
# Βουλή των Ελλήνων

## Αυτόματη Κατηγοριοποίση Πολιτικού Λόγου

---

Άρης Φεργάδης, ΑΜ 03002718
Αντώνης Κορκοφίγκας, ΑΜ 03002703

# Περιγραφή του προβλήματος

- Αναγνώριση πολιτικού κόμματος από κείμενο ομιλίας
- Αναγνώριση ομιλητή από κείμενο ομιλίας
- Διερεύνηση σημασιολογικών αποστάσεων μεταξύ ομιλιών

# Δεδομένα

Περιγραφή - Ανάλυση

# Ιδιαιτερότητες

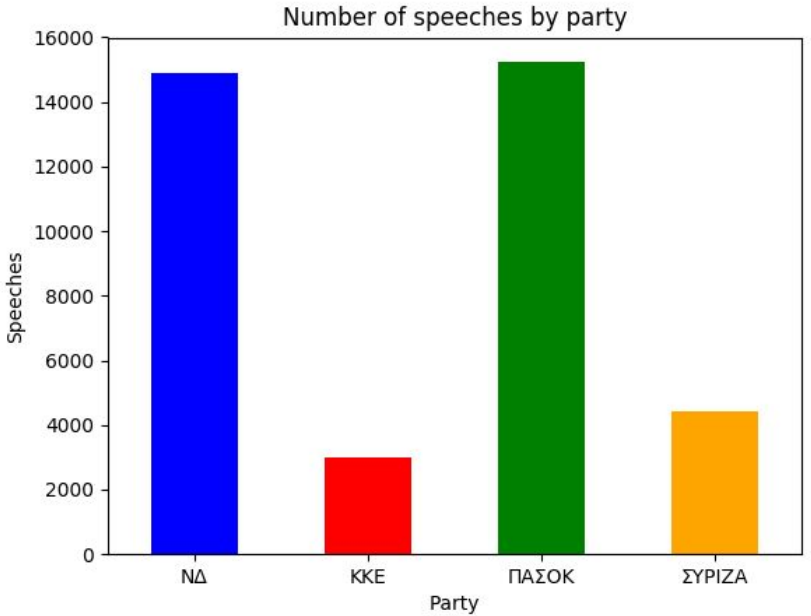
- Διαχωρισμός ομιλιών

ΖΩΗ ΚΩΝΣΤΑΝΤΟΠΟΥΛΟΥ: Ευχαριστώ, κύριε Πρόεδρε. …
ΓΕΩΡΓΙΟΣ ΚΑΛΑΝΤΖΗΣ (Β΄ Αντιπρόεδρος της Βουλής): Ολοκληρώστε, κυρία συνάδελφε.
ΖΩΗ ΚΩΝΣΤΑΝΤΟΠΟΥΛΟΥ: Ολοκληρώνω, κύριε Πρόεδρε. … *(συνέχεια ομιλίας)*

- Διαχωρισμός προτάσεων
  - Ελληνικό ερωτηματικό,
  - Τρεις τελείες,
  - Εκτενής χρήση της τελείας στα ακρωνύμια και στους αριθμούς.
  - Μη συνεπής γραφή (δισ. / δισεκατομμύρια)

# Πλήθος ομιλιών ανά κόμμα

| | |
|---|---|
| ΝΔ | 14874 |
| ΚΚΕ | 3001 |
| ΠΑΣΟΚ | 15233 |
| ΣΥΡΙΖΑ | 4431 |

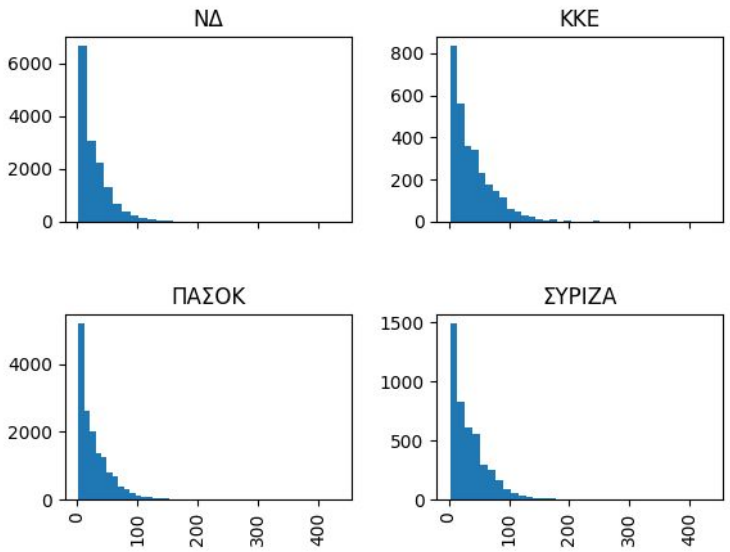


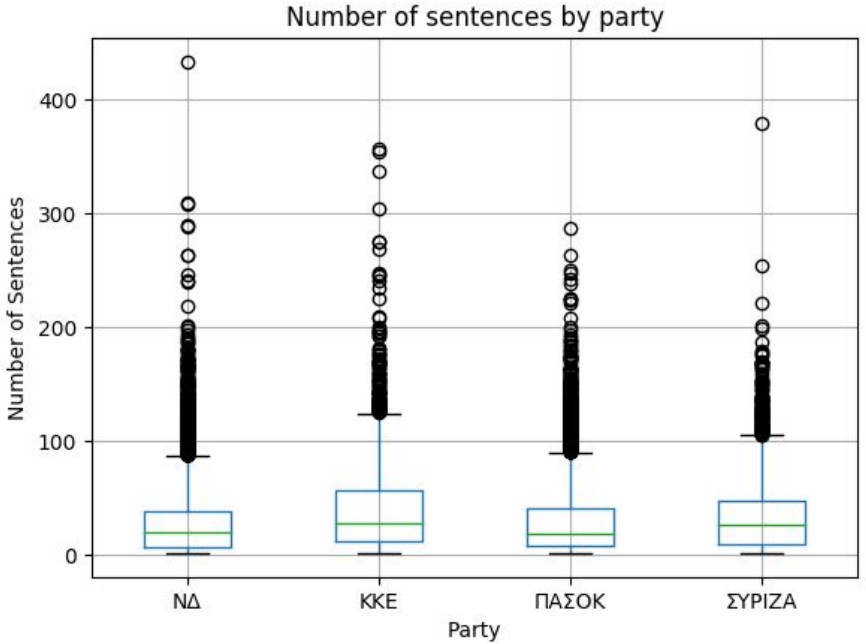
# Στατιστικά προτάσεων

Distribution of number of sentences by party.

| | count | mean | std | min | 25.00% | 50.00% | 75.00% | max |
|---|---|---|---|---|---|---|---|---|
| ΝΔ | 14874 | 27.96 | 27.77 | 2 | 7 | 20 | 39 | 433 |
| ΚΚΕ | 3001 | 39.94 | 38.49 | 2 | 12 | 28 | 57 | 357 |
| ΠΑΣΟΚ | 15233 | 28.43 | 27.47 | 2 | 8 | 19 | 41 | 287 |
| ΣΥΡΙΖΑ | 4431 | 33.61 | 30.39 | 2 | 10 | 26 | 48 | 379 |

# Στατιστικά προτάσεων



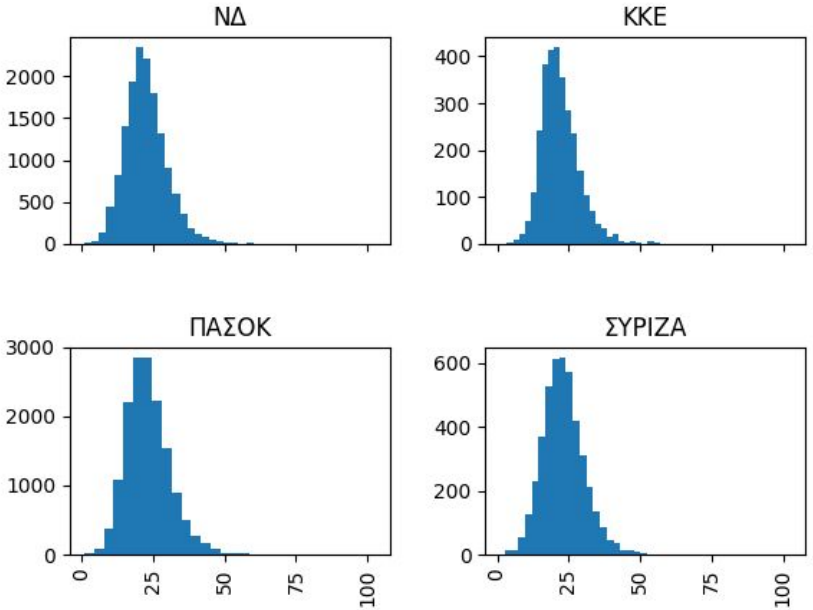



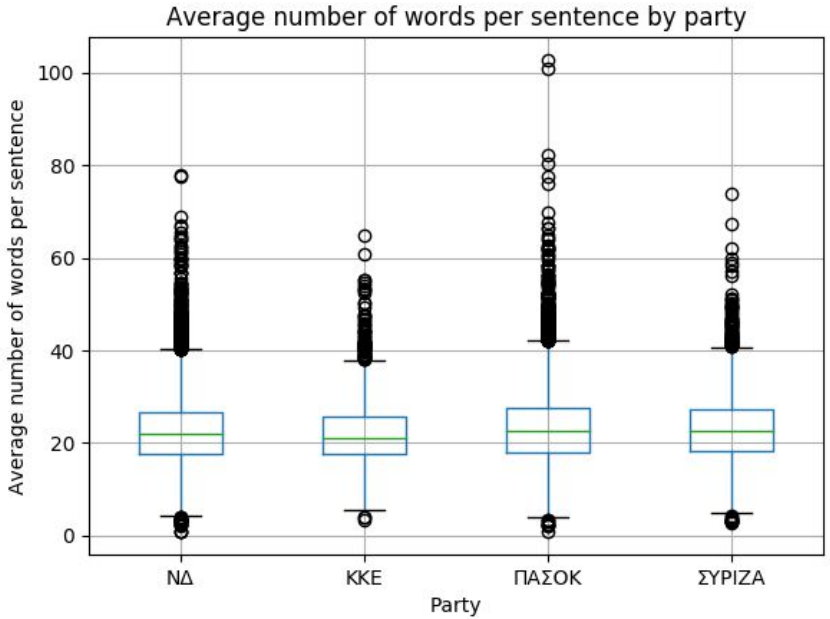
# Στατιστικά λέξεων

Distribution of number of words by party.

| | count | mean | std | min | 25.00% | 50.00% | 75.00% | max |
|---|---|---|---|---|---|---|---|---|
| ΝΔ | 14874 | 636.02 | 636.64 | 2 | 146 | 445 | 932 | 7886 |
| ΚΚΕ | 3001 | 863.55 | 813.02 | 7 | 254 | 604 | 1234 | 7508 |
| ΠΑΣΟΚ | 15233 | 657.95 | 639.37 | 2 | 166 | 464 | 966 | 6363 |
| ΣΥΡΙΖΑ | 4431 | 784.51 | 710.18 | 6 | 202 | 605 | 1147.5 | 8644 |

# Μέσος αριθμός λέξεων ανά πρόταση



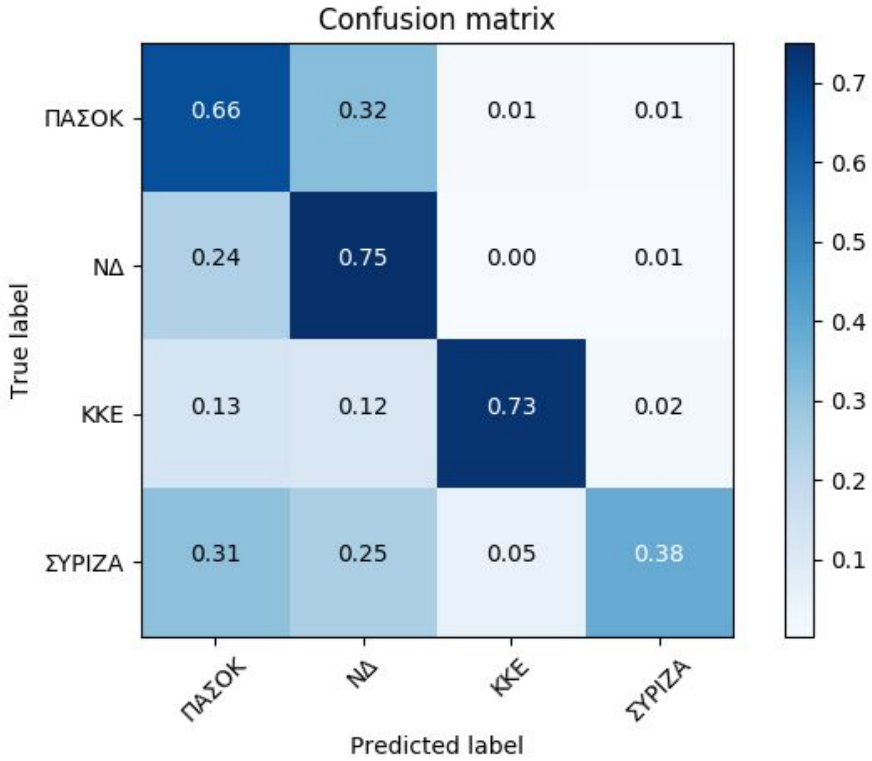
# Προσέγγιση του προβλήματος

SVM & Recurrent Neural Network

# SVM

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| ΠΑΣΟΚ | 0.6528 | 0.6608 | 0.6567 | 3841 |
| ΝΔ | 0.6361 | 0.7489 | 0.6879 | 3747 |
| ΚΚΕ | 0.8316 | 0.7344 | 0.78 | 753 |
| ΣΥΡΙΖΑ | 0.8697 | 0.3833 | 0.5321 | 1114 |
| avg / total | 0.686 | 0.6689 | 0.6642 | 9455 |

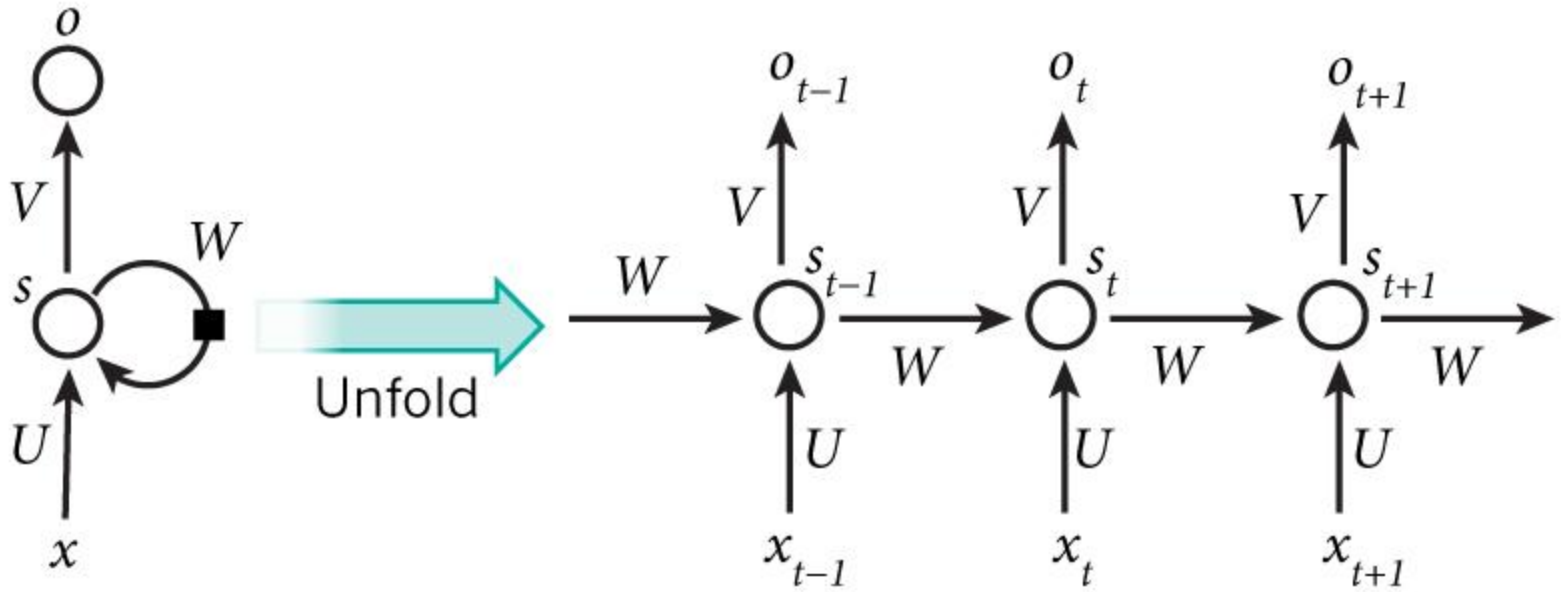
# Confusion Matrix



# Recurrent Neural Network (RNN)

- Ένα νευρωνικό δίκτυο λέγεται αναδρομικό, εάν υπάρχει έστω και μια σύνδεση από έναν νευρώνα επιπέδου i προς έναν νευρώνα επιπέδου j, όπου $j \leq i$.
- Οι αναδρομικές συνδέσεις καθιστούν τα δίκτυα αυτά ικανά να αναγνωρίζουν χρονικές ή και χωρικές συσχετίσεις μεταξύ των δεδομένων.

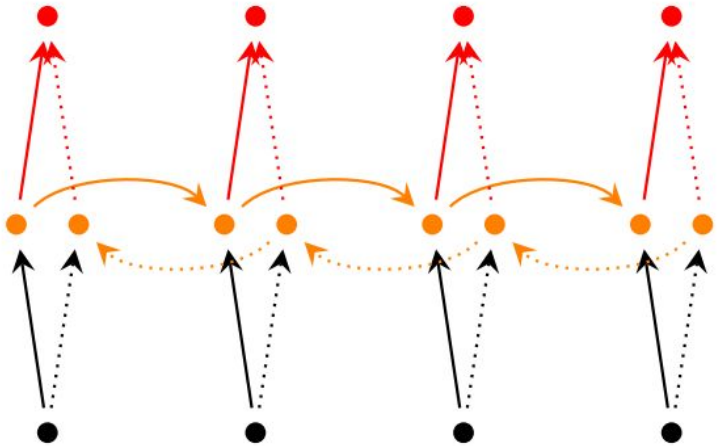
# Recurrent Neural Networks (RNN)




- x: είσοδος, s: κατάσταση (hidden state), o: έξοδος
- Λόγω του ότι η $s_t$ έχει είσοδο και την $s_{t-1}$ λέμε ότι το δίκτυο έχει μνήμη.

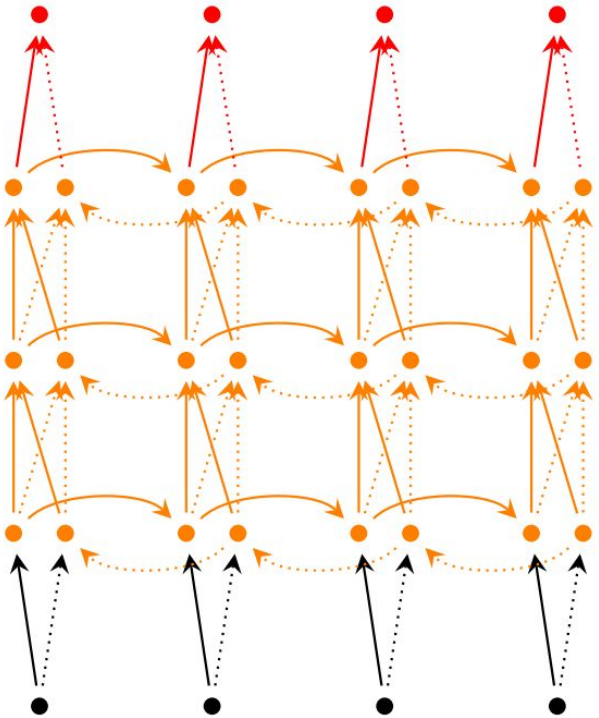
# Δικατευθυνόμενα (Bidirectional) RNN

- Βασίζονται στην ιδέα ότι η έξοδος τη χρονική στιγμή t μπορεί να μη βασίζεται μόνο στις προηγούμενες καταστάσεις αλλά και στις επόμενες.
- Για παράδειγμα, για την πρόβλεψη μιας λέξης που λείπει απ' την ακολουθία θα θέλαμε να ξέρουμε τι προηγείται αλλά και τι έπεται.



# Πολυεπίπεδα δικατευθυνόμενα RNN

- Όμοια με τα δικατευθυνόμενα μόνο που τώρα έχουμε περισσότερα επίπεδα (η έξοδος του ενός RNN γίνεται είσοδος στο επόμενο επίπεδο).
- Μπορούμε να μάθουμε πιο περίπλοκες ακολουθιακές δομές.

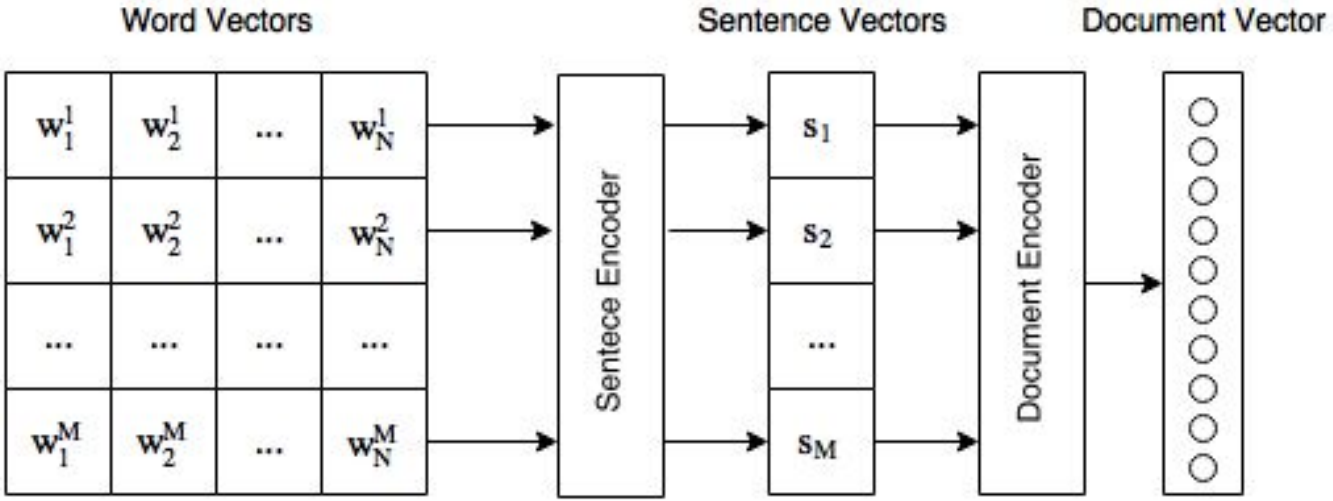


# Αρχιτεκτονική Προτεινόμενου Δικτύου

- Δημιουργία ενός διεπίπεδου δικατευθυνόμενου RNN.
- Το πρώτο επίπεδο έχει είσοδο τα τις λέξεις των προτάσεων.
  - Παράγει διανύσματα τα οποία αντιστοιχούν στις προτάσεις του κειμένου.
- Το δεύτερο επίπεδο έχει είσοδο τις προτάσεις του εγγράφου
  - Παράγει ένα διάνυσμα που αντιπροσωπεύει το έγγραφο.
- Το διάνυσμα εγγράφου χρησιμοποιείται ως feature vector για ταξινόμηση σε μια από τις κλάσεις από ένα layer με softmax activation.

# Αρχιτεκτονική Προτεινόμενου Δικτύου

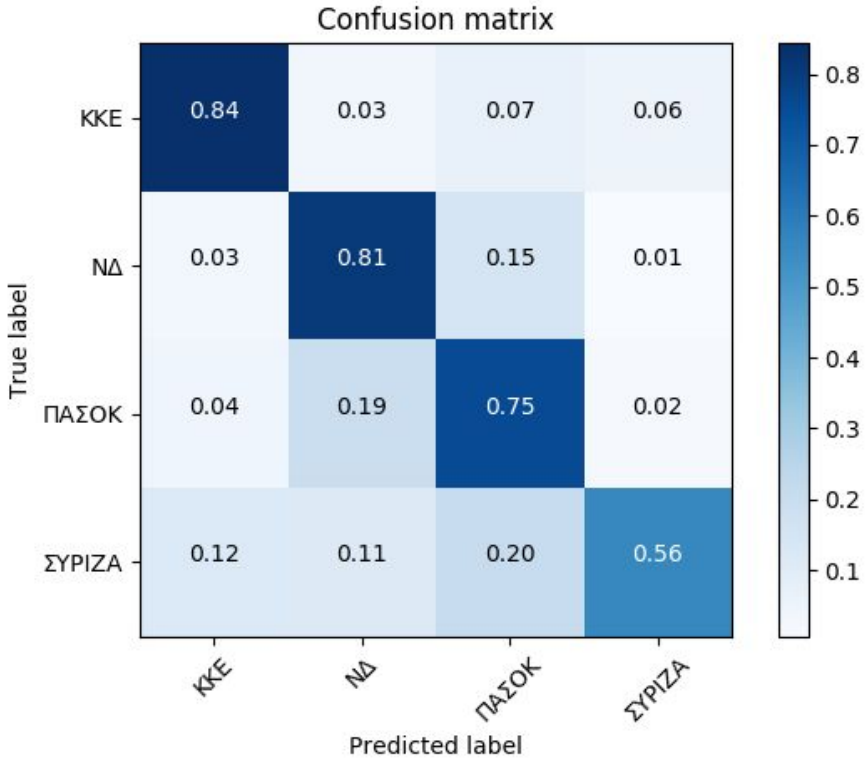
- Η αρχιτεκτονική αυτή περιγράφει τη δομή ενός εγγράφου
  - Οι προτάσεις απαρτίζονται από λέξεις, και
  - Το έγγραφο απαρτίζεται από προτάσεις.



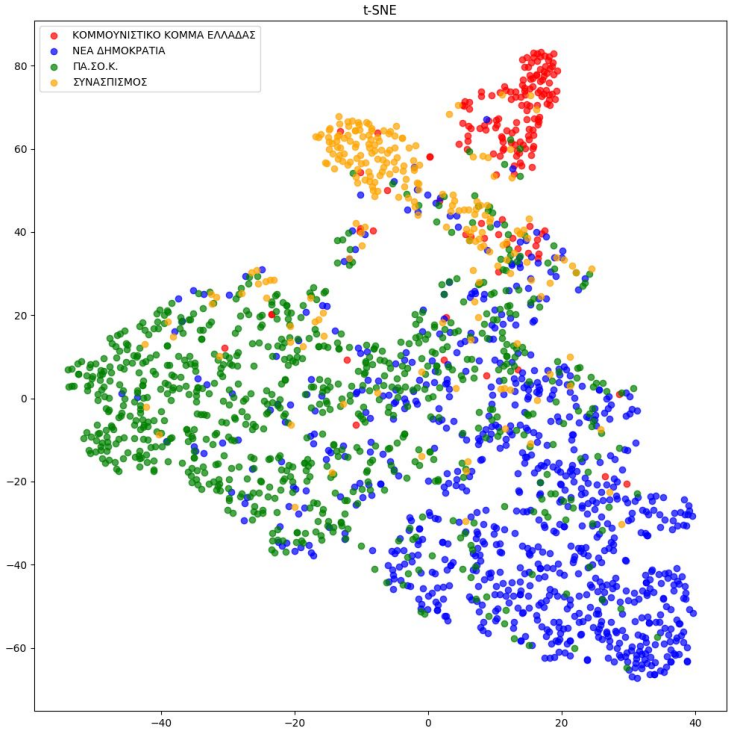
# Αξιολόγηση του προτεινόμενου δικτύου

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| ΚΚΕ | 0.6049 | 0.8429 | 0.7044 | 1248 |
| ΝΔ | 0.7722 | 0.8106 | 0.7909 | 6182 |
| ΠΑΣΟΚ | 0.7761 | 0.7525 | 0.7641 | 6392 |
| ΣΥΡΙΖΑ | 0.8325 | 0.5610 | 0.6703 | 1852 |
| avg/total | 0.7676 | 0.7600 | 0.7589 | 15674 |

# Confusion Matrix



# Γραφική αναπαράσταση



# Εναλλακτική προσέγγιση

Διανύσματα λέξεων με το μοντέλο fastText

# Ταξινόμηση Κειμένου

- Bag of Words + SVM/logistic regression
  - Απώλεια πληροφορίας για τη σειρά των λέξεων
  - Αδυναμία αντιμετώπισης λέξεων εκτός λεξιλογίου
- Νευρωνικά Δίκτυα
  - Συνελικτικά, Αναδρομικά, ή συνδυασμός αρχιτεκτονικών (ενδεικτικά char-CNN, char-RNN, VDCNN, Conv-GRNN)
  - Είσοδος σε επίπεδο χαρακτήρα ή λέξης
- fastText

# fastText (unsupervised)

- CBOW
- Εμπλουτισμός λέξεων με αναπαραστάσεις των N-grams

Ταχύτητα

- Αρνητική Δειγματοληψία (Negative Sampling)
- Ιεραρχικό Softmax
- N-gram hashing

# fastText (unsupervised) N-grams

- Σπάνιες Λέξεις
- Λέξεις εκτός λεξιλογίου (OOV)
- Εγκλίσεις, πτώσεις
- Μορφήματα
- Σύνθετες λέξεις

# fastText (supervised)

- CBOW με την κλάση ως λέξη στόχο
- Εμπλουτισμός λέξεων με αναπαραστάσεις των N-grams

Ταχύτητα

- Αρνητική Δειγματοληψία (Negative Sampling)
- Ιεραρχικό Softmax
- N-gram hashing

# Αναγνώριση Κόμματος

Είσοδος: Ολόκληρη ομιλία

- 76M λέξεις
- 24k μοναδικές (μέγεθος λεξιλογίου)

Επιλογή μετα-παραμέτρων: Grid Search σε

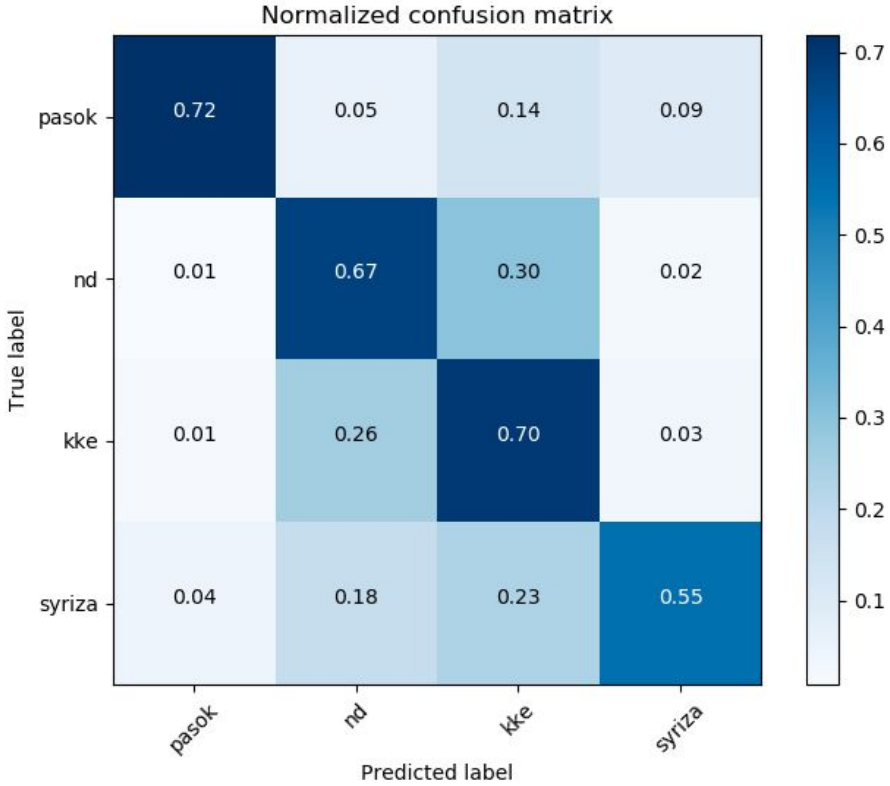
- Learning Rate
- Αριθμό εποχών
- Διάσταση διανύσματος λέξης
- Εμπλουτισμός με bigrams ή trigrams

# Αποτελέσματα

Τελικό μοντέλο (LR: 1.0, epochs: 5, word vector dim: 10, bigrams)

| | precision | recall | f1-score |
|---|---|---|---|
| ΠΑΣΟΚ | 0.51 | 0.70 | 0.59 |
| ΝΔ | 0.58 | 0.67 | 0.62 |
| ΚΚΕ | 0.92 | 0.72 | 0.81 |
| ΣΥΡΙΖΑ | 0.79 | 0.55 | 0.65 |
| Μ.Ο. | 0.70 | 0.66 | 0.67 |

# Confusion Matrix



# Αναγνώριση Κόμματος (Παράγραφοι)

Είσοδος: Παράγραφος ομιλίας

Ομοίως με πριν

- 76M λέξεις
- 24k μοναδικές (μέγεθος λεξιλογίου)

Επιλογή μετα-παραμέτρων: Grid Search σε

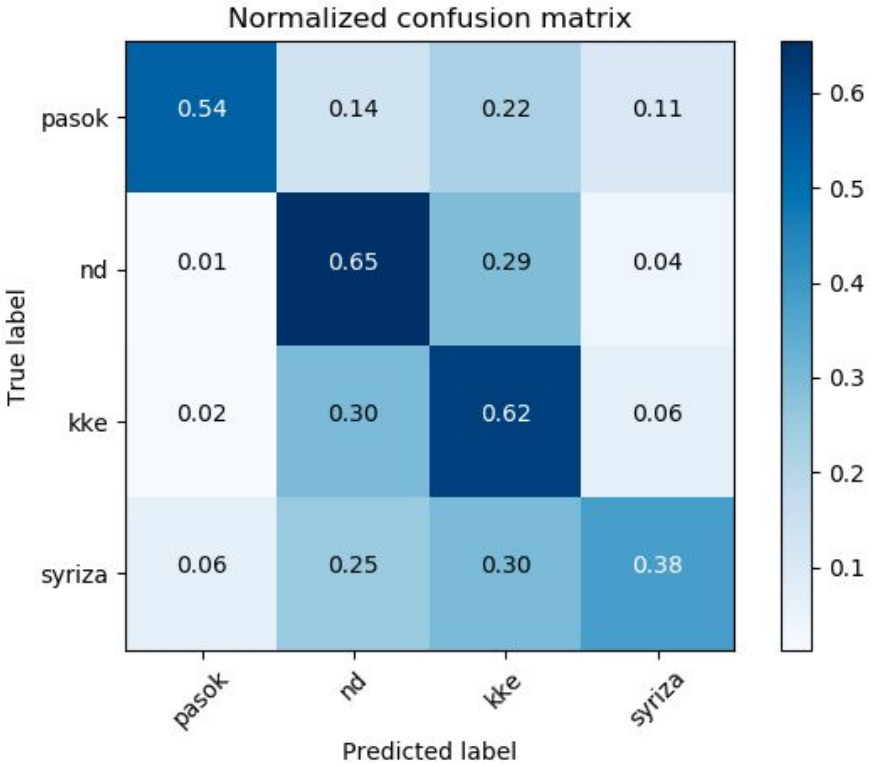
- Learning Rate
- Αριθμό εποχών
- Διάσταση διανύσματος λέξης
- Εμπλουτισμός με bigrams ή trigrams

# Αποτελέσματα

Τελικό μοντέλο (LR: 1.0, epochs: 5, word vector dim: 10, bigrams)

| | precision | recall | f1-score |
|---|---|---|---|
| ΠΑΣΟΚ | 0.43 | 0.62 | 0.51 |
| ΝΔ | 0.49 | 0.65 | 0.56 |
| ΚΚΕ | 0.85 | 0.54 | 0.66 |
| ΣΥΡΙΖΑ | 0.64 | 0.38 | 0.48 |
| Μ.Ο. | 0.60 | 0.55 | 0.55 |

# Confusion Matrix



# Αναγνώριση Ομιλητή

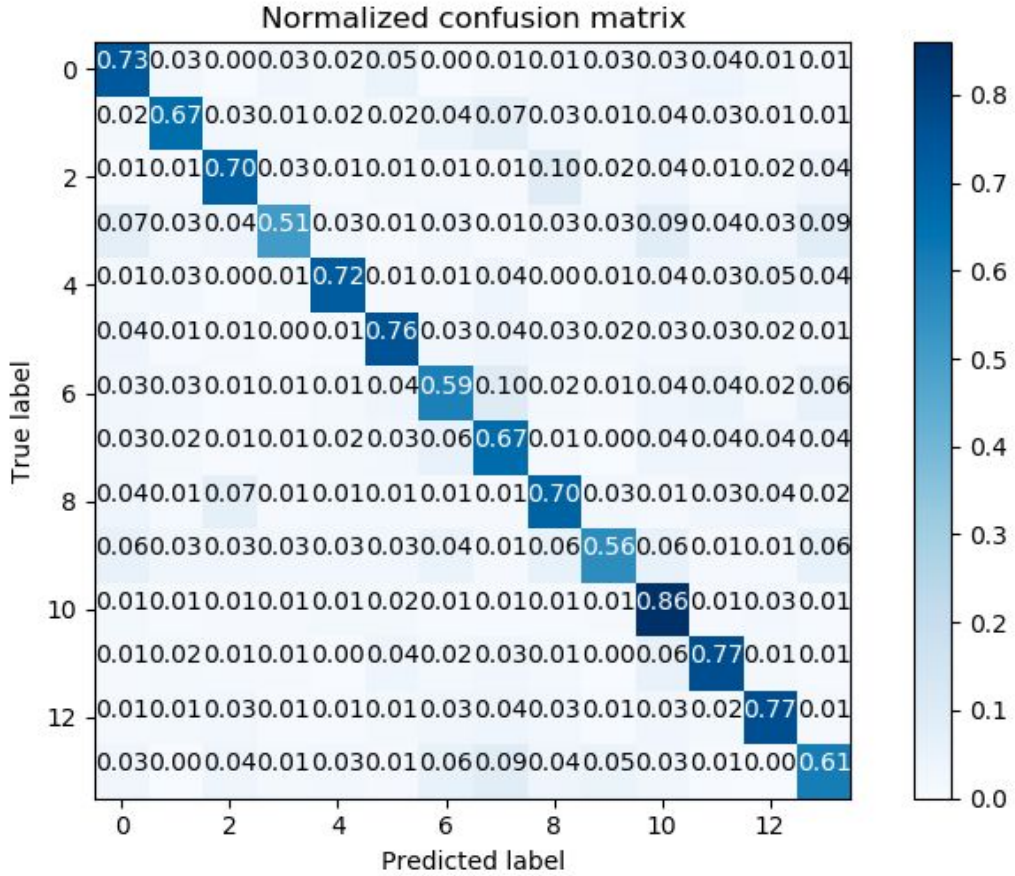
- 14 Βουλευτές με πάνω από 2000 αποσπάσματα (παραγράφους) ομιλιών

1. Παυλόπουλος Προκόπιος Βασιλείου 4638
2. Βενιζέλος Ευάγγελος Βασιλείου 3356
3. Λαφαζάνης Παναγιώτης Γεωργίου 3333
4. Ξηροτύρη Αικατερινάρη Ασημίνα Γεωργίου 3245
5. Σουφλιάς Γεώργιος Αθανασίου 3011
6. Κουβέλης Φώτιος Φανούριος Ευαγγέλου 2954
7. Παπανδρέου Γεώργιος Ανδρέα 2815
8. Σκυλλάκος Αντώνιος Ηλία 2747
9. Λεβέντης Αθανάσιος Σωτηρίου 2604
10. Καραμανλής Κωνσταντίνος Αλεξάνδρου 2447
11. Τζάκρη Θεοδώρα Εμμανουήλ 2277
12. Γείτονας Κωνσταντίνος Ιωάννη 2091
13. Καστανίδης Χαράλαμπος Γεωργίου 2086
14. Παπουτσής Χρήστος Δημητρίου 2010

# Αποτελέσματα

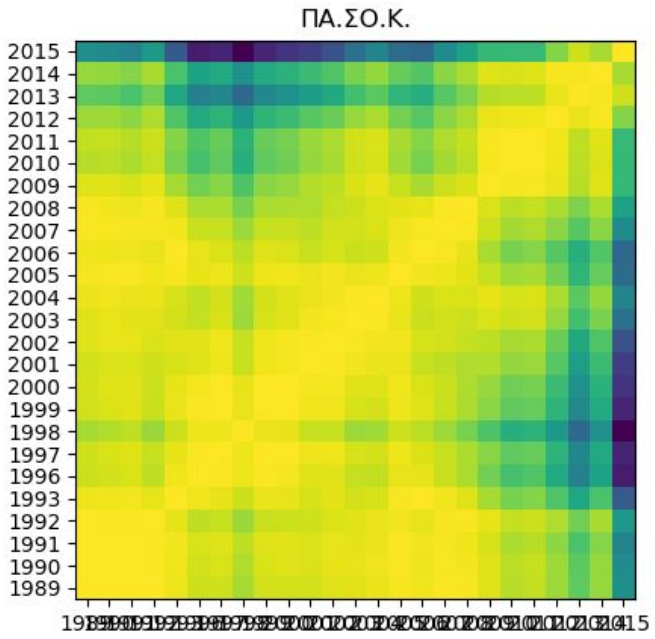
- Accuracy 0.69
- Precision 0.69
- Recall 0.688
- F1-score 0.688

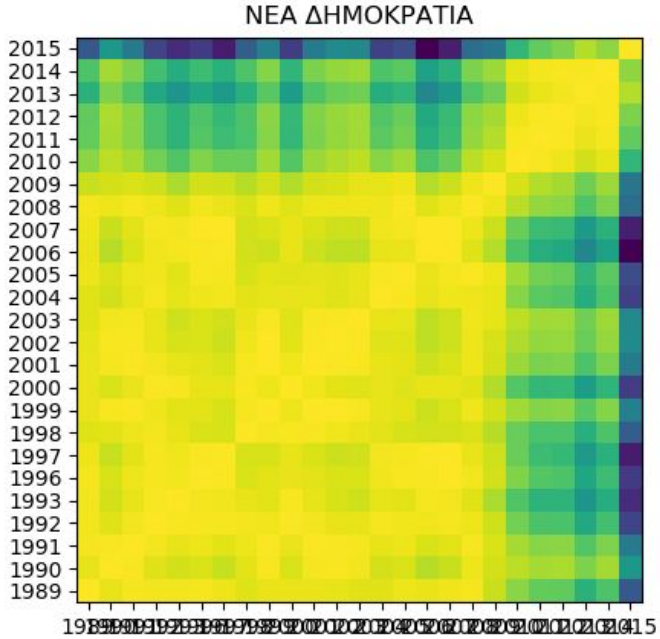
# Confusion Matrix



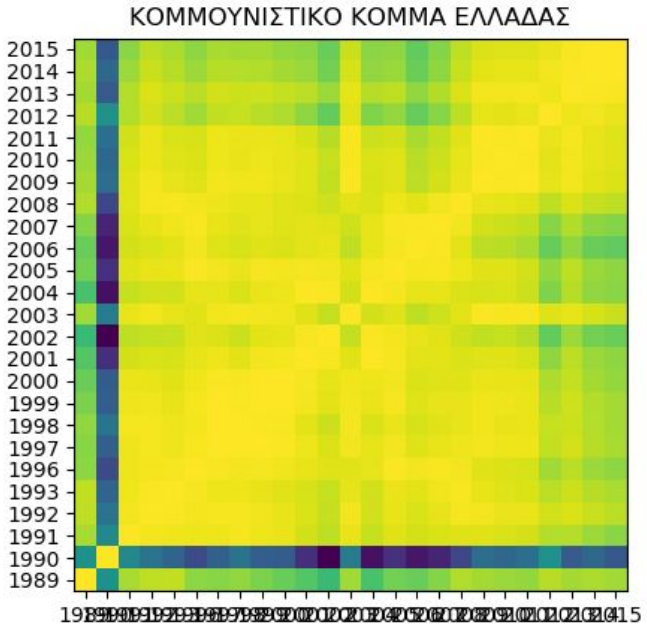
# Διερεύνηση σημασιολογικών αποστάσεων μεταξύ ομιλιών

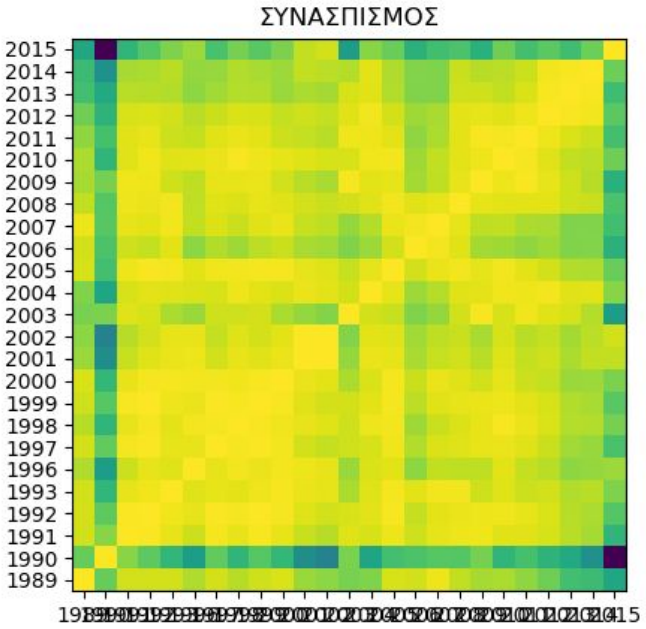
# Ομοιότητα ομιλιών κόμματος ανά έτος



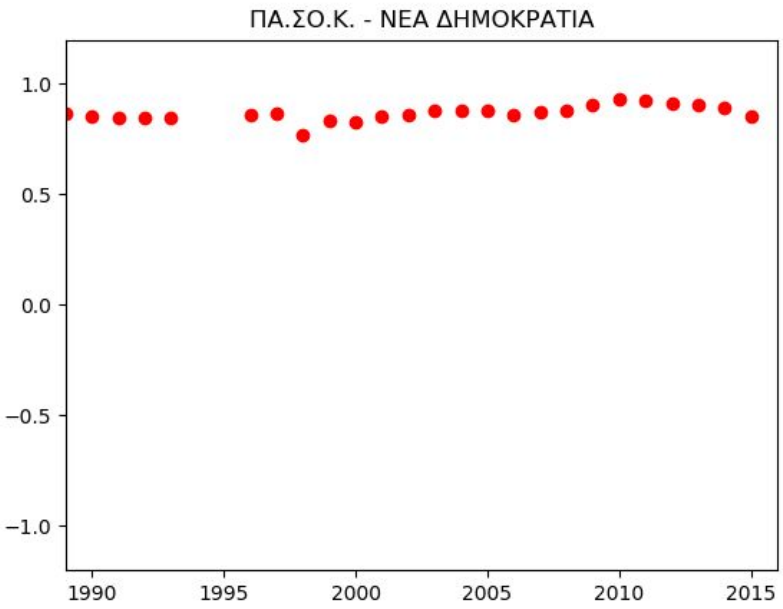



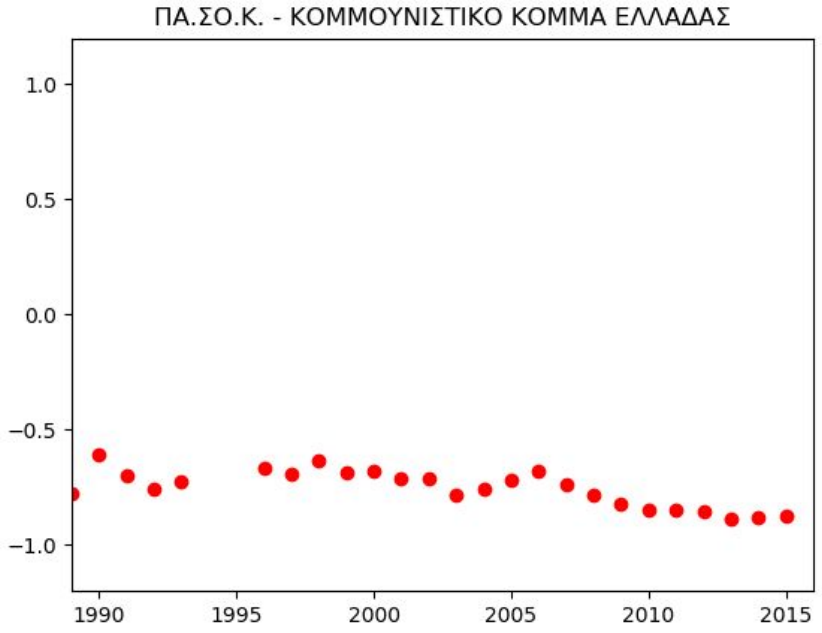
# Ομοιότητα ομιλιών κόμματος ανά έτος



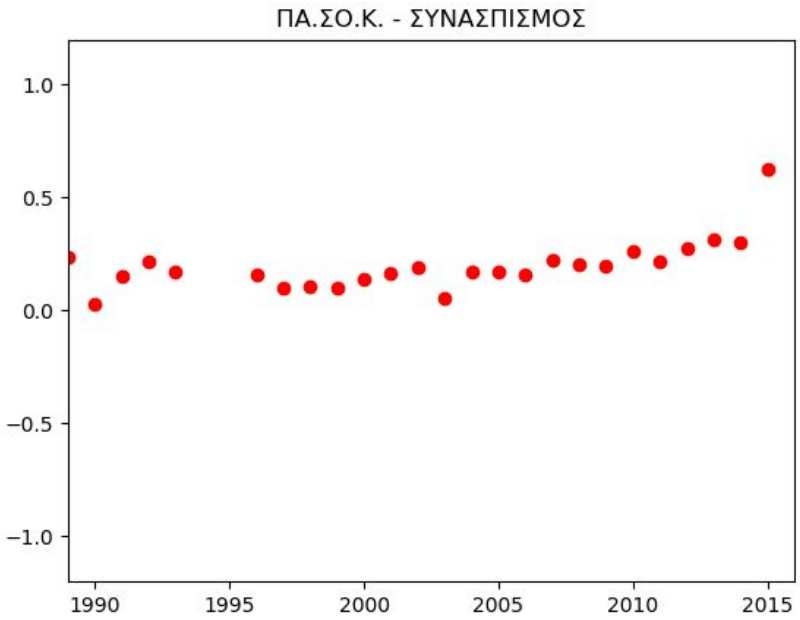



# Ομοιότητα ομιλιών ανά ζεύγος κομμάτων



# Ομοιότητα ομιλιών ανά ζεύγος κομμάτων




# Ομοιότητα ομιλιών ανά ζεύγος κομμάτων