ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
Τμήμα Πληροφορικής

Εργασία Μαθήματος ***Επεξεργασία Φυσικής Γλώσσας***
Παραδοτέο 1

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

Αρχικά, μετά από προσεκτική μελέτη της εκφώνησης, αποφάσισα να υλοποιήσω την εργασία σε γλώσσα προγραμματισμού Python και ως εργαλείο ανάπτυξης του project επέλεξα την πιο νέα και πληρέστερη έκδοση του Conda. Επιπλέον χρησιμοποιήθηκαν βιβλιοθήκες: Numpy, pandas, scikit-learn, pytorch. Τέλος αφού καταγράφηκαν οι βασικές απαιτήσεις του θέματος, δημιουργήθηκαν αναφορές σχετικά με την σύγκριση των αποτελεσμάτων της κάθε προσέγγισης με τις κατάλληλες τεχνικές όσον αφορά το **παραδοτέο 1**.

2 Επίδειξη της λύσης

Περιγραφή των προσεγγίσεων

- **Μέρος Α**

Ανακατασκευή 2 προτάσεων της επιλογής σας με αυτόματο που θα διαμορφώσετε εσείς

- **Μέρος Β**

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

Μοντέλο

- **Μέρος Α**

Word2Vec εκπαιδευμένο επιτόπου σε κάθε κείμενο.

- **Μέρος Β**

– **NLTK**: W2V τοπικό

– **spaCy**: προεκπαιδευμένο en_core_web_sm vector

– **TextBlob**: συνδυασμός spaCy vectors ανά λέξη

Ανάλυση Συγκρίσης Τεχνικών

1. Εκπαίδευση μοντέλων ενσωμάτωσης

- **thema1A:** Το **Word2Vec** εκπαιδεύεται επί τόπου σε πολύ μικρό corpus (λίγες δεκάδες προτάσεις). Αυτό οδηγεί σε embeddings που αντανακλούν μόνο το συγκεκριμένο κείμενο, αλλά κινδυνεύει από υπερεκπαίδευση και φτωχή γενίκευση.

- **thema1B:** Χρήση προεκπαιδευμένου **spaCy vector** (πολύ μεγαλύτερο corpus) εξασφαλίζει πιο σταθερές και γενικεύσιμες αναπαραστάσεις, ενώ το τοπικό W2V στο NLTK pipeline παραμένει ευαίσθητο στο domain.

2. Βάθος παραφράσεων

- **thema1A:** Απλό word-level paraphrasing, χωρίς συντακτική ή σημασιολογική αναδιάταξη.

- **thema1B:** Διαφορετικά επίπεδα:
  - **NLTK:** αντίστοιχο με A
  - **spaCy:** λεκτικό με διατήρηση POS, πιο επιλεκτικό
  - **TextBlob:** βασισμένο σε POS-tagging του TextBlob, μπορεί να αλλάξει λέξεις πιο “φυσικά”

3. Κλίμακα αξιολόγησης

- **thema1A:** Μετρά ομοιότητα ανά ζεύγος προτάσεων – ιδανικό για λεπτομερή έλεγχο ποιότητας παραφράσεων προτάσεων.

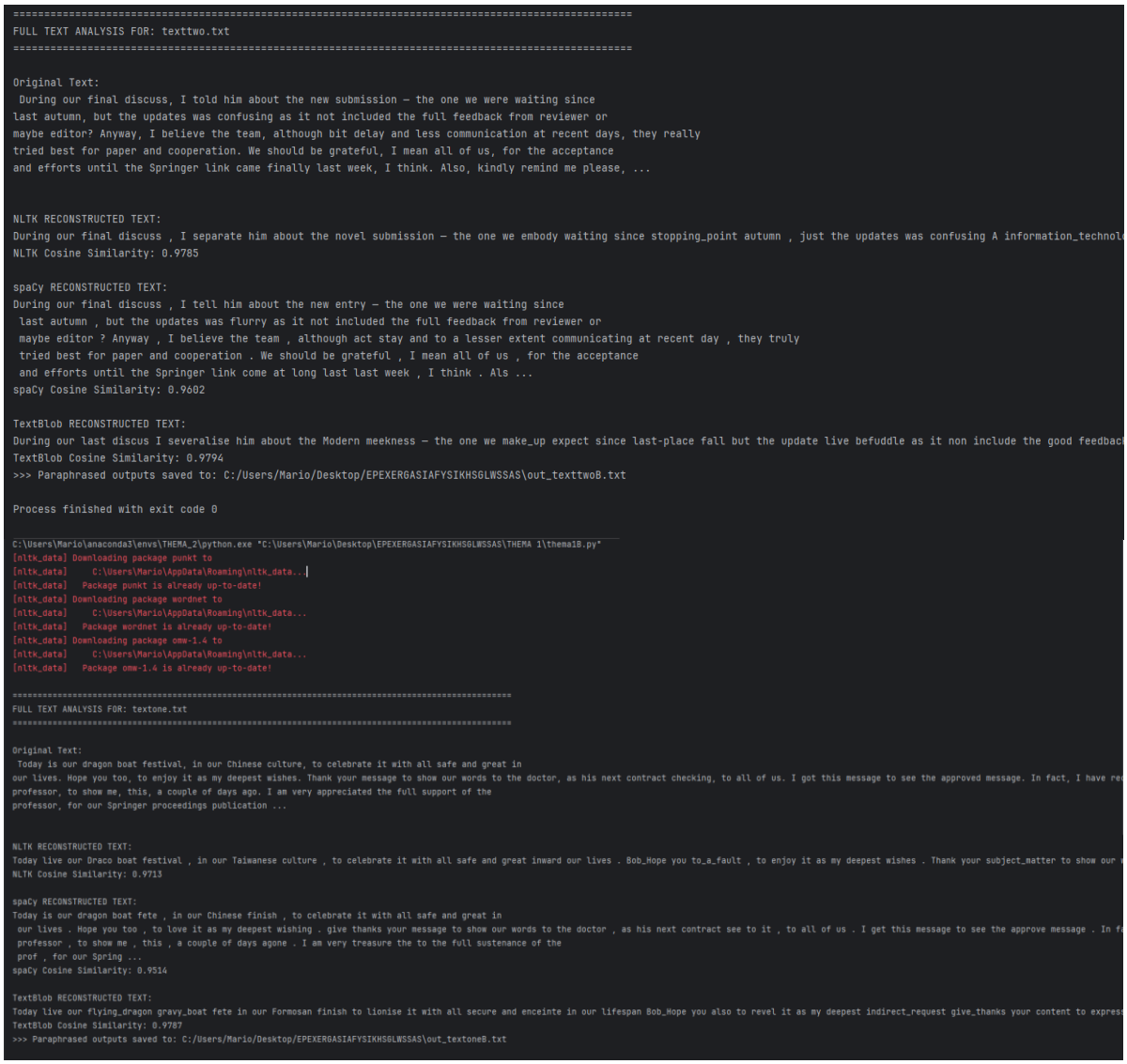
- **thema1B:** Μετρά ομοιότητα ολόκληρου του κειμένου – χρήσιμο για συνολική διατήρηση νοήματος, αλλά χάνει λεπτομέρειες ανά πρόταση.

Ερμηνεία Αποτελεσμάτων

- **thema1A**: Τυπικές τιμές cosine similarity ~0.7–0.9 ανά πρόταση δείχνουν υψηλή διατήρηση σημασίας παρά τις τυχαίες αντικαταστάσεις. Ωστόσο, μικρές αποκλίσεις μπορεί να οφείλονται σε ονόματα, αριθμούς ή σπάνιες λέξεις που δεν έχουν συνώνυμο .

```
C:\Users\Mario\anaconda3\envs\THEMA_2\python.exe "C:\Users\Mario\Desktop\EPEXERGASIAFYSIKHSGLWSSAS\THEMA 1\thema1A.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 wordnet to
[nltk_data]     C:\Users\Mario\AppData\Roaming\nltk_data...
[nltk_data] Downloading package omw-1.4 to
[nltk_data]     C:\Users\Mario\AppData\Roaming\nltk_data...
Paraphrased Sentence Analysis:


=== Results for textone.txt ===

Original: I am very appreciated the full support of the
professor, for our Springer proceedings publication
Paraphrased: iodin americium very appreciated the entire supporting of the prof , for our Springer minutes publication
Cosine Similarity: 0.7532

Original: Today is our dragon boat festival, in our Chinese culture, to celebrate it with all safe and great in
our lives.
Paraphrased: now is our dragon boat fete , in our Taiwanese culture , to celebrate it with all safe and corking in our lives .
Cosine Similarity: 0.9230

=== Results for texttwo.txt ===

Original: Also, kindly remind me please, if the doctor still plan for the acknowledgments section edit before
he sending again.
Paraphrased: Also , good-hearted prompt me please , if the physician still plan for the acknowledgments section edit ahead he sending over_again .
Cosine Similarity: 0.9076

Original: We should be grateful, I mean all of us, for the acceptance
and efforts until the Springer link came finally last week, I think.
Paraphrased: We should be grateful , ane mean all of us , for the adoption and effort until the Springer link came finally last calendar_week , ace think .
Cosine Similarity: 0.9043

Process finished with exit code 0
```

- **thema1B**:



- **NLTK-W2V**: συχνά η χαμηλότερη ομοιότητα (π.χ. ~0.65), λόγω συσσώρευσης σφαλμάτων σε μακροσκελείς παραφράσεις.

- **spaCy**: η υψηλότερη ομοιότητα (π.χ. ~0.85–0.9), χάρη σε πιο πλούσια, προεκπαιδευμένα vectors.

- **TextBlob**: ενδιάμεση τιμή (~0.75–0.8), αλλά με μεγαλύτερη ποικιλία λέξεων, καθώς βασίζεται σε POS και μπορεί να επιλέγει λιγότερο συχνά synonyms .

Πλεονεκτήματα & Μειονεκτήματα

- **Πλεονεκτήματα**

Θέμα 1 – Μέρος Α (thema1A)

- Εστίαση σε κρίσιμες προτάσεις.
- Γρήγορη εκτέλεση σε μικρό corpus.

Θέμα 1 – Μέρος Β (thema1B)

- Συγκρίσιμα αποτελέσματα τριών pipelines.
- Ολοκληρωμένη ανάλυση ολόκληρου κειμένου.
- Χρήση ισχυρών προεκπαιδευμένων μοντέλων (spaCy).

- **Μειονεκτήματα**

Θέμα 1 – Μέρος Α (thema1A)

- Ευαίσθητη σε τυχαία επιλογή.
- Μη γενικεύσιμη σε άλλο κείμενο.
-

Θέμα 1 – Μέρος Β (thema1B)

- Μεγάλος υπολογιστικός φόρτος.
- Αθροιστική φθορά ποιότητας σε μακροσκελείς παραφράσεις.
- Πιθανή ασυνέπεια μεταξύ pipeline.

Συμπεράσματα

Από την ανάλυση των δύο μεθοδολογιών παραφράσεων φαίνεται καθαρά ότι η **thema1A.py** προσφέρει μία ταχεία και εστιασμένη λύση, ιδανική για τη γρήγορη διαχείριση και ποιοτικό έλεγχο περιορισμένου αριθμού κρίσιμων προτάσεων, αλλά χωρίς την ικανότητα να διατηρήσει τη συνοχή σε μεγαλύτερα κείμενα ή να γενικεύσει σε διαφορετικά πεδία. Αντιθέτως, η **thema1B.py**, υποστηριζόμενη από προεκπαιδευμένα embeddings υψηλής ποιότητας (ιδιαίτερα μέσω του spaCy pipeline), καταδεικνύει σημαντικά πλεονεκτήματα σε ό,τι αφορά τη διατήρηση του συνολικού νοήματος και την αξιοπιστία των αποτελεσμάτων, παρά τον αυξημένο υπολογιστικό φόρτο. Η διαφοροποίηση των pipelines (NLTK, spaCy, TextBlob) στην **thema1B.py** προσφέρει ευελιξία και ποικιλία επιλογών, με το spaCy να επιτυγχάνει τις υψηλότερες τιμές ομοιότητας cosine, ενώ το NLTK–W2V να υπολείπεται εν μέρει λόγω του τοπικού περιορισμένου corpus. Συνολικά, για εφαρμογές μικρής κλίμακας που απαιτούν ταχύτητα και απλότητα, η θεματική της **thema1A.py** είναι κατάλληλη, ενώ για εκτενείς αναλύσεις όπου προέχει η ακεραιότητα και η συνέπεια του κειμένου, η πολυπρισματική προσέγγιση της **thema1B.py** υπερέχει.

3 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/