ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
Τμήμα Πληροφορικής

Εργασία Μαθήματος ***Επεξεργασία Φυσικής Γλώσσας***
Παραδοτέο 2

Αριθμός εργασίας – Τίτλος εργασίας	***Απαλλακτική***
Όνομα φοιτητή	ΜΑΡΙΟΣ ΚΥΡΟΓΛΟΥ
Αρ. Μητρώου	Π21080

Εκφώνηση εργασίας

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
B. Ανακατασκευή του συνόλου των 2 κειμένων με χρήση 3 διαφορετικών αυτόματων βιβλιοθηκών python pipelines
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.)*και δικές σας -custom- αυτόματες ροές εργασίας NLP (προεπεξεργασία, λεξιλόγιο, ενσωμάτωση λέξεων, εννοιολογικά δέντρα κλπ) 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 του παραδοτέου 1.

Οπτικοποιήστε τις ενσωματώσεις λέξεων για τα A,B 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*

Bonus - Masked Clause Input
Ακολουθώντας τις μεθοδολογίες που αναπτύξατε στα προηγούμενο ερωτήματα με τις απαραίτητες τροποποιήσεις, προσπαθήστε να εφαρμόστε μια αντίστοιχη προσέγγιση χρησιμοποιώντας μοντέλα open source για να επιλύσετε το παρακάτω πρόβλημα το οποίο εντάσετε στην κατηγορία **Masked Clause Input**. Να συμπληρωθούν οι λέξεις που λείπουν (στα ελληνικά) και να συγκρίνετε τα μοντέλα που θα χρησιμοποιήσετ μεταξύ τους αξιολογώντας τα σε σχέση με την ικανότητα τους να προσδίδουν νόημα σε σχέση με τα αντίστοιχα άρθρα του αστικού κώδικα (διαθέσιμα online) - τα οποία θα θεωρήσετε ως ground truth-premise για τις συγκρίσεις σας. Για την κατανόηση του προβλήματος χρησιμοποιήστε ένα συντακικό αναλυτή (nltk ή αντίστοιχο) και εξαγάγετε συμπεράσματα ως προς τις ελλείψεις των μοντέλων που έχετε χρησιμοποιήσει. **Για όσους ασχοληθούν με το bonus θα το εντάξουν στο παραδοτέο κείμενο 3 - Δομημένη Αναφορά.**

ΣΥΓΚΥΡΙΟΤΗΤΑ
Άρθρο 1113. Κοινό πράγμα. — Αν η κυριότητα του [MASK] ανήκει σε περισσότερους [MASK] αδιαιρέτου κατ'ιδανικά [MASK], εφαρμόζονται οι διατάξεις για την κοινωνία.

Άρθρο 1114. Πραγματική δουλεία σε [MASK] η υπέρ του κοινού ακινήτου. — Στο κοινό [MASK] μπορεί να συσταθεί πραγματική δουλεία υπέρ του [MASK] κύριου άλλου ακινήτου και αν ακόμη αυτός είναι [MASK] του ακινήτου που βαρύνεται με τη δουλεία. Το ίδιο ισχύει και για την [MASK] δουλεία πάνω σε ακίνητο υπέρ των εκάστοτε κυρίων κοινού ακινήτου, αν [MASK] από αυτούς είναι κύριος του [MASK] που βαρύνεται με τη δουλεία.

παράδειγμα
https://huggingface.co/google-bert/bert-base-uncased?text=i+am+[MASK]+a+nice+time

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

1 Εισαγωγή

Στο δεύτερο παραδοτέο της εργασίας εξετάστηκε η χρήση τεχνικών παραφραστικής αναδιατύπωσης δύο κειμένων μέσω δύο διαφορετικών μεθόδων: (α) μια βασική τοπική μέθοδο (Method A) και (β) τη χρήση μεγάλου γλωσσικού μοντέλου GPT. Εφαρμόστηκαν δύο διαφορετικοί τύποι μοντέλων ενσωμάτωσης (SpaCy και BERT) για την εκτίμηση της ομοιότητας.

Η ανάπτυξη πραγματοποιήθηκε σε περιβάλλον Python με χρήση των βιβλιοθηκών: spaCy, transformers, sentence-transformers, torch, numpy, nltk. Η εκτέλεση έγινε σε περιβάλλον Anaconda με ξεχωριστό περιβάλλον (env).

2 Περιγραφή Υλοποίησης

Η υλοποίηση πραγματοποιήθηκε στο αρχείο analysis.py. Το πρόγραμμα αναλύει δύο ζεύγη κειμένων: ένα πρωτότυπο και το παραφρασμένο μέσω δύο μεθόδων (A και GPT). Οι μετρικές που μετρούνται είναι:
Ομοιότητα μεταξύ αρχικού και παραφρασμένου κειμένου (doc-level cosine similarity).
Μέση λέξη προς λέξη ομοιότητα.
Διαφορά μέσων διανυσμάτων (vector difference).
Για κάθε μοντέλο (SpaCy και BERT) παράγονται οι τιμές για τα δύο κείμενα (TEXT1 και TEXT2). Επίσης, γίνεται παράθεση ενδεικτικών συγκρίσεων λέξεων.
Η μέθοδος A χρησιμοποιεί στατική αντικατάσταση συνωνύμων, ενώ η μέθοδος GPT βασίζεται σε αυτόματη δημιουργία παραφράσεων από το προεκπαιδευμένο μοντέλο γλώσσας.
Οι τελικές τιμές αποθηκεύονται στο analysis_results.json και προβάλλονται στον χρήστη μέσω εκτυπώσεων στην κονσόλα.

3 Συγκριτική Ανάλυση Αποτελεσμάτων

SpaCy Embeddings

- **TEXT1**: Η μέθοδος A είχε υψηλότερη ομοιότητα εγγράφου (0.82) έναντι της GPT (0.71), αλλά η GPT είχε τέλεια αντιστοιχία στις λέξεις (1.00). Το διανυσματικό σφάλμα της GPT ήταν μεγαλύτερο (22.1 έναντι 13.4).

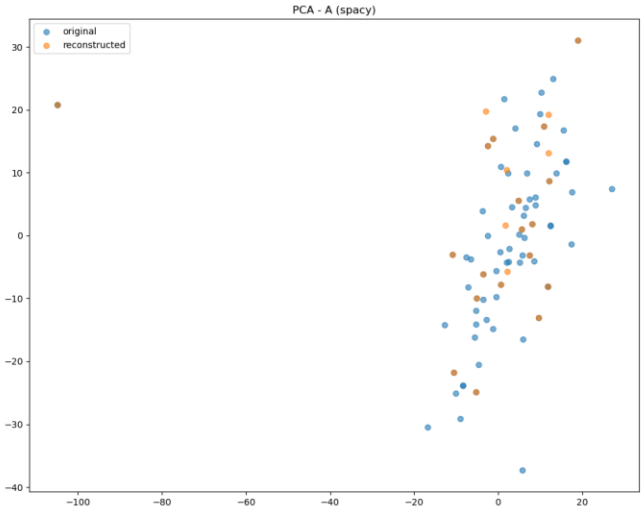
```
C:\Users\Mario\anaconda3\envs\THEMA_2\python.exe "C:\Users\Mario\Desktop\EPEXERGASIAFYSIKHSGLWSSAS\THEMA 2\analysis.py"
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\Mario\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\Mario\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!

========= MODEL: SPACY =========
Device set to use cpu
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.

--- ANALYSIS FOR TEXT1 ---

Method          Doc Similarity  Word Similarity Vector Diff
A               0.8238        0.5899        13.3970
GPT             0.7121        1.0000        22.1576

Example word comparisons:

Method A:
  'today' → 'today' (score: 1.00)
  'dragon' → 'dragon' (score: 1.00)
  'boat' → 'boat' (score: 1.00)
```

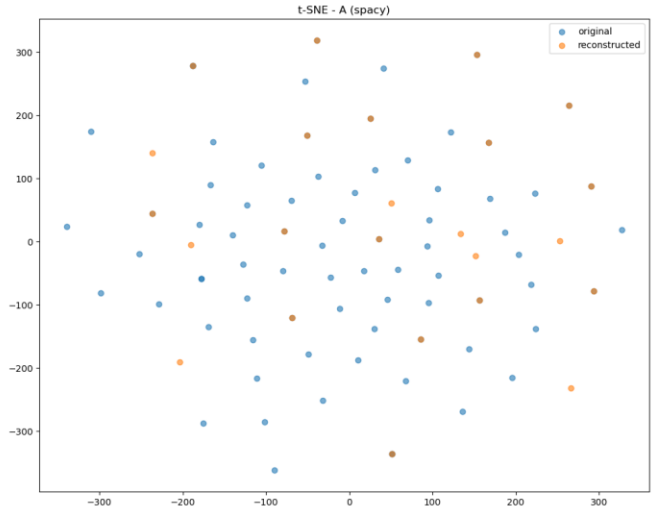
```
Method GPT:
  'today' → 'today' (score: 1.00)
  'dragon' → 'dragon' (score: 1.00)
  'boat' → 'boat' (score: 1.00)
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
```

- **TEXT2**: Η GPT ξεπέρασε τη μέθοδο A τόσο σε ομοιότητα εγγράφου (0.97 έναντι 0.94), όσο και σε word similarity (1.00 έναντι 0.64) και vector difference (μικρότερη τιμή σημαίνει μεγαλύτερη ομοιότητα).

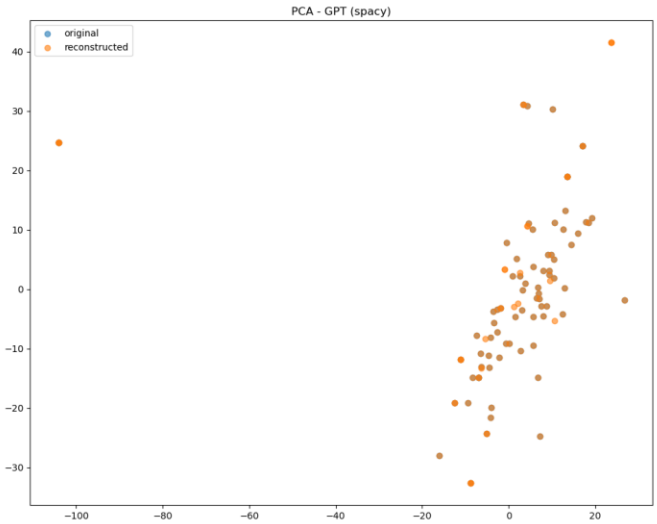
```
--- ANALYSIS FOR TEXT2 ---

Method          Doc Similarity  Word Similarity Vector Diff
A               0.9474          0.6438          8.2351
GPT             0.9710          1.0000          6.8891

Example word comparisons:

Method A:
  'final' → 'last' (score: 0.56)
  'discuss' → 'acknowledgments' (score: 0.56)
  'told' → 'think' (score: 0.35)

Method GPT:
  'final' → 'final' (score: 1.00)
  'discuss' → 'discuss' (score: 1.00)
  'told' → 'told' (score: 1.00)
```

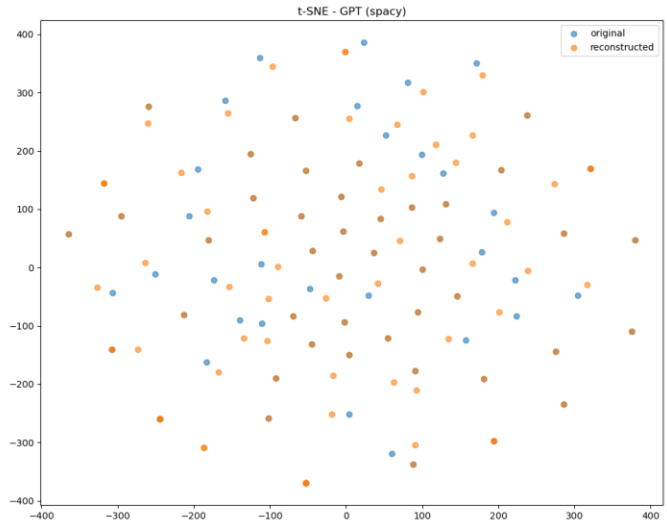
Embeddings_A_SpaCy:





Embeddings_GPT_Spacy:





BERT Embeddings

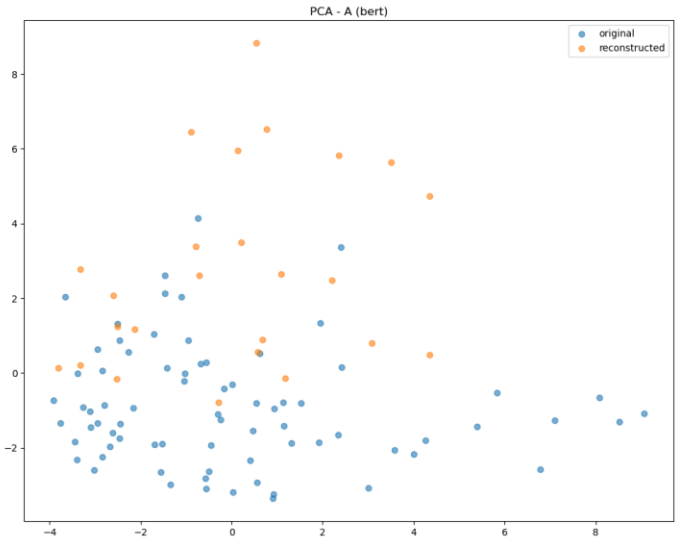
- **TEXT1**: Πιο ισορροπημένα αποτελέσματα. Η μέθοδος A έδωσε υψηλότερη doc similarity (0.84), αλλά η GPT υπερείχε σε word similarity (0.94).

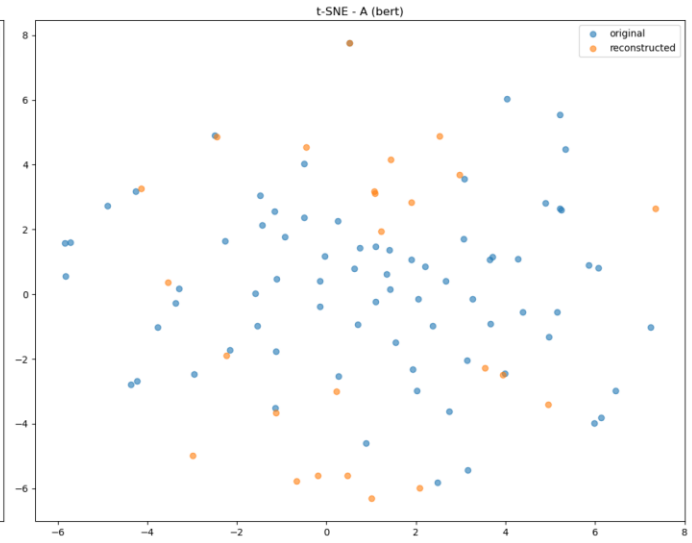
```
========= MODEL: BERT =========
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.

--- ANALYSIS FOR TEXT1 ---

Method          Doc Similarity  Word Similarity Vector Diff
A               0.8408          0.6115          4.8925
GPT             0.8097          0.9445          5.7447

Example word comparisons:

Method A:
  'today' → 'today' (score: 0.74)
  'dragon' → 'dragon' (score: 0.90)
  'boat' → 'boat' (score: 0.84)

Method GPT:
  'today' → 'today' (score: 0.95)
  'dragon' → 'dragon' (score: 0.98)
  'boat' → 'boat' (score: 0.94)
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
```

- **TEXT2**: Η GPT και πάλι υπερέχει συνολικά με υψηλότερες τιμές ομοιότητας και μικρότερη διαφορά διανυσμάτων.

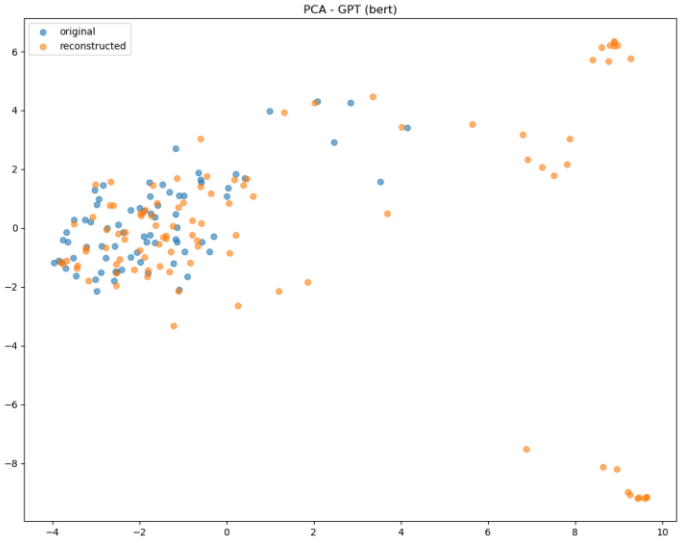
```
--- ANALYSIS FOR TEXT2 ---

Method          Doc Similarity  Word Similarity Vector Diff
A               0.8722          0.6537          4.5975
GPT             0.9005          0.9616          4.0187

Example word comparisons:

Method A:
  'final' → 'finally' (score: 0.61)
  'discuss' → 'springer' (score: 0.65)
  'told' → 'springer' (score: 0.66)

Method GPT:
  'final' → 'final' (score: 0.96)
  'discuss' → 'discuss' (score: 0.99)
  'told' → 'told' (score: 0.98)

Process finished with exit code 0
```

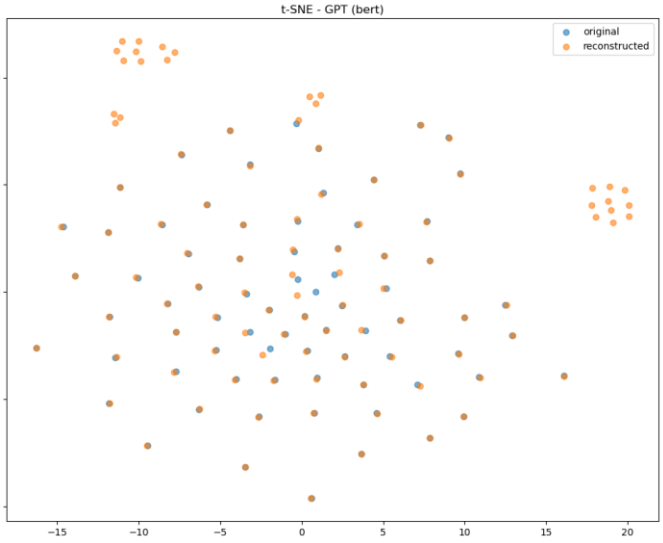
Embeddings_A_Bert:





Embeddings_GPT_Bert:





Γενικά, η GPT διατηρεί καλύτερη συνοχή στις λέξεις, ενώ η μέθοδος A εμφανίζει υψηλότερη συνολική ομοιότητα όταν οι λέξεις διατηρούνται σταθερές, αλλά μειονεκτεί όταν απαιτείται ρεαλιστική και νοηματική παραφράση.

Συμπεράσματα

Η χρήση προχωρημένων γλωσσικών μοντέλων όπως το GPT οδηγεί σε παραφράσεις με υψηλή πιστότητα νοήματος και μεγάλη λεκτική ομοιότητα. Ωστόσο, μπορεί να υπάρχουν διαφοροποιήσεις στο ύφος ή στη σύνταξη που οδηγούν σε μικρότερη doc-level ομοιότητα. Αντιθέτως, οι πιο «μηχανικές» μέθοδοι αντικατάστασης όπως η μέθοδος A έχουν προβλέψιμη συμπεριφορά αλλά μειωμένη φυσικότητα.

Ο συνδυασμός διαφορετικών embeddings (SpaCy και BERT) προσφέρει χρήσιμα insights: τα SpaCy vectors είναι λιγότερο ευαίσθητα, ενώ τα BERT embeddings συλλαμβάνουν λεπτές σημασιολογικές αποχρώσεις καλύτερα.

Η υλοποίηση είναι επεκτάσιμη και μπορεί να προσαρμοστεί για περαιτέρω ανάλυση με χρήση BLEU, ROUGE ή άλλων NLP metrics.

4 Πηγές

Spacy: https://spacy.io/

Word2Vec: https://www.tensorflow.org/text/tutorials/word2vec

NLTK: https://www.nltk.org/

TextBlob: https://textblob.readthedocs.io/en/dev/