

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ  
Τμήμα Επιστήμης Υπολογιστών

ΗΥ-317: Εφαρμοσμένες Στοχαστικές Διαδικασίες - Εαρινό Εξάμηνο 2025  
Διδάσκων: Π. Τσακαλίδης

Προγραμματιστική άσκηση στις Μαρκοβιανές Αλυσίδες

Ημερομηνία Ανάθεσης: 30/04/2025

Ημερομηνία Παράδοσης: 20/06/2025

**ΣΗΜΑΝΤΙΚΕΣ ΟΔΗΓΙΕΣ:** Στο πλαίσιο αυτής της προγραμματιστικής άσκησης θα πρέπει να παραδώσετε ένα συμπίεμένο αρχείο (.rar, .zip κτλ.) το οποίο θα περιλαμβάνει όλα τα αρχεία πηγαίου κώδικα (.m, .py) που έχετε υλοποιήσει, καθώς και το PDF της αναφοράς σας. Συμπεριλάβετε στο όνομα του αρχείου και στην αναφορά σας το ονοματεπώνυμό σας, τον ΑΜ σας και το τμήμα σας. **Δεν χρειάζεται να λύσετε αναλυτικά κάτι στο χαρτί, εκτός αν σας ζητείται.** Σε περίπτωση αντιγραφής (μικρής ή μεγάλης εμβέλειας) ολόκληρη η βονυς άσκηση θα μηδενίζεται, για όλους τους εμπλεκόμενους. Η υλοποίησή σας μπορεί να γίνει είτε με τη χρήση *MATLAB* είτε με τη χρήση *Python*. Η παράδοση της άσκησης θα γίνει ηλεκτρονικά, **ΑΠΟΚΛΕΙΣΤΙΚΑ** μέσω της σελίδας *e – learn* του μαθήματος, **εδώ**.

## Περιγραφή

Σε αυτήν την άσκηση θα προγραμματίσετε έναν γεννήτορα κειμένου (*text generator*) βασιζόμενο σε ένα Μαρκοβιανό μοντέλο. Το μοντέλο σας θα εκπαιδευτεί σε ένα δείγμα κειμένου, και στη συνέχεια θα αξιοποιηθεί ώστε να παράγει τυχαία λέξεις, η πιθανότητα των οποίων εξαρτάται μόνο από την προηγούμενη λέξη (Μαρκοβιανή ιδιότητα). Το αποτέλεσμα θα είναι η παραγωγή ενός κειμένου το οποίο με μια πρώτη ματιά θα μοιάζει αληθινό αφού δύο γειτονικές λέξεις θα μπορούν να συναντηθούν και σε ένα πραγματικό κείμενο, όμως στο σύνολό του δεν θα βγάζει κανένα νόημα! Τα βήματα που θα ακολουθήσετε δίνονται παρακάτω.

1. Διαβάζετε το δείγμα κειμένου και εξαγάγετε τις λέξεις που το απαρτίζουν. Μας ενδιαφέρει η ακολουθία των λέξεων, και συγκεκριμένα τα ζευγάρια λέξεων που είναι γειτονικά μέσα στο κείμενο.
2. Το μοντέλο μας αποτελείται από ένα Λεξικό, το οποίο θα περιέχει μία καταχώρηση για κάθε διαφορετική λέξη που υπάρχει μέσα στο κείμενο. Ονομάζουμε την λέξη που αφορά την καταχώρηση «κλειδί». Κάθε τέτοια καταχώρηση θα συνοδεύεται από τις λέξεις που ακολουθούν την λέξη-κλειδί μέσα στο κείμενο. Αυτές τις λέξεις τις ονομάζουμε «ακόλουθους». Για παράδειγμα, με την πρόταση «είπα ναι, μετά είπα όχι», θα πρέπει να υπάρχουν καταχωρήσεις για τις λέξεις «είπα», «ναι,», «μετά», «όχι». Οι καταχωρήσεις θα πρέπει να μπορούν να μας δώσουν την εξής πληροφορία:

«είπα» → «ναι,» , «όχι»

«ναι,» → «μετά»

«μετά» → «είπα»

3. Το λεξικό θα πρέπει να μας δίνει την δεσμευμένη πιθανότητα του ακόλουθου, δεδομένου του κλειδιού. Υπάρχουν δύο προφανείς τρόποι να το κάνουμε αυτό: α) Να συνοδεύουμε κάθε ακόλουθο με το πλήθος των φορών που έχουν συναντηθεί ως ακόλουθοι στο συγκεκριμένο κλειδί μέσα στο κείμενο. β) Να καταχωρούμε τον ακόλουθο όσες φορές τον έχουμε συναντήσει ως ακόλουθο του κλειδιού μέσα στο κείμενο. Για παράδειγμα, με την πρόταση «είπα ναι και μετά είπα ναι», το λεξικό θα δομείται ως εξής με τους δύο τρόπους:

α)

«είπα» → «ναι» (2)

«ναι» → «και» (1)

«και» → «μετά» (1)

«μετά» → «είπα» (1)

β)

«είπα» → «ναι», «ναι»

«ναι» → «και»

«και» → «μετά»

«μετά» → «είπα»

Για μεγάλα κείμενα, ο (α) τρόπος εξοικονομεί περισσότερη μνήμη, όμως ο (β) τρόπος είναι απλούστερος στην υλοποίηση. Υλοποιήστε όποια εκδοχή προτιμάτε.

Παρατηρήστε ότι και με τους δύο τρόπους μπορούμε να βρούμε την πιθανότητα κάθε λέξη-ακόλουθος να ακολουθήσει κάθε λέξη-κλειδί. Συγκεκριμένα, αν  $F_k$  είναι το σύνολο ακολούθων της λέξης  $k$ , με τον (α) τρόπο, η πιθανότητα η λέξη  $u_j$  να ακολουθεί την λέξη  $k$  είναι

$$P(u_j | k) = \frac{\#(u_j | k)}{\sum_{u_i \in F_k} \#(u_i | k)}, \quad (1)$$

όπου  $\#(u_i | k)$  είναι το πλήθος των φορών που η  $u_i$  ακολουθεί την  $k$ .

Με τον (β) τρόπο, η πιθανότητα είναι

$$P(u_j | k) = \frac{1}{|F_k|}, \quad (2)$$

όπου  $|F_k|$  είναι ο πληθάριθμος του  $F_k$ .

Προγραμματιστικά, η επιλογή της επόμενης λέξης με τον (α) τρόπο απαιτεί τον υπολογισμό της δεσμευμένης πιθανότητας, ενώ με τον (β) τρόπο απλώς επιλέγετε μία τυχαία λέξη από την λίστα των ακολούθων.

4. Έχοντας κατασκευάσει το λεξικό, η παραγωγή κειμένου είναι πολύ απλή. Ξεκινήστε με μία λέξη με κεφαλαίο το πρώτο γράμμα, και κατόπιν συνεχίστε να διαλέγετε τυχαία την επομένη λέξη από τις λέξεις-ακόλουθους της προηγούμενης. Η επιλογή της επόμενης λέξης πρέπει να γίνεται σύμφωνα με την δεσμευμένη πιθανότητα.

## Οδηγίες για την υλοποίηση

- Την υλοποίηση μπορείτε να την κάνετε είτε σε *MATLAB* είτε σε *Python*. Δίνονται βοηθητικά αντίστοιχα αρχεία, τα οποία περιέχουν και βοηθητικά σχόλια. Σαν δείγμα κειμένου, σας δίνεται ένα αρχείο το οποίο περιέχει μία συλλογή παραμυθιών. Μπορείτε να πειραματιστείτε με δικά σας κείμενα.
- Η αναφορά σας πρέπει να περιέχει 1 παράδειγμα κειμένου που παρήχθη από τον αλγόριθμό σας, μήκους 100 λέξεων, καθώς και τα σχόλιά σας που αφορούν την υλοποίηση του αλγορίθμου και της ποιότητας των αποτελεσμάτων. Δεν χρειάζεται να αναλύσετε βήμα-προς-βήμα τον αλγόριθμο, απλώς αναφέρετε τι δυσκολίες συναντήσατε και πώς τις αντιμετωπίσατε.