούμε πόσο παρόμοια είναι τα ανακατασκευασμένα κείμενα με τα πρωτότυπα.

Με το NLP η διαδικασία γίνεται πιο αποτελεσματική και επεκτάσιμη χωρίς να υπάρχει ανάγκη για χειροκίνητη παρέμβαση.

Github: https://github.com/kp267/p22146_ex

ΜΕΘΟΔΟΛΟΓΙΑ

Η στρατηγικές ανακατασκευής για τα A, B, C:

A: Σε αυτό το ερώτημα έγινε ανακατασκευή μεμονομένων προτάσεων χειροκίνητα με συνθήκες if. Δηλαδή αν μέσα σε μια πρόταση εντωπυζόταν μια λέξη ή σειρά λέξεων τότε η ρπόταση αλλάζει έτσι ώστε να είναι γραμματικά και συντακτικά σωστή διατηρώντας το αρχικό νόημα.

B: Σε αυτό το ερώτημα χρησιμοποιήθηκαν αυτόματα piplines NLP, πιο συγκεκριμένα: T5, Pegasus, BART καιτο κείμενο βελτιώθηκε τόσο γραμματικά όσο και συντακτικά σε έναν αποδεκτό βαθμό διατηρώντας και πάλι το αρχικό του νόημα.

C: Εδώ εφαρμόστηκαν τεχνικές σύγκρισης για να αξιολογήσουμε την σημασιολογική ομοιότητα μεταξύ των αρχικών κειμένων και των ανακατασκευασμένων τους εκδοχών.

Ανάλυση υπολογιστικών τεχνικών:

Sentence embeddings:

Χρησιμοποιήθηκε το μοντέλο all-MiniLM-L6-v2 για να μετατρέψουμε κάθε πρόταση σε διάνυσμα χαρακτηριστικών που αποτυπώνει την σημασιολογική τους πληροφορία και έτσι μπορούμε στη συνέχεια να συγκρίνουμε αυτά τα διανύσματα μεταξύ τους για νε δούμε πόσο παρόμοιες είναι δυο προτάσεις νοηματικά.

Cosine similarity:

Για να υπολογιστεί η ομοιότητα που αναφέρθηκε παραπάνω χρησιμοποιήθηκε η συνάρτηση συνημιτόνου (cosine similarity). Η οποία δίνει τιμές από -1 έως 1 και όσο πιο κοντά στο 1 είναι η τιμή τόσο πιο παρόμοιες είναι οι προτάσεις νοηματικά.

Word embeddings:

Με τη χρήση του προεκπαιδευμένου μοντέλου glove-wiki-gigaword-100 κάθε λέξη των κειμένων μετατρέπεται σε διάνυσμα. Για κάθε πρόταση υπολογίστηκε ο μέσος όρος των διανυσμάτων των λέξεων της και πάλι με τη χρήση του cosine similarity συγκρίθηκε με την αντίστοιχη ανακατασκευασμένη πρότασή της.

Οπτικοποιήση:

Χρησιμοποιήθηκε το PCA(Principal Component Analysis) για γραμμική μείωση σε 2D

Και το t-SNE(t-distribute Stochastic Neighbor Embedding) για μη γραμμική μείωση

Ανάλυση λεξιλογικών διαφορών με TF-IDF:

Χρησιμοποιώντας το TfidVectorizer υπολογίστηκαν οι λέξεις που διαφοροποιούνται πιο σημαντικά μεταξύ στο αρχικό κείμενο και το ανακατασκευασμένο.

ΠΕΙΡΑΜΑΤΑ & ΑΠΟΤΕΛΕΣΜΑΤΑ

Παραδείγματα πριν και μετά την ανακατασκευή:

Text 1:

Original Text 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

Reocnstructed text 1 using T5:

Today is our dragon boat festival in our Chinese culture, to celebrate it with all safe and great in our lives. I hope you too to enjoy it as my deepest wishes. Thank you for 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 received the message from the professor a couple of days ago to show me this. I am very grateful for the full support of the professor for our Springer proceedings publication

Reconstructed text 1 using Pegasus:

Our Chinese culture has a dragon boat festival that we celebrate today. I hope you enjoy it as much as I do. Thank you for your message and for showing it to the doctor. I received this message to see the approved one. The professor sent me a message a few days ago, to show me. The professor supported the Springer proceedings publication.

Reconstructed text 1 using BART:

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.

Text 2:

Original Text 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

Reocnstructed text 2 using T5:

During our final discussion, I told him about the new submission — the one we were waiting for since last autumn , but the updates were confusing as it did not include the full feedback from reviewer or maybe editor ? Anyway, I believe that the team, although a bit delayed and less communication in recent days, really tried the best for paper and cooperation . We should be thankful, I mean all of us, for the acceptance and efforts until the Springer link finally came last week, I think . Please also remind me if the doctor still plan for the acknowledgments section edit before he sends again . Because I didn't see that part final yet, or maybe I missed it, I apologize if so . Let us make sure that all are safe and celebrate the outcome with strong coffee and future targets .

Reocnstructed text 2 using Pegasus:

I told him the new submission we were waiting for, but the updates were confusing because it didn't include the full feedback from the reviewer or editor. The team is trying best for paper and cooperation despite some delay and less communication recently. I think we should all be grateful, for the acceptance and the efforts, until the Springer link

came last week. Please remind me if the doctor still plans to edit the acknowledgments section before sending again. I apologize if I missed the part final. Let us make sure all are safe and celebrate the outcome with coffee and targets.

Reocnstructed text 2 using BART:

"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 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 of the Springer link. Also, kindly remind me please, if the doctor still plan for the acknowledgments section edit before 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.

Πλήρης αναφορά και ανάλυση του παραδοτέου 2:

Απαραίτητες βιβλιοθήκες:

pip install transformers

pip install sentencepiece

pip install sentence-transformers

pip install scikit-learn

pip install matplotlib

pip install nltk

pip install gensim

Επίσης πρέπει να τρέξει και μια φορά ο κώδικας nltk.download('stopwords').

Output μετά την εκτέλεση του 2:

Word Embedding Similarity (GloVe)

Text 1 vs T5: 0.9995709

Text 1 vs Pegasus: 0.96964544

Text 1 vs Bart: 1.0

Text 2 vs T5: 0.9991656

Text 2 vs Pegasus: 0.992698

Text 2 vs Bart: 0.99924093

Top changed words for Text 1 with T5:

for - 0.135

will - 0.095

grateful - 0.095

that - 0.095

you - 0.067

your - -0.0

dragon - -0.0

his - -0.0

have - -0.0

great - -0.0

Top changed words for Text 1 with Pegasus:

pad - 0.596

able - 0.099

are - 0.099

letting - 0.099

supported - 0.099

celebrated - 0.099

be - 0.099

know - 0.099

wish - 0.099

about - 0.099

Top changed words for Text 1 with Bart:

your - 0.0

in - 0.0

his - 0.0

have - 0.0

great - 0.0

got - 0.0

full - 0.0

from - 0.0

for - 0.0

festival - 0.0

Top changed words for Text 2 with T5:

include - 0.088

in - 0.088

discussion - 0.088

did - 0.088

delayed - 0.088

sends - 0.088

for - 0.067

that - 0.064

were - 0.064

think - 0.064

Top changed words for Text 2 with Pegasus:

pad - 0.498

despite - 0.083

good - 0.083

in - 0.083

plans - 0.083

thankful - 0.083

sends - 0.083

latest - 0.083

editing - 0.083

on - 0.083

Top changed words for Text 2 with Bart:

said - 0.089

the - 0.017

and - 0.008

for - 0.005

us - 0.003

or - 0.003

if - 0.003

last - 0.003

all - 0.003

we - 0.003

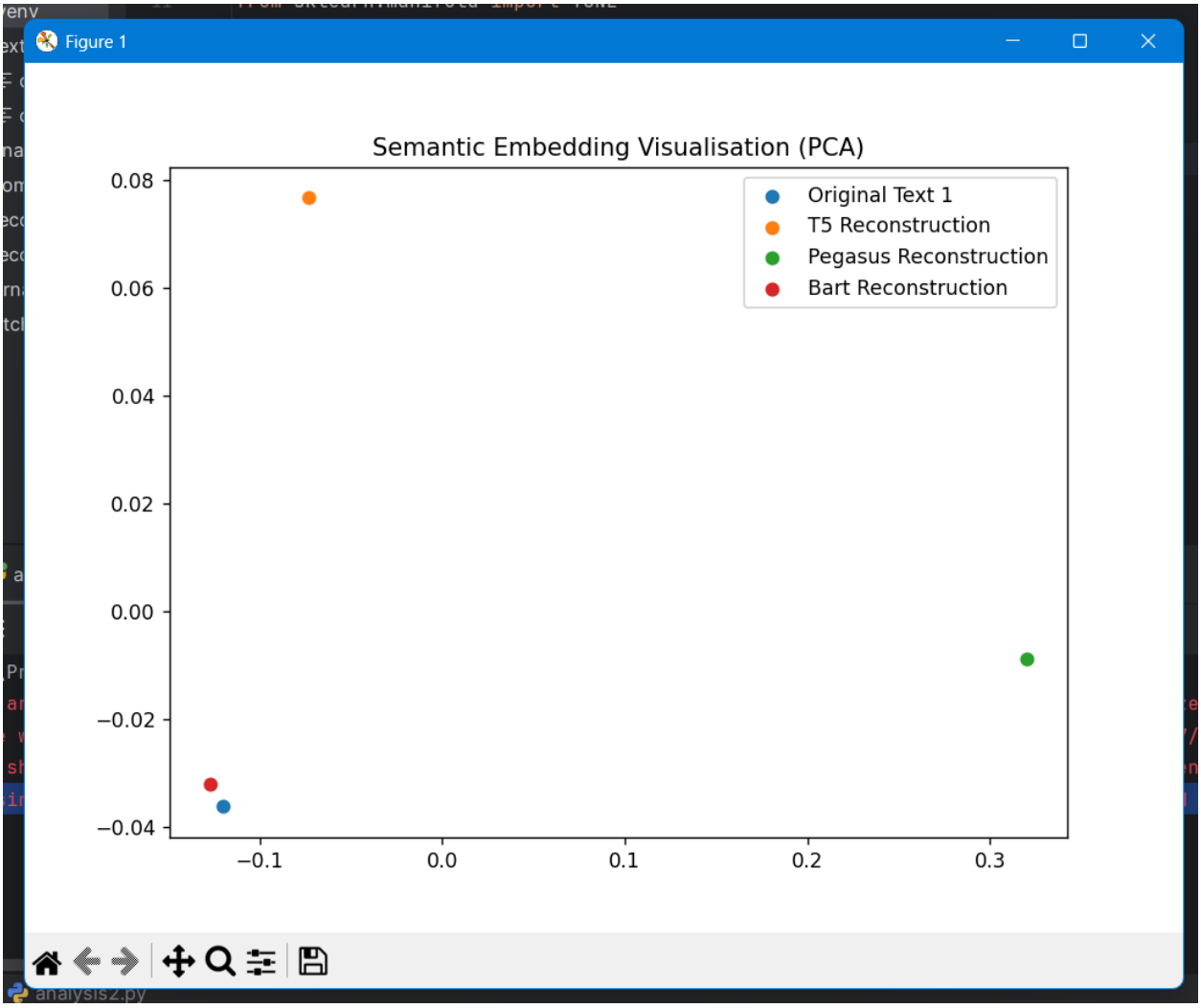


Figure 1
Semantic Embedding Visualisation (PCA)
Original Text 1
T5 Reconstruction
Pegasus Reconstruction
Bart Reconstruction
0.08
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.1
0.0
0.1
0.2
0.3
analysis2.py

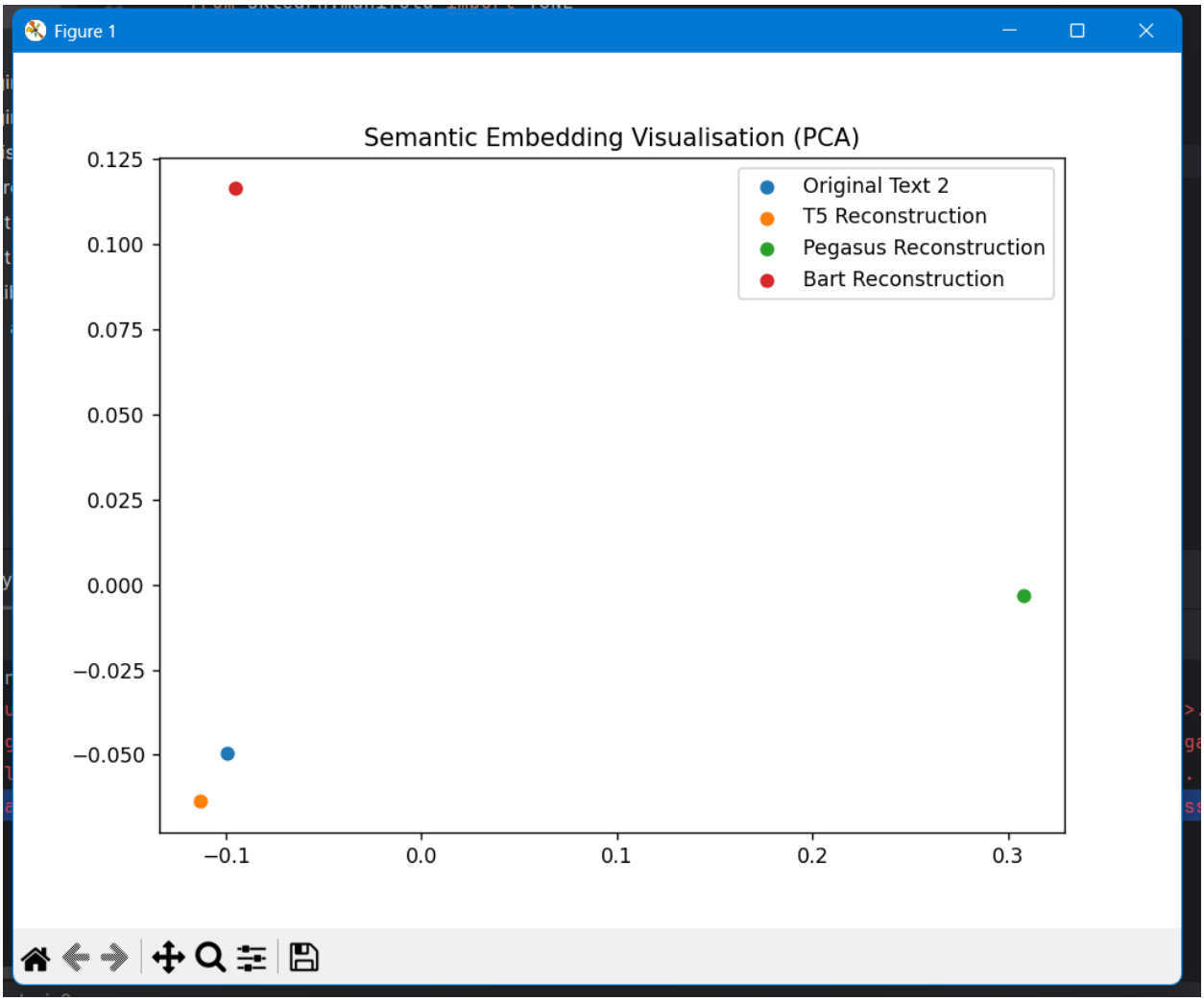


Figure 1
Semantic Embedding Visualisation (PCA)
Original Text 2
T5 Reconstruction
Pegasus Reconstruction
Bart Reconstruction
0.125
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.1
0.0
0.1
0.2
0.3

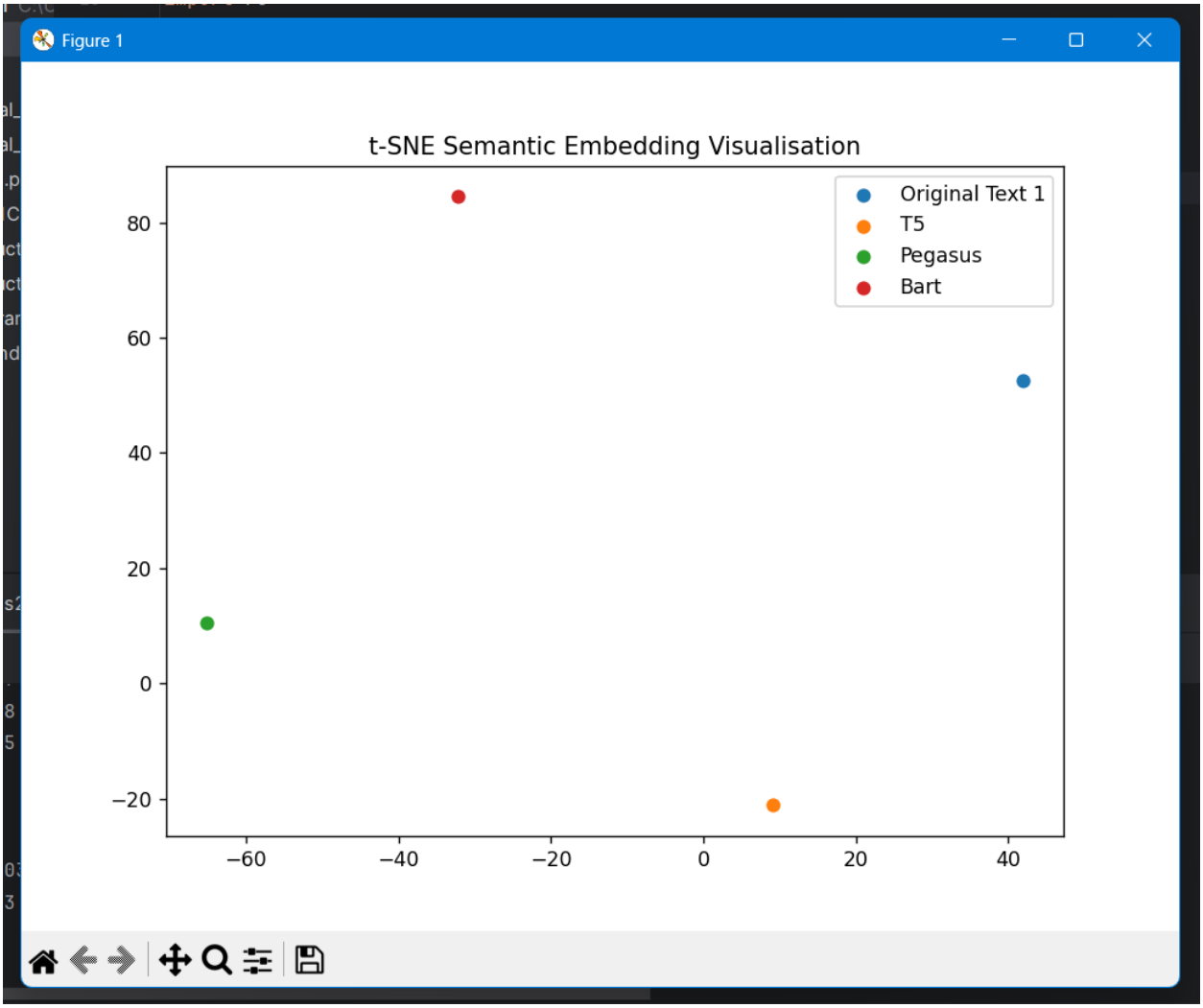
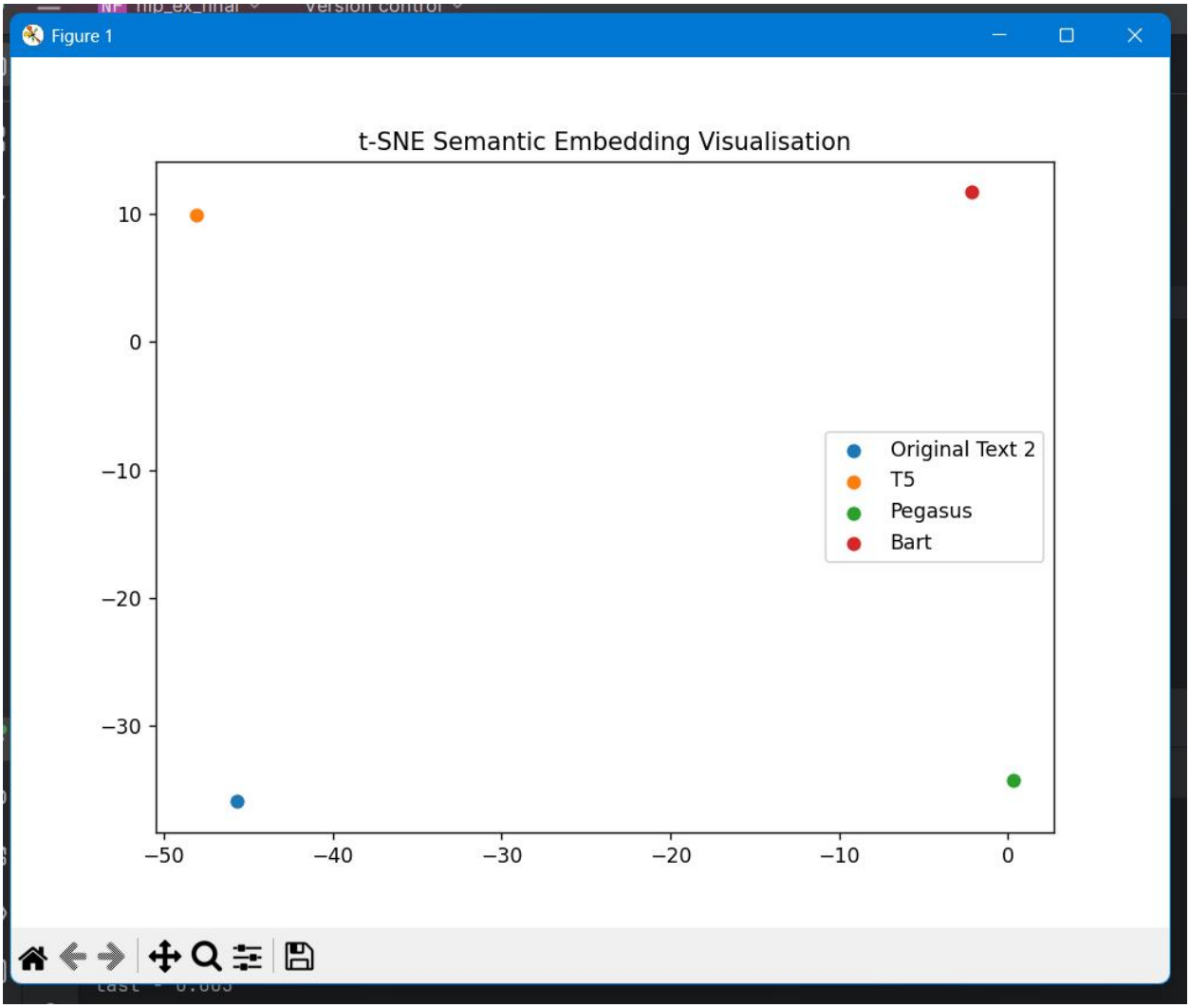


Figure 1
t-SNE Semantic Embedding Visualisation
Original Text 1
T5
Pegasus
Bart
80
60
40
20
0
−20
−60
−40
−20
0
20
40



Με τα γραφήματα βλέπουμε την σύγκριση των χωρικών αναπαραστάσεων των κειμένων και για κάθε ανακατασκευή παρουσιάζονται οι λέξεις που διαφοροποιήθηκαν περισσότερο με βάσει το TF-IDF score.

Ανάλυση ομοιότητας:

Κάθε κείμενο «καθαρίστηκε» μετατρεποντάς το σε πεζό, αφαιρώντας τα σημεία στίξης και τα stopwords(nltk). Στη συνέχεια με τη χρήση του προεκπαιδευμένου λεξικού glove-wiki-gigaword-100, υπολογίζοντας τα μέσα διανύσματα όρων και υπολογίζοντας το cosine similarity μεταξύ των διανυσμάτων υπολογίσαμε τη λεξιλογική ομοιότητα ανάμεσα στα αρχικά και στα ανακατασκευασμένα κείμενα.

Οπτικοποίηση:

Για να εξετάσουμε το πως τα διαφορετικά μοντέλα χειρίζονται τα κείμενα σημασιολογικά, χρησιμοποιήσαμε το μοντέλο all-MiniLM-L6-v2 ώστε να δημιουργηθούν sentence embeddings. Στη συνέχεια με τη χρήση του PCA και του t-SNE αναπαρηστήθηκαν τα αποτελέσματα.

Λεξιλόγιο:

Με τη συνάρτηση show_top_words() εντοπίσθηκαν και εκτυπώθηκαν οι 10 πιο διαφορετικές λέξεις μεταξύ του αρχικού και του ανακατασκευασμένου κειμένου σύμφωνα με την διαφορά των τιμών TF-IDF. Αυτό αποκαλύπτει ποια κομμάτια του λεξιλογίου αλλάζουν από μοντέλο σε μοντέλο και σε ποιο βαθμό.

Συμπέρασμα:

Όλα τα μοντέλα διατήρησαν με σχετική ακρίβεια το νόημα των αρχικών κειμένων. Η οπτικοποίηση επιβεβαίωσε τη συνοχή μεταξύ των προτότυπων κειμένων και των ανακατασκευασμένων τους. Η T5 διατήρησε το αρχικό νόημα χωρίς μεγάλες αποκλίσεις. Η Pegasus παρουσίασε ελαφρώς μεγαλύτερη λεξιλογική και σημασιολογική απόκλιση. Και η BART είχε την μεγαλύτερη πιστότητα ως προς τα αρχικά κείμενα με το GloVe similarity.

ΣΥΖΗΤΗΣΗ

Πόσο καλά αποτύπωσαν οι ενσωματώσεις λέξεων το νόημα;

Οι ενσωματώσεις λέξεων GloVe και προτάσεων SentenceTransformer αποδείωθηκαν πολύ χρήσιμες για την σημασιολογική αξιολόγηση των ανακατασκευάσεων. Χρησιμοποιώντας το cosine similarity αντικατοπτρίστηκαν με ακρίβεια οι σημασιολογικές αποστάσεις μεταξύ των αρχικών και των ανακατασκευασμένων κειμένων. Επιπλέον, τα GloVe embeddings ήταν χρήσιμα στην εντόπηση απώλειας πληροφορίας όταν άλλαζε πολύ το λεξιλόγιο.

Ποιες ήταν οι μεγαλύτερες προκλήσεις στην ανακατασεκυή;

Αρχικά, μια από τις μεγαλύτερες προκλήσεις ήταν η εύρεση των κατάλληλων βιβλιοθηκών έτσι ώστε τα ανκατασκευασμένα κείμενα να είναι γραμματικά και συντακτικά σωστά χώρις να χάνουν το αρχικό νόημα. Τα μοντέλα που χρησιμοποιήθηκαν τελικά (T5, Pegasus, BART) ήταν τα πιο αποδοτικά. Αλλά ακόμα και αυτά δεν ήταν αψεγάδιαστα. Για παράδειγμα το Pegasus είχε την τάση να προσθέτει κάποιες πληροφορίες που δεν υπήρχαν στο αρχικό κείμενο ενώ το T5 σε κάποιες περιπτώσεις ήταν υπερβολικά απλοϊκό στο λεξιλόγιο και το νόημα. Κάποια αποτελέσματα περιήχαν tokens (<pad>, </s>) τα οποία αφαιρέθηκαν με τη χρήση της skip_special_tokens=True. Ακόμα, τα μοντέλα BART, Pegasus απαιτούν αρκετούς υπολογιστικούς πόρους κάτι που καθιστά το πρόγραμμα πιο αργό.

Πώς μπορεί να αυτοματοποιηθεί αυτή η διαδικασία χρησιμοποιώντας μοντέλα NLP;

Η διαδικασία μπορεί να αυτοματοποιηθεί με ένα pipeline που περιλαμβάνει:

- Ανάγνωση και προεπξεργασία των κειμένων
- Ανακατασκευή από ήδη προκαθορισμένα μοντέλα όπως T5, Pegasus, BART
- Αξιολόγηση με cosine similarity, TF-IDF
- Οπτικοποίηση με PCA, t-SNE για την σύγκριση embeddings

Υπήρξαν διαφορές στην ποιότητα ανακατασκευής μεταξύ τεχνικών, μεθόδων, βιβλιοθηκών κλπ;
Ναι υπήρξαν διαφορές.
Αρχικά όσο αναφορά τα μοντέλα ανακατασκευής:

- Το μοντέλο T5 είχε σταθερό λεξιλόγιο και καλή διατήρηση πληροφορίας, ήταν πιστό στο αρχικό νόημα αλλά όχι πολύ δημιουργικό.

- Το μοντέλο Pegasus κάποιες φορές απομακρυνόταν από την αρχική σημασία και έχανε λίγο το νόημα αλλά ήταν πολύ δημιοργικό.
- Το μοντέλο BART ήταν δημιουργικό, διατηρούσε το αρχικό νόημα και γενικά μάλλον ήταν το καλύτερο από τα τρια μοντέλα.

Επίσης φαίνεται ότι η ανακατασκευή του δεύτερου κειμένου ήταν καλύτερη από του πρώτου και με τα τρία μοντέλα.

Ακόμα, το μοντέλο all-MiniLM-L6-v2 του SentenceTransformer αποδείχθηκε γρήγορο για semantic encoding και τα GloVe embeddings βοήθησαν στην επεξηγηματική ανάλυση.

ΣΥΜΠΕΡΑΣΜΑ

Η εργασία ανέδειξε τις δυνατότητες και τους περιορισμούς των παραφραστικών μοντέλων στη διαδικασία ανακατασκευής κειμένων όταν πρέπει να διατηρήσουν το αρχικό νόημα. Από την εκτέλεση των ερωτημάτων προέκυψαν τα παρακάτω αποτελέσματα:

- Τα μοντέλα T5, Pegasus, BART ήταν γενικά αποτελεσματικά στην δημιουργία ανακατασκευασμένων κειμένων που σε σημαντικό βαθμό διατηρούσαν το αρχικό νόημα.
- Υπήρχαν ωστόσο αρκετές προκλήσεις. Αρχικά, η απώλεια πληροφορίας ,ιδιαίτερα από το Pegasus. Κάποιες φορές εκέι που θα έπρεπε να εισαχθεί τελεία (.) αυτό δεν γινόταν. Επίσης υπήρχαν περιορισμοί σε μήκος εξόδου και έτσι μειωνόταν η πιστότητα στο αρχικό κείμενο. Τέλος, μια από τις μεγαλύτερες προκλήσεις ήταν η εύρεση των κατάλληλων και των καλύτερων pipelines και βιβλιοθηκών για την υλοποίηση της εργασίας καθώς έπρεπε να δοκιμαστούν πολλά διαφορετικά pipelines μέχρι να καταλήξουμε στα τρια που αναφέρθηκαν παραπάνω.
- Τα pipelines ήταν και τα 3 αποδοτικά, σε διαφορετικούς βαθμούς, ενώ είχε και το καθένα τα δικά του προβλήματα στην ανακατασκευή των δυο κειμένων.

Κλείνοντας, η εργασία ανέδειξε πόσο χρήσιμα μπορούν να γίνουν όλα τα εργαλεία που προσφέρει το NLP, αλλά και την ανάγκη για ανθρώπινη παρέμβαση σε συγκεκριμένα σημεία καθώς επίσης και προσεκτικό σχεδιασμό ώστε να επιτευχθεί το καλύτερο δυνατό αποτέλεσμα στην ανακατασκευή κειμένων.

ΒΙΒΛΙΟΓΡΑΦΙΑ

- Έγγραφα Gunet του μαθήματος
- Github nlp_lab_unipi
- Hugging Face Transformers
  https://huggingface.co/docs/transformers/index
- T5 model page
  https://huggingface.co/Vamsi/T5_Paraphrase_Paws
- Pegasus model page
  https://huggingface.co/tuner007/pegasus_paraphrase
- BART model page
  https://huggingface.co/facebook/bart-large-cnn
- SentenceTransformers
  https://www.sbert.net/
- Gensim data
  https://github.com/RaRe-Technologies/gensim-data
- GloVe
  https://nlp.stanford.edu/projects/glove/
- Cosine similarity, PCA, t-SNE, TF-IDF
  https://scikit-learn.org/stable/documentation.html