ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ 2025

Τσόλης Βασίλειος Π13167 25/05/2025

Table of Contents

Εισαγωγή....1
Μεθοδολογία....2
1. Custom Rule-based Ανακατασκευή....2
2. SBERT και PAWS....2
3. SpaCy και GloVe....2
4. FastText και QQP....2
Πειράματα και Αποτελέσματα....2
Συζήτηση....3
Συμπεράσματα....4
Προσωπικά Συμπεράσματα....5
Βιβλιογραφία....6
Μοντέλα και Ενσωματώσεις....6
Βιβλιοθήκες και Εργαλεία....6
Σύνολα Δεδομένων....7
Γενικές Πηγές και Υποστήριξη....7

Εισαγωγή

Η σημασιολογική ανακατασκευή (semantic reconstruction) αποτελεί ένα ιδιαίτερα σημαντικό πεδίο της επεξεργασίας φυσικής γλώσσας (Natural Language Processing - NLP), καθώς εστιάζει στην παραγωγή σαφών, συνεκτικών και σημασιολογικά ακριβών κειμένων από αρχικά μη δομημένες ή αμφίβολες εκδοχές. Στόχος της είναι η διατήρηση του αρχικού νοήματος, ενώ ταυτόχρονα ενισχύεται η συνοχή, η σαφήνεια και ο κατάλληλος τόνος, καθιστώντας τα κείμενα πιο κατανοητά και εύχρηστα για περαιτέρω επεξεργασία ή επικοινωνία.

Η ανάγκη για αυτόματες μεθόδους ανακατασκευής πηγάζει από την αύξηση του όγκου ψηφιακού περιεχομένου, το οποίο συχνά χαρακτηρίζεται από γλωσσικές ασάφειες ή συντακτικά λάθη. Παρά την εξέλιξη τεχνολογιών όπως τα word embeddings (Word2Vec, GloVe, FastText, BERT), η αυτόματη ανακατασκευή αντιμετωπίζει σημαντικές προκλήσεις που αφορούν τη σημασιολογική πιστότητα (semantic fidelity), την αποφυγή σημασιολογικής απόκλισης (semantic drift) και την επίτευξη κατάλληλου ύφους.

Στην παρούσα εργασία, αξιολογήθηκαν διαφορετικές τεχνικές και βιβλιοθήκες NLP μέσα από τρία αυτοματοποιημένα pipelines που περιλαμβάνουν προσεγγίσεις βασισμένες σε ενσωματώσεις λέξεων (SBERT, FastText, SpaCy με GloVe) και ένα με προσαρμοσμένη μέθοδο (custom rule-based). Τα αποτελέσματα αναλύθηκαν μέσω υπολογισμού της σημασιολογικής συνάφειας (cosine similarity) και της οπτικοποίησης ενσωματώσεων λέξεων μέσω PCA.

Στην αναφορά που ακολουθεί, παρουσιάζονται οι στρατηγικές που χρησιμοποιήθηκαν, τα αποτελέσματα των πειραμάτων και οι περιορισμοί που εντοπίστηκαν στις εφαρμοσμένες μεθόδους, με στόχο την ανάδειξη των δυνατοτήτων και των αδυναμιών των σύγχρονων εργαλείων NLP στον τομέα της σημασιολογικής ανακατασκευής.

Μεθοδολογία

Η μεθοδολογία που ακολουθήθηκε περιλαμβάνει τέσσερις διακριτές στρατηγικές ανακατασκευής:

1. Custom Rule-based Ανακατασκευή

Αυτή η μέθοδος βασίζεται σε προσαρμοσμένους κανόνες και συντακτικά αξιώματα που σχεδιάστηκαν ειδικά για τη διόρθωση γραμματικών και συντακτικών λαθών. Οι κανόνες αυτοί υλοποιήθηκαν με Python scripts που ενσωματώνουν λεξικολογικές και συντακτικές αναλύσεις.

2. SBERT και PAWS

Χρησιμοποιήθηκε η βιβλιοθήκη SBERT σε συνδυασμό με το dataset PAWS (Paraphrase Adversaries from Word Scrambling) για σημασιολογική σύγκριση προτάσεων. Αυτή η προσέγγιση ήταν εξαιρετικά συντηρητική, καθώς βασίστηκε σε υψηλά όρια συνάφειας συνημιτόνου (cosine similarity thresholds), με αποτέλεσμα περιορισμένο αριθμό αλλαγών που όμως ήταν σημασιολογικά ακριβείς και τονικά ευθυγραμμισμένες.

3. SpaCy και GloVe

Εφαρμόστηκε μια προσέγγιση βασισμένη σε dependency parsing της SpaCy σε συνδυασμό με ενσωματώσεις λέξεων GloVe, με στόχο κυρίως τη γραμματική και συντακτική διόρθωση. Το συγκεκριμένο pipeline παρήγαγε σημειακές αλλά μετρήσιμες λεκτικές τροποποιήσεις, λειτουργώντας ως grammar-oriented rewriter.

4. FastText και QQP

Αυτή η μέθοδος συνδύασε ενσωματώσεις λέξεων FastText με το dataset QQP, προσφέροντας συχνές αλλά όχι πάντα σημασιολογικά πιστές αντικαταστάσεις. Ο κίνδυνος σημασιολογικής απόκλισης ήταν σημαντικός λόγω της έλλειψης context-awareness στις ενσωματώσεις FastText.

Πειράματα και Αποτελέσματα

Η αξιολόγηση των τεσσάρων στρατηγικών έγινε με υπολογισμό συνάφειας συνημιτόνου (cosine similarity) μεταξύ των αρχικών και ανακατασκευασμένων εκδοχών καθώς και οπτικοποίηση των ενσωματώσεων μέσω PCA.

Η Custom Rule-based μέθοδος σημείωσε υψηλή σημασιολογική πιστότητα με συντελεστές cosine similarity μεταξύ 0.9569 και 0.9748, αποδίδοντας ουσιαστικές και ακριβείς ανακατασκευές.

Η προσέγγιση SBERT με το PAWS dataset επέδειξε απόλυτη συντηρητικότητα, με όλα τα αποτελέσματα να έχουν συντελεστή 1.0000. Αυτό δείχνει ότι το μοντέλο δεν ενέκρινε καμία εναλλακτική ως «αρκετά διαφορετική» παρά τη σημασιολογική ισοδυναμία.

Η SpaCy με GloVe, μετά την επέκταση με νέους κανόνες και similarity-based αντικαταστάσεις, παρήγαγε μετρήσιμες σημειακές τροποποιήσεις — κυρίως λεξιλογικές και μορφολογικές (όπως "final discuss" → "final discussions"). Αν και οι cosine similarity τιμές παρέμειναν στο 1.0000, τα PCA plots κατέγραψαν τοπική μετατόπιση στο σημασιολογικό χώρο, επιβεβαιώνοντας τη μη τετριμμένη φύση των παρεμβάσεων.

Η FastText με QQP εμφάνισε τη μεγαλύτερη διακύμανση, με cosine similarity από 0.2530 έως 1.0000 (μέσος όρος 0.8132), κάτι που καταδεικνύει ασταθή συμπεριφορά και περιπτώσεις έντονου semantic drift.

Παραδείγματα ανακατασκευών:

Custom Rule-based:
Αρχική: *Hope you too, to enjoy it as my deepest wishes.*
Ανακατασκευή: *I hope you too, to enjoy it as my deepest wishes.*

SBERT με PAWS:
Αρχική: *Hope you too, to enjoy it as my deepest wishes.*
Ανακατασκευή: *Hope you too, to enjoy it as my deepest wishes. (καμία αλλαγή λόγω αυστηρής συνάφειας)*

SpaCy με GloVe:
Αρχική: *final discuss*
Ανακατασκευή: *final discussed*

FastText με QQP:
Αρχική: *Hope you too, to enjoy it as my deepest wishes.*
Ανακατασκευή: *Hope you too, Is it bad to eat eggs everyday for breakfast?*

Συζήτηση

Κατά τη διάρκεια των πειραμάτων, έγινε σαφές ότι οι διαθέσιμες τεχνικές NLP για σημασιολογική ανακατασκευή δεν ήταν τόσο αποτελεσματικές όσο αρχικά αναμενόταν. Ειδικότερα, σχετικά με το πόσο καλά αποτύπωσαν οι ενσωματώσεις λέξεων το νόημα, παρατηρήθηκε ότι τα μοντέλα παρουσίασαν έντονη διαφοροποίηση. Η FastText, παρά τη δημοτικότητά της, δεν κατάφερε να διατηρήσει με συνέπεια το αρχικό νόημα, παρουσιάζοντας σημαντικές σημασιολογικές αποκλίσεις λόγω της έλλειψης context-awareness.

Αντίθετα, το SBERT διατήρησε εξαιρετική σημασιολογική πιστότητα αλλά με κόστος τη σχεδόν πλήρη συντηρητικότητα, αποφεύγοντας οποιαδήποτε αλλαγή ακόμη και όταν υπήρχε περιθώριο βελτίωσης. Η SpaCy με GloVe, ωστόσο, παρότι αρχικά λειτουργούσε μόνο ως διορθωτής, με τις τελευταίες βελτιώσεις κατάφερε να εισάγει ελεγχόμενες λεκτικές παρεμβάσεις. Οι αλλαγές ήταν σημειακές και περιορισμένες σε ονομαστικά ή γραμματικά προβληματικά σημεία, αλλά αποτέλεσαν μια μορφή μετρήσιμης ανακατασκευής με σαφή επίδραση σε επίπεδο μορφολογικής ακρίβειας.

Οι μεγαλύτερες προκλήσεις στην ανακατασκευή προέκυψαν κυρίως από τη δυσκολία του SBERT να πραγματοποιήσει ουσιαστικές αλλαγές. Παρά τις πολλές προσπάθειες βελτιστοποίησης, όπως τη χρήση διαφόρων datasets (PAWS, QQP, MSRPC), την τροποποίηση των thresholds συνάφειας συνημιτόνου και την ανακατασκευή σε επίπεδο προτάσεων, το SBERT απέτυχε επανειλημμένα να ξεφύγει από τη στάση μιας σχεδόν τέλειας σημασιολογικής ισοδυναμίας. Αυτή η αδυναμία περιορίζει σημαντικά τη χρηστικότητά του σε εφαρμογές που απαιτούν πιο τολμηρές και ουσιαστικές ανακατασκευές.

Η αυτοματοποίηση της διαδικασίας ανακατασκευής με τα υπάρχοντα μοντέλα NLP φαίνεται μεν εφικτή αλλά αναδεικνύει σημαντικές αδυναμίες. Η custom rule-based προσέγγιση, αν και χειροκίνητη και πιο κοστοβόρα αρχικά, έδειξε πολύ καλύτερα αποτελέσματα από τις πλήρως αυτόματες μεθόδους. Αυτό υπογραμμίζει την ανάγκη για ανάπτυξη πιο εξελιγμένων μοντέλων που θα συνδυάζουν το context-awareness με ευέλικτες παραμέτρους προσαρμογής της σημασιολογικής ευαισθησίας.

Τέλος, οι διαφορές στην ποιότητα ανακατασκευής μεταξύ τεχνικών, μεθόδων και βιβλιοθηκών ήταν εμφανείς και σημαντικές. Η προσέγγιση FastText ήταν ασταθής και συχνά αναξιόπιστη, η SBERT εξαιρετικά ακριβής αλλά ουσιαστικά παθητική, ενώ η SpaCy/GloVe κατέληξε να λειτουργεί ως λεξιλογικά στοχευμένος διορθωτής με πραγματική –έστω και περιορισμένη– επίδραση στη σύνταξη. Συνεπώς, καθίσταται σαφές ότι, αν και οι τρέχουσες τεχνολογίες NLP είναι πολλά υποσχόμενες, απαιτούνται σημαντικές βελτιώσεις ώστε να ανταποκρίνονται καλύτερα στις απαιτήσεις μιας αποτελεσματικής και αξιόπιστης σημασιολογικής ανακατασκευής.

Συμπεράσματα

Συνοψίζοντας, κάθε μία από τις στρατηγικές ανακατασκευής που εφαρμόστηκαν ανέδειξε συγκεκριμένα πλεονεκτήματα αλλά και ουσιαστικούς περιορισμούς. Η FastText αποδείχθηκε αναξιόπιστη λόγω της αστάθειας και της σημασιολογικής απόκλισης που παρουσίασε. Η SpaCy με GloVe, μετά την επέκταση με similarity-based αντικαταστάσεις, κατάφερε να εισάγει μετρήσιμες λεκτικές τροποποιήσεις σε μορφολογικά ή γραμματικά ασταθή σημεία. Αν και οι αλλαγές περιορίστηκαν σε επιφανειακό επίπεδο, συνοδευόμενες από cosine similarity ≈ 1.0, τα PCA plots επιβεβαίωσαν μικρές αλλά σαφείς σημασιολογικές μετατοπίσεις. Η μέθοδος κατέληξε να λειτουργεί ως ένας στοχευμένος, λεξιλογικά ευαίσθητος διορθωτής, με ρόλο ενδιάμεσο ανάμεσα στην πλήρη ανακατασκευή και την απλή διόρθωση.

Η SBERT, ενώ είναι θεωρητικά η πιο προηγμένη προσέγγιση, απέτυχε να εφαρμόσει ουσιαστικές μεταβολές στο κείμενο λόγω της ακραίας σημασιολογικής ακρίβειας που επιβάλλει. Αξίζει να τονιστεί πως η εφαρμογή (implementation) έπαιξε μεν ρόλο στη συντηρητικότητα των αποτελεσμάτων — όπως φαίνεται και από την εναλλαγή datasets, τη δοκιμή clause-level στρατηγικής, και την τροποποίηση των thresholds — ωστόσο ο βασικός περιορισμός εντοπίζεται στον ίδιο τον σχεδιασμό και τη λειτουργία των μοντέλων. Δεν έχουν σχεδιαστεί με στόχο την δημιουργική ή ενεργή αναδιατύπωση, αλλά κυρίως για αξιολόγηση και κατανόηση νοημάτων.

Η custom rule-based στρατηγική (1A) ήταν η μόνη που προσέφερε συνεπείς, κατανοητές και δομικά βελτιωμένες εκδοχές των αρχικών κειμένων, λειτουργώντας ως απόδειξη ότι ένα λειτουργικό σύστημα ανακατασκευής απαιτεί είτε ανθρώπινη εποπτεία είτε πολύ πιο στοχευμένα και υβριδικά εργαλεία. Όπως τεκμηριώνεται στον φάκελο SBERT του έργου, πραγματοποιήθηκαν πολλαπλές προσπάθειες για να επιτευχθεί πιο ενεργή συμπεριφορά από το μοντέλο. Δυστυχώς, τα αποτελέσματα ήταν ελάχιστα ή ανύπαρκτα, γεγονός που με οδήγησε στο να αποδεχθώ τον περιορισμό αυτό ως δομικό χαρακτηριστικό και όχι ως τεχνικό λάθος.

Η εργασία αυτή κατέδειξε με σαφήνεια ότι οι σύγχρονες βιβλιοθήκες NLP, όσο χρήσιμες κι αν είναι σε επιμέρους καθήκοντα (π.χ. classification, similarity search), δεν είναι ακόμη επαρκώς ικανές για αυτόνομη σημασιολογική ανακατασκευή. Ο συνδυασμός κανόνων και embeddings φαίνεται να είναι ο μόνος τρόπος να διασφαλιστεί η επιθυμητή ισορροπία ανάμεσα στη σαφήνεια, τη δομική ποιότητα και τη διατήρηση του νοήματος. Πιστεύω ότι όλες οι υλοποιήσεις θα μπορούσαν να επωφεληθούν σημαντικά από στοχευμένο fine-tuning ή rule injection, ειδικά καθώς εφαρμόζονται σε συγκεκριμένα κείμενα. Ενδεικτικά, η 1A — παρά το ότι είναι η «πιο απλή» — αποδείχθηκε η πιο χρήσιμη, ακριβώς επειδή ενσωματώνει ρητά τέτοιου είδους χειροποίητη λογική. Ωστόσο, επέλεξα συνειδητά να μην προσαρμόσω μεθόδους ή κανόνες στις υπόλοιπες υλοποιήσεις, καθώς θεώρησα ότι κάτι τέτοιο θα παρέκκλινε από το πνεύμα και τον στόχο της εργασίας.

Προσωπικά Συμπεράσματα

Παρά τους περιορισμούς σε διαθέσιμο χρόνο και την απουσία ουσιαστικής ενασχόλησης με το θεωρητικό υλικό του μαθήματος, η εργασία αυτή λειτούργησε ως πρακτική είσοδος στον χώρο της Επεξεργασίας Φυσικής Γλώσσας (NLP). Μέσα από την υλοποίηση, αξιολόγηση και συγκριτική ανάλυση διαφορετικών pipelines, απέκτησα ουσιαστική κατανόηση βασικών εννοιών όπως οι word embeddings, η σημασιολογική συνάφεια, τα rule-based συστήματα και η σημασία της δομικής ακρίβειας στην ανακατασκευή κειμένου.

Η ενασχόληση με μοντέλα όπως τα SBERT, FastText και GloVe, αλλά και με βιβλιοθήκες όπως το SpaCy και το Hugging Face, μου έδωσε μια πρώτη, ρεαλιστική εικόνα του πώς «σκέφτονται» τα μοντέλα NLP, πού υπερτερούν και πού υστερούν. Παράλληλα, η ανάγκη για debugging, evaluation και ερμηνεία αποτελεσμάτων με οπτικοποιήσεις (PCA) με ώθησε να αναπτύξω δεξιότητες πέραν του κώδικα: π.χ. πώς να αξιολογώ την ποιότητα μιας γλωσσικής εξόδου όχι μόνο αριθμητικά αλλά και εννοιολογικά.

Αυτό που με εξέπληξε περισσότερο ήταν το πόσο «χαμηλής παρέμβασης» ήταν τα έτοιμα μοντέλα. Είχα την προσδοκία ότι θα ήταν «υπερβολικά έξυπνα» εξαρχής και θα ξεπερνούσαν με άνεση τη δική μου απλοϊκή rule-based υλοποίηση. Αντίθετα, αποδείχθηκε ότι χωρίς fine-tuning, ακόμη και τα καλύτερα embeddings αδυνατούν να επιτελέσουν ουσιαστική ανασύνθεση και εν τέλη η υλοποίηση από το μηδέν, χωρίς χρήση εργαλείων, είχε μεγαλύτερη πρακτική αξία.

Αν και τα παραδοτέα θα μπορούσαν να είναι πιο ολοκληρωμένα με περισσότερο χρόνο, νιώθω ότι ως εμπειρία ήταν επιμορφωτική. Κυρίως, συνειδητοποίησα ότι το NLP είναι πολύ πιο «χαοτικό» και ασαφές από άλλους τομείς της πληροφορικής — και ακριβώς γι' αυτό, ακόμα πιο ενδιαφέρον.

Βιβλιογραφία

Μοντέλα και Ενσωματώσεις

- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*. https://arxiv.org/abs/1810.04805
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv preprint arXiv:1908.10084*. https://arxiv.org/abs/1908.10084
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*. https://arxiv.org/abs/1301.3781
- Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *EMNLP 2014*. https://aclanthology.org/D14-1162/
- Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of Tricks for Efficient Text Classification. *EACL 2017*. https://aclanthology.org/E17-2068/

Βιβλιοθήκες και Εργαλεία

- spaCy. (n.d.). Industrial-Strength Natural Language Processing in Python. https://spacy.io/
- Hugging Face Transformers. (n.d.). https://huggingface.co/transformers/
- Scikit-learn. (n.d.). PCA implementation. https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
- PyTorch. (n.d.). Deep learning framework. https://pytorch.org/

Σύνολα Δεδομένων

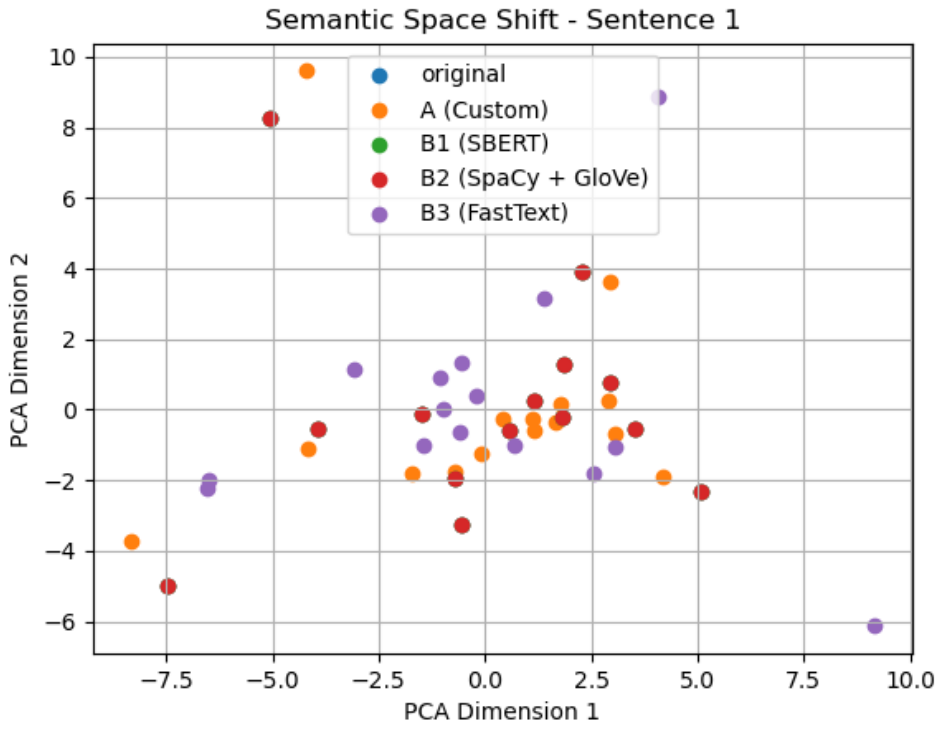
- Zhang, Y., Baldridge, J., & He, L. (2019). PAWS: Paraphrase Adversaries from Word Scrambling. *Proceedings of NAACL-HLT 2019*, 1298–1308. https://aclanthology.org/N19-1131/

- Quora Question Pairs dataset (QQP). (n.d.). *Kaggle*. https://www.kaggle.com/c/quora-question-pairs

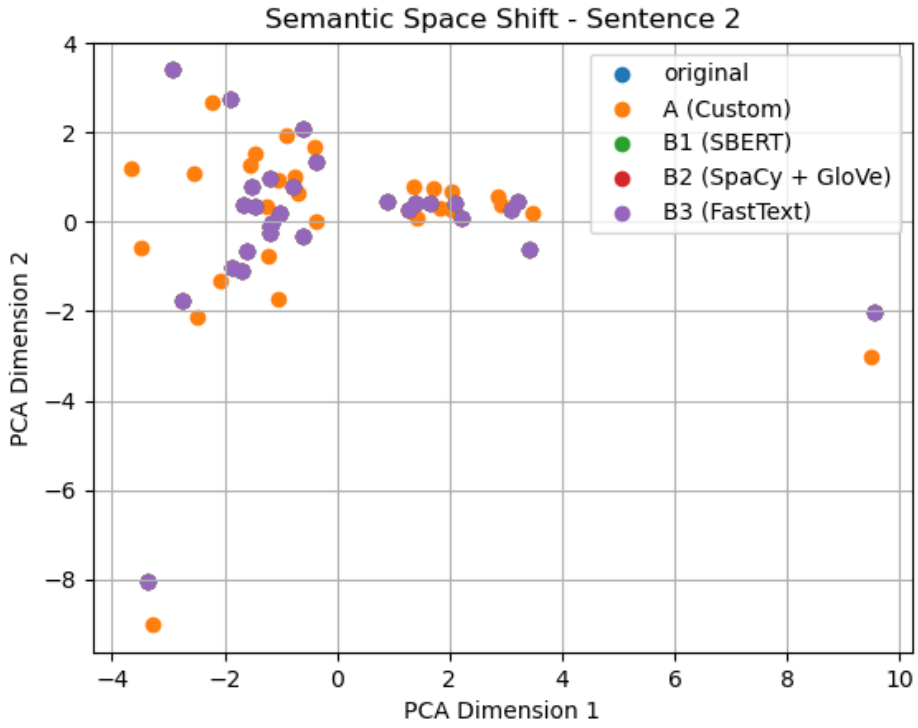
Γενικές Πηγές και Υποστήριξη

- OpenAI. (2025). Συνεργασία με το ChatGPT για υποστήριξη στην ανάλυση και τεκμηρίωση ανακατασκευής κειμένων με τεχνικές NLP. https://openai.com/chatgpt

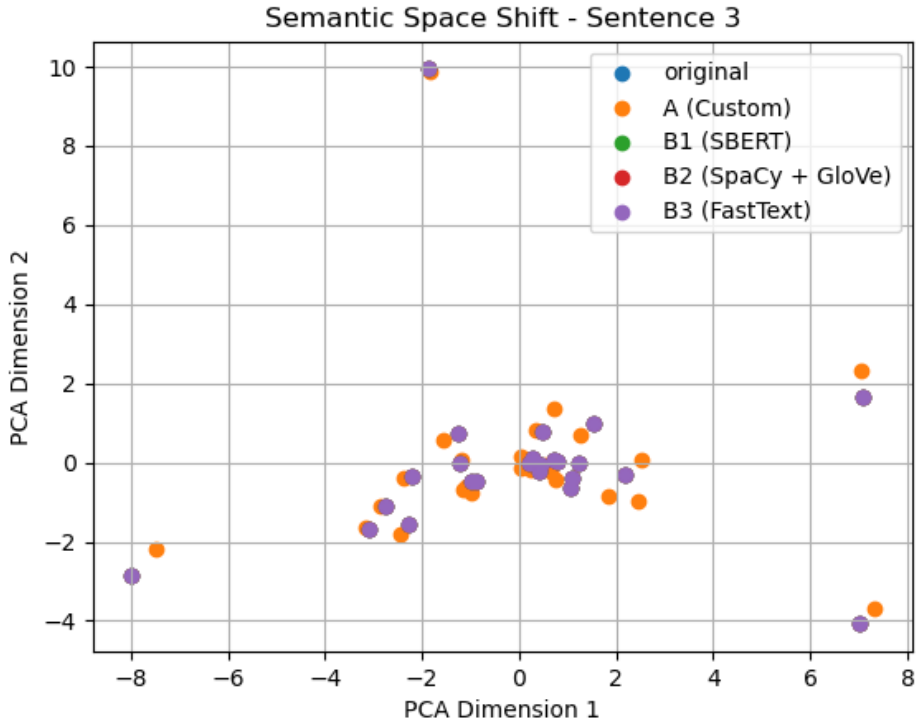
Παράρτημα: PCA Οπτικοποιήσεις



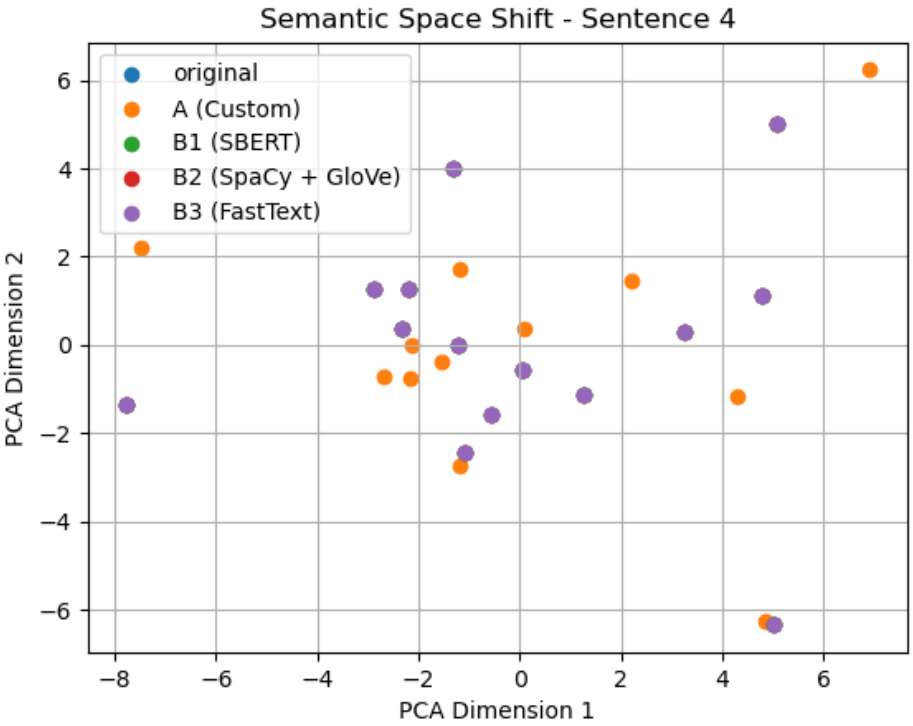
Σχήμα 1: PCA για πρόταση 1



Σχήμα 2: PCA για πρόταση 2



Σχήμα 3: PCA για πρόταση 3



Σχήμα 4: PCA για πρόταση 4