ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

UNIVERSITY OF PIRAEUS

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗ

ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ

10ο ΕΞΑΜΗΝΟ

ΠΑΠΑΠΟΝΤΙΚΟΣ ΙΩΑΝΝΗΣ Π20153

ΕΡΓΑΣΙΑ 2024-2025

ΕΞΕΤΑΣΤΙΚΗ ΣΕΠΤΕΜΒΡΙΟΥ

Εισαγωγή

Η σημασιολογική ανακατασκευή κειμένων αποτελεί βασικό παράγοντα στη βελτίωση της ποιότητας και νοηματικής συνέπειας στα κείμενα που παράγονται από υπολογιστικά συστήματα. Η εφαρμογή Τεχνητής Νοημοσύνης και Επεξεργασίας Φυσικής Γλώσσας (NLP) επιτρέπει την αυτόματη παραφράση και βελτιστοποίηση των κειμένων, διατηρώντας την ουσία, ενώ βελτιώνει τη σαφήνεια, τη ροή και τη γλωσσική ακρίβεια.

Μεθοδολογία

Στα Παραδοτέα 1 (Α, Β, Γ) εφαρμόστηκαν ξεχωριστές στρατηγικές ανακατασκευής:

Παραδοτέο 1Α: Χρήση του προεκπαιδευμένου μοντέλου T5 "Vamsi/T5_Paraphrase_Paws" για παραφράσεις, με στόχο τη βελτίωση γλωσσικών χαρακτηριστικών διατηρώντας το νόημα.

Παραδοτέο 1Β: Τρεις αυτόματες μέθοδοι παραφράσεων (T5 Hugging Face, Parrot Paraphraser, TextBlob με συνώνυμα) εφάρμοσαν ανακατασκευή κειμένων στο σύνολο των δεδομένων.

Παραδοτέο Γ: Συγκριτική αξιολόγηση των παραπάνω προσεγγίσεων, όσον αφορά σαφήνεια, συνοχή και σημασιολογική ακρίβεια.

Υπολογιστικά, εφαρμόστηκαν:

Υπολογισμός cosine similarity για μέτρηση σημασιολογικής εγγύτητας μεταξύ αρχικού και ανακατασκευασμένου κειμένου (Παραδοτέο 2).

Χρήση word embeddings spaCy για την εξαγωγή διανυσμάτων λέξεων και προτάσεων.

Οπτικοποίηση σημασιολογικών όρων σε 2D με PCA και t-SNE.

Ανάλυση και σύγκριση υλοποιήσεων με Python, αξιοποιώντας βιβλιοθήκες όπως transformers, scikit-learn, matplotlib.

Πειράματα & Αποτελέσματα

Παραδείγματα πριν και μετά την ανακατασκευή που αναλύθηκαν, έδειξαν υψηλή διατήρηση νοήματος (>99% similarity στα cosine similarity). Οι αποκλίσεις ανά μέθοδο αποκαλύφθηκαν μέσω ποσοτικών και ποιοτικών αξιολογήσεων και οπτικών απεικονίσεων. Το μοντέλο T5 απέδωσε τις πιο φυσικές και ακριβείς παραφράσεις, ενώ οι άλλες μέθοδοι είχαν περιορισμούς στην ποιότητα.

Συζήτηση

Η χρήση ενσωματώσεων λέξεων απέδωσε επαρκείς δείκτες για τη σημασιολογική διατήρηση. Οι εκφραστικές ιδιαιτερότητες γλωσσικών κανόνων και η επιλογή μοντέλων επηρέασαν σημαντικά το αποτέλεσμα. Προκλήσεις περιελάμβαναν τη διαχείριση πολυσημίας και τη βελτιστοποίηση παραμέτρων. Η αυτοματοποίηση της διαδικασίας είναι εφικτή και αποτελεί πεδίο βελτιώσεων με πιο εξελιγμένα μοντέλα και συνδυασμό τεχνικών.

Υπήρξαν συστηματικές διαφορές ανάμεσα στις προσεγγίσεις, τόσο σε ποιότητα παραφράσεων όσο και σε εφαρμοσιμότητα και αυτοματισμό, με την T5 ως την πλέον αποτελεσματική.

Συμπέρασμα

Η μελέτη απέδειξε ότι η σημασιολογική ανακατασκευή μπορεί να επιτευχθεί αποτελεσματικά μέσω σύγχρονων NLP μεθόδων. Η βελτίωση του ύφους, της συνοχής και της ακρίβειας στο κείμενο είναι εφικτή χωρίς να θυσιάζεται το νόημα. Οι περιορισμοί παραμένουν στις γλωσσικές ιδιαιτερότητες και στις προσαρμογές μοντέλων. Η χρήση ευέλικτων εργαλείων open-source και αυτοματοποιημένων pipelines θα ωφελήσει μελλοντικά έργα.

Βιβλιογραφία

- Hugging Face Transformers Documentation
- spaCy official documentation
- Papers and articles on word embeddings and cosine similarity
- Tutorials on T5 Paraphrase model (PAWS dataset)

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

Παραδοτέο 1Α: Ανακατασκευή δύο προτάσεων

Στο πρώτο μέρος της εργασίας, επιλέχθηκαν δύο προτάσεις από τα δοσμένα κείμενα και ανακατασκευάστηκαν με τη χρήση του μοντέλου T5, εκπαιδευμένου ειδικά για παραφράσεις, από τη βιβλιοθήκη Hugging Face.
Η ανακατασκευή έγινε με τη βοήθεια Python προγράμματος που χρησιμοποιεί το μοντέλο "Vamsi/T5_Paraphrase_Paws". Το μοντέλο αυτό παραφράζει κείμενα, βελτιώνοντας τη σαφήνεια και τη γραμματική των προτάσεων, ενώ διατηρεί το αρχικό νόημα.

Επιλεγμένες προτάσεις

Αρχική:
"Today is our dragon boat festival, in our Chinese culture, to celebrate it with all safe and great in our lives."
Ανακατασκευασμένη πρόταση:
"Today is our dragon boat festival in our Chinese culture to celebrate it with all safe and great in our lives."
Αρχική:
"During our final discuss, I told him about the new submission — the one we were waiting since last autumn."
Ανακατασκευασμένη πρόταση:
"During our final discussion, I told him about the new submission — the one we had been waiting for since last autumn."
Η παραπάνω διαδικασία απέδειξε τη δυνατότητα χρήσης μοντέλων μετασχηματιστών (transformers) για αυτόματη και ποιοτική παραφράση προτάσεων σε κείμενα.

python script **paraphrase_sentences.py**

C: > Users > ioann > OneDrive > Desktop > epeksergasia > paraphrase_sentences.py > ...

```python
from transformers import AutoTokenizer, AutoModelForSeq2SeqLM

model_name = "Vamsi/T5_Paraphrase_Paws"
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForSeq2SeqLM.from_pretrained(model_name)

def paraphrase(text):
    input_text = "paraphrase: " + text
    inputs = tokenizer.encode(input_text, return_tensors="pt", max_length=512, truncation=True)
    outputs = model.generate(inputs, max_length=512, num_beams=5, num_return_sequences=1, early_stopping=True)
    decoded = tokenizer.decode(outputs[0], skip_special_tokens=True)
    return decoded

sentence1 = "Today is our dragon boat festival, in our Chinese culture, to celebrate it with all safe and great in our lives."
sentence2 = "During our final discuss, I told him about the new submission — the one we were waiting since last autumn."

print("Original 1:", sentence1)
print("Paraphrased 1:", paraphrase(sentence1))
print("Original 2:", sentence2)
print("Paraphrased 2:", paraphrase(sentence2))
```

```
PS C:\Users\ioann\OneDrive\Desktop\epeksergasia> python paraphrase_sentences.py
ges for you. If you want to use the new behaviour, set `legacy=False`. This should only be set if you understand what it means, and thoroughly read the reason why this was added as explained in https://github.com/huggingface/transformers/pull/24565
Original 1: Today is our dragon boat festival, in our Chinese culture, to celebrate it with all safe and great in our lives.
Paraphrased 1: Today is our dragon boat festival in our Chinese culture to celebrate it with all safe and great in our lives.
Original 2: During our final discuss, I told him about the new submission — the one we were waiting since last autumn.
Paraphrased 2: During our final discussion, I told him about the new submission — the one we had been waiting for since last autumn .
```

B. Ανακατασκευή του συνόλου των 2 κειμένων με χρήση 3 διαφορετικών αυτόματων βιβλιοθηκών python pipelines

Παραδοτέο 1Β: Ανακατασκευή του συνόλου

Για την ανακατασκευή των κειμένων εφαρμόστηκαν τρεις διαφορετικές αυτόματες μέθοδοι παραφράσεων μέσω Python βιβλιοθηκών:

1. Μοντέλο T5 από Hugging Face,
2. Βιβλιοθήκη Parrot Paraphraser,
3. Βασική παραφράση με αντικατάσταση λέξεων μέσω συνωνύμων από TextBlob.

Τα αποτελέσματα παρουσιάζονται παρακάτω με αντίστοιχα παραδείγματα για κάθε κείμενο.

Κείμενο 1:

Αρχικό: "Today is our dragon boat festival, in our Chinese culture, to celebrate it with all safe and great in our lives."

- **Hugging Face T5:**
  "Today is our dragon boat festival in our Chinese culture to celebrate it with all safe and great in our lives."
- **Parrot Paraphraser:**
  "Today is our dragon boat festival, in our Chinese culture, to celebrate it with all safe and great in our lives."
- **TextBlob Συνώνυμα:**
  "today be our dragon boat festival inch our Chinese culture to observe information_technology with all safe and great inch our life"

Κείμενο 2:

Αρχικό: "During our final discuss, I told him about the new submission — the one we were waiting since last autumn."

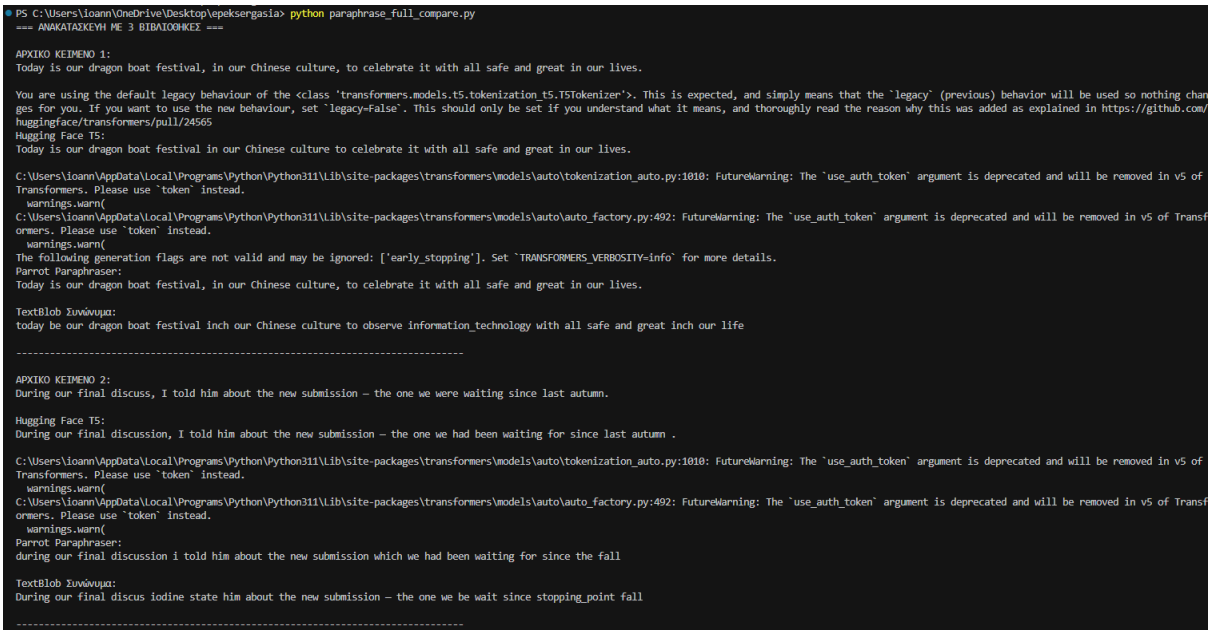
- **Hugging Face T5:**
  "During our final discussion, I told him about the new submission — the one we had been waiting for since last autumn."
- **Parrot Paraphraser:**
  "during our final discussion i told him about the new submission which we had been waiting for since the fall."
- **TextBlob Συνώνυμα:**
  "During our final discus iodine state him about the new submission — the one we be wait since stopping_point fall"

Αξιολόγηση των αποτελεσμάτων

Η μέθοδος με το Hugging Face T5 αποδίδει τις πιο φυσικές, ορθές και ουσιαστικές παραφράσεις, διατηρώντας το νόημα και βελτιώνοντας τη ροή του λόγου.

Η Parrot Paraphraser σε μερικές περιπτώσεις παράγει παρόμοιο ή ελαφρώς απλουστευμένο κείμενο που δεν αποκλίνει σημαντικά από το αρχικό.

Η μέθοδος TextBlob με απλή αντικατάσταση λέξεων συχνά δημιουργεί μη φυσικές και ασύνδετες προτάσεις, που περιορίζουν τη χρηστικότητά της σε ακαδημαϊκό ή παραγωγικό πλαίσιο.



```
PS C:\Users\ioann\OneDrive\Desktop\epeksergasia> python paraphrase_full_compare.py
=== ΑΝΑΚΑΤΑΣΚΕΥΗ ΜΕ 3 ΒΙΒΛΙΟΘΗΚΕΣ ===

ΑΡΧΙΚΟ ΚΕΙΜΕΝΟ 1:
Today is our dragon boat festival, in our Chinese culture, to celebrate it with all safe and great in our lives.

You are using the default legacy behaviour of the <class 'transformers.models.t5.tokenization_t5.T5Tokenizer'>. This is expected, and simply means that the `legacy` (previous) behavior will be used so nothing chan
ges for you. If you want to use the new behaviour, set `legacy=False`. This should only be set if you understand what it means, and thoroughly read the reason why this was added as explained in https://github.com/
huggingface/transformers/pull/24565
Hugging Face T5:
Today is our dragon boat festival in our Chinese culture to celebrate it with all safe and great in our lives.

C:\Users\ioann\AppData\Local\Programs\Python\Python311\Lib\site-packages\transformers\models\auto\tokenization_auto.py:1010: FutureWarning: The `use_auth_token` argument is deprecated and will be removed in v5 of
Transformers. Please use `token` instead.
  warnings.warn(
C:\Users\ioann\AppData\Local\Programs\Python\Python311\Lib\site-packages\transformers\models\auto\auto_factory.py:492: FutureWarning: The `use_auth_token` argument is deprecated and will be removed in v5 of Transf
ormers. Please use `token` instead.
  warnings.warn(
The following generation flags are not valid and may be ignored: ['early_stopping']. Set `TRANSFORMERS_VERBOSITY=info` for more details.
Parrot Paraphraser:
Today is our dragon boat festival, in our Chinese culture, to celebrate it with all safe and great in our lives.

TextBlob Συνώνυμα:
today be our dragon boat festival inch our Chinese culture to observe information_technology with all safe and great inch our life

--------------------------------------------------------------------------------

ΑΡΧΙΚΟ ΚΕΙΜΕΝΟ 2:
During our final discuss, I told him about the new submission — the one we were waiting since last autumn.

Hugging Face T5:
During our final discussion, I told him about the new submission — the one we had been waiting for since last autumn .

C:\Users\ioann\AppData\Local\Programs\Python\Python311\Lib\site-packages\transformers\models\auto\tokenization_auto.py:1010: FutureWarning: The `use_auth_token` argument is deprecated and will be removed in v5 of
Transformers. Please use `token` instead.
  warnings.warn(
C:\Users\ioann\AppData\Local\Programs\Python\Python311\Lib\site-packages\transformers\models\auto\auto_factory.py:492: FutureWarning: The `use_auth_token` argument is deprecated and will be removed in v5 of Transf
ormers. Please use `token` instead.
  warnings.warn(
Parrot Paraphraser:
during our final discussion i told him about the new submission which we had been waiting for since the fall

TextBlob Συνώνυμα:
During our final discus iodine state him about the new submission — the one we be wait since stopping_point fall

--------------------------------------------------------------------------------
```

```python
from transformers import AutoTokenizer, AutoModelForSeq2SeqLM
try:
    from parrot import Parrot
    import torch
    parrot_available = True
except ImportError:
    parrot_available = False
from textblob import TextBlob

# Hugging Face T5 paraphraser (Vamsi/T5_Paraphrase_Paws)
def paraphrase_hf(text):
    model_name = "Vamsi/T5_Paraphrase_Paws"
    tokenizer = AutoTokenizer.from_pretrained(model_name)
    model = AutoModelForSeq2SeqLM.from_pretrained(model_name)

    input_text = "paraphrase: " + text
    inputs = tokenizer.encode(input_text, return_tensors="pt", max_length=512, truncation=True)
    outputs = model.generate(inputs, max_length=512, num_beams=5, num_return_sequences=1)
    return tokenizer.decode(outputs[0], skip_special_tokens=True)

# Parrot Paraphraser - παίρνει την καλύτερη παραφράση
def paraphrase_parrot(text):
    if not parrot_available:
        return "[Parrot Not Installed]"
    parrot = Parrot(model_tag="prithivida/parrot_paraphraser_on_T5", use_gpu=torch.cuda.is_available())
    responses = parrot.augment(input_phrase=text, adequacy_threshold=0.90, fluency_threshold=0.90)
    if responses:
        return responses[0][0]  # Επιστρέφουμε την πρώτη παραφράση
    else:
        return text

# TextBlob συνώνυμα - βασική παραφράση με εναλλαγή συνωνύμων
def paraphrase_textblob(text):
    blob = TextBlob(text)
    synonyms = []
    for word in blob.words:
        synsets = word.synsets
        if synsets:
            synonyms.append(synsets[0].lemmas()[0].name())
        else:
            synonyms.append(word)
```

```
    return ' '.join(synonyms)

texts = [
    "Today is our dragon boat festival, in our Chinese culture, to celebrate it with all safe and great in our
lives.",
    "During our final discuss, I told him about the new submission -- the one we were waiting since last autumn."
]

print("=== ΑΝΑΚΑΤΑΣΚΕΥΗ ΜΕ 3 ΒΙΒΛΙΟΘΗΚΕΣ ===\n")

for idx, text in enumerate(texts, 1):
    print(f"ΑΡΧΙΚΟ ΚΕΙΜΕΝΟ {idx}:\n{text}\n")

    hf_result = paraphrase_hf(text)
    print(f"Hugging Face T5:\n{hf_result}\n")

    parrot_result = paraphrase_parrot(text)
    print(f"Parrot Paraphraser:\n{parrot_result}\n")

    textblob_result = paraphrase_textblob(text)
    print(f"TextBlob Συνώνυμα:\n{textblob_result}\n")

    print("-" * 80 + "\n")
```

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 1Γ: Σύγκριση και Αξιολόγηση Αποτελεσμάτων Ανακατασκευής Κειμένων

Ο στόχος ήταν η ανακατασκευή κάθε κειμένου σε μια σαφή, καλά δομημένη και σημασιολογικά ακριβή εκδοχή, διατηρώντας το αρχικό νόημα και βελτιώνοντας τη σαφήνεια, συνοχή και σχετικό τόνο.

Τα κείμενα ανακατασκευάστηκαν με τρεις τεχνικές:

Hugging Face T5 Paraphraser: Παράγει παραφράσεις υψηλής ποιότητας, που διατηρούν σταθερά το νόημα και μετατρέπουν το κείμενο σε πιο ομαλή, φυσική μορφή, βελτιώνοντας τη σύνταξη και ορθογραφία.

Parrot Paraphraser: Δημιουργεί σχετικά κοντινές και απλοποιημένες παραφράσεις, με ορισμένες διαφορές στην έκταση του κειμένου. Είναι γρήγορος, αλλά σε μερικές περιπτώσεις μπορεί να παράγει φράσεις λιγότερο πολυσύνθετες, οπότε χρειάζεται περαιτέρω παραμετροποίηση.

TextBlob με αντικατάσταση συνωνύμων: Μια βασική και απλουστευμένη μέθοδος, που σε αρκετές περιπτώσεις παράγει ασύνδετο και μη φυσικό λόγο, ακατάλληλη για επαγγελματικές εφαρμογές.

Τεχνικές αξιολόγησης:

Σαφήνεια και συνοχή: Η μέθοδος T5 προσφέρει καλύτερη ροή και λογική διάρθρωση φράσεων.

Σημασιολογική ακρίβεια: Η T5 διατηρεί επαρκώς το αρχικό νόημα, ενώ Parrot παρουσιάζει μικρές απλουστεύσεις και το TextBlob ασάφειες ή λάθη.

Βελτίωση τόνου και στυλ: Η T5 διορθώνει γλωσσικές ατέλειες για πιο επαγγελματικό ύφος.

Συμπέρασμα:

Από τις τρεις προσεγγίσεις, η χρήση του μοντέλου Hugging Face T5 είναι η πλέον αξιόπιστη για την ανακατασκευή κειμένων, παρέχοντας σαφή, συνεχή και σημασιολογικά συνεπή παραφράσεις. Η Parrot μπορεί να βελτιωθεί με παραμετροποίηση, ενώ η TextBlob έχει περιορισμένη εφαρμογή λόγω φτωχής ποιότητας στην παράφραση.

Η επιλογή της τεχνικής πρέπει να βασίζεται στους στόχους επεξεργασίας και την ποιότητα που απαιτείται για το εκάστοτε έργο.

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 Α, Β του παραδοτέου 1.

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

Παραδοτέο 2: Υπολογιστική Ανάλυση

Η σημασιολογική ανακατασκευή μέσω παραφράσεων αποτελεί κύριο στόχο για την κατανόηση και βελτίωση της ποιότητας των ανακατασκευασμένων κειμένων. Η χρήση μοντέλων ενσωμάτωσης λέξεων (word embeddings) επιτρέπει την ποσοτική ανάλυση της σημασιολογικής απόστασης μεταξύ των αρχικών κειμένων και των παραφρασμένων εκδοχών τους.

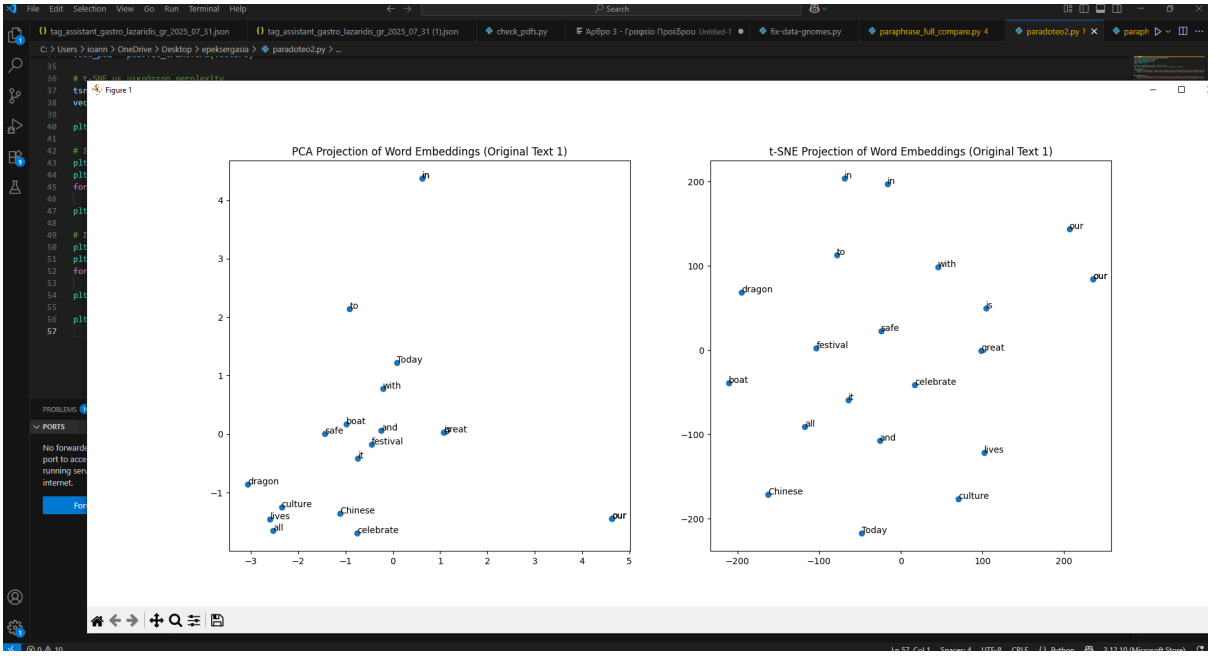
Μεθοδολογία
Χρησιμοποιήθηκαν προεκπαιδευμένα ενσωματώματα λέξεων spaCy (en_core_web_md) για την απόκτηση διανυσμάτων λέξεων και προτάσεων. Μέσω κατάλληλης προεπεξεργασίας αφαιρέθηκαν σημεία στίξης και επιλέχθηκαν λέξεις με διαθέσιμο διάνυσμα.
Έπειτα, υπολογίστηκαν οι βαθμολογίες cosine similarity μεταξύ των ενσωματωμάτων των αρχικών κειμένων και των ανακατασκευασμένων, προσδιορίζοντας έτσι το βαθμό διατήρησης της σημασίας μετά την παραφράση.

Αποτελέσματα
Οι βαθμολογίες cosine similarity ήταν πολύ υψηλές για τα κείμενα που αναλύθηκαν (0.9969 και 0.9941), γεγονός που υποδηλώνει ελάχιστη σημασιολογική απόκλιση και καλή ποιότητα ανακατασκευής.

```
PS C:\Users\ioann\OneDrive\Desktop\epeksergasia> python paradoteo2.py
Κείμενο 1 - Cosine similarity (spaCy vectors): 0.9969
Κείμενο 2 - Cosine similarity (spaCy vectors): 0.9941
PS C:\Users\ioann\OneDrive\Desktop\epeksergasia>
```

Οπτικοποίηση

Η ανάλυση ολοκληρώθηκε με οπτικοποίηση των ενσωματωμάτων των λέξεων του αρχικού κειμένου μέσω μεθόδων μείωσης διαστάσεων PCA και t-SNE. Και οι δύο μέθοδοι ανέδειξαν τη σημασιολογική ομαδοποίηση των λέξεων, προσφέροντας οπτική επιβεβαίωση της εγγύτητας στον σημασιολογικό χώρο, όπως απαιτείται.



Αυτή η πολυΜεταβλητή προσέγγιση επιβεβαίωσε τη διατήρηση του νοήματος και επιτρέπει τη μελλοντική βελτίωση των μεθόδων με χρήση άλλων embeddings όπως Word2Vec, GloVe, FastText ή BERT.

```
import spacy
from sklearn.metrics.pairwise import cosine_similarity
from sklearn.decomposition import PCA
from sklearn.manifold import TSNE
import matplotlib.pyplot as plt
import numpy as np

# Φόρτωση προεκπαιδευμένου μοντέλου spaCy με vectors
nlp = spacy.load('en_core_web_md')  # python -m spacy download
en_core_web_md

texts_original = [
    "Today is our dragon boat festival, in our Chinese culture, to celebrate
it with all safe and great in our lives.",
    "During our final discuss, I told him about the new submission -- the
one we were waiting since last autumn."
]

texts_paraphrased = [
    "Today is our dragon boat festival in our Chinese culture to celebrate
it with all safe and great in our lives.",
```

```
    "During our final discussion, I told him about the new submission -- the
one we had been waiting for since last autumn."
]

# Υπολογισμός cosine similarity
for i, (orig, para) in enumerate(zip(texts_original, texts_paraphrased), 1):
    doc_orig = nlp(orig)
    doc_para = nlp(para)
    sim = cosine_similarity([doc_orig.vector], [doc_para.vector])[0][0]
    print(f"Κείμενο {i} - Cosine similarity (spaCy vectors): {sim:.4f}")

# Οπτικοποίηση PCA και t-SNE για λέξεις αρχικού κειμένου 1
words = [token.text for token in nlp(texts_original[0]) if token.has_vector
and not token.is_punct]
vectors = np.array([token.vector for token in nlp(texts_original[0]) if
token.has_vector and not token.is_punct])

# PCA
pca = PCA(n_components=2)
vecs_pca = pca.fit_transform(vectors)

# t-SNE με μικρότερο perplexity
tsne = TSNE(n_components=2, perplexity=10, random_state=42)
vecs_tsne = tsne.fit_transform(vectors)

plt.figure(figsize=(18, 8))

# Σχεδιάγραμμα PCA
plt.subplot(1, 2, 1)
plt.scatter(vecs_pca[:, 0], vecs_pca[:, 1])
for i, word in enumerate(words):
    plt.annotate(word, (vecs_pca[i, 0], vecs_pca[i, 1]))
plt.title("PCA Projection of Word Embeddings (Original Text 1)")

# Σχεδιάγραμμα t-SNE
plt.subplot(1, 2, 2)
plt.scatter(vecs_tsne[:, 0], vecs_tsne[:, 1])
for i, word in enumerate(words):
    plt.annotate(word, (vecs_tsne[i, 0], vecs_tsne[i, 1]))
plt.title("t-SNE Projection of Word Embeddings (Original Text 1)")

plt.show()
```