Αν. Καθηγητής Π. Λουρίδας

Τμήμα Διοικητικής Επιστήμης και Τεχνολογίας

Οικονομικό Πανεπιστήμιο Αθηνών

Ευγένιος Όνεγκιν, Αλυσίδες Μάρκοφ, Πατρότητα Κειμένου (και Λίγη Εντροπία στο Τέλος)

Πριν προχωρήσουμε στην εκφώνηση, μια λεκτική παρατήρηση. Χρησιμοποιούμε τον όρο «πατρότητα κειμένου», αφού αυτός έχει ιστορικά επικρατήσει. Βεβαίως δεν υπάρχει λόγος να μη μιλάμε για «μητρότητα κειμένου», ή να προταθεί κάποιος ουδέτερος όρος.

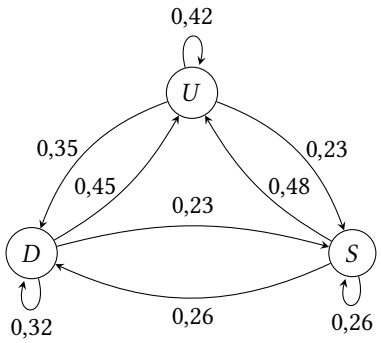
Το 1913 ο Ρώσος μαθηματικός Αντρέι Αντρέγιεβιτς Μάρκοφ (1856–1922) παρουσίασε στη Βασιλική Ακαδημία της Αγίας Πετρούπολης μια στατιστική μελέτη του κειμένου του μυθιστορήματος σε στίχους «Ευγένιος Όνεγκιν» του Αλεξάντερ Σεργκέγιεβιτς Πούσκιν. Ο Μάρκοφ πήρε τα πρώτα 20.000 γράμματα του ποιήματος και μέτρησε τις εμφανίσεις φωνηέντων. Με βάση αυτές, υπολόγισε τις παρακάτω πιθανότητες:

- $p(\Phi)$: η πιθανότητα ένα γράμμα να είναι φωνήεν.
- $p(\Sigma)$: η πιθανότητα ένα γράμμα να είναι σύμφωνο.
- $p(\Phi\Phi)$: η πιθανότητα να έχουμε δύο φωνήεντα στη σειρά.
- $p(\Sigma\Phi)$: η πιθανότητα να έχουμε ένα σύμφωνο και ένα φωνήεν στη σειρά.
- $p(\Phi\Phi\Phi)$: η πιθανότητα να έχουμε φωνήεν, φωνήεν, φωνήεν στη σειρά.
- $p(\Phi\Sigma\Phi)$: η πιθανότητα να έχουμε φωνήεν, σύμφωνο, φωνήεν στη σειρά.
- $p(\Sigma\Phi\Phi)$: η πιθανότητα να έχουμε σύμφωνο, φωνήεν, φωνήεν στη σειρά.
- $p(\Sigma\Sigma\Phi)$: η πιθανότητα να έχουμε σύμφωνο, σύμφωνο, φωνήεν στη σειρά.

Ο Μάρκοφ βρήκε ότι στα 20.000 γράμματα τα 8.638 ήταν φωνήεντα και τα 11.362 ήταν σύμφωνα, άρα η πιθανότητα $p(\Phi) = \frac{8638}{20000} \approx 0.43$. Αν τώρα το κατά πόσο ένα γράμμα είναι φωνήεν ή σύμφωνο δεν εξαρτάται από το προηγούμενο γράμμα, η πιθανότητα να έχουμε δύο φωνήεντα στη σειρά, $p(\Phi\Phi)$ θα ήταν $(0.43)^2 \approx 0{,}19$. Αφού έχουμε 20.000 γράμματα, άρα 19.999 ζευγάρια γραμμάτων, θα περιμέναμε να βρούμε $(\frac{8638}{20000})^2 \times 19.999 = 3.731$ φωνήεντα στη σειρά. Πλην όμως, στον Ευγένιο Όνεγκιν αυτό δεν συμβαίνει. Στα πρώτα 20.000 γράμματα υπάρχουν 1.104 ζεύγη φωνηέντων, το ένα τρίτο δηλαδή από αυτά που θα περιμέναμε αν το κάθε γράμμα δεν εξαρτώταν από το προηγούμενό του. Έτσι, $p(\Phi\Phi) = \frac{1104}{8638} \approx 0.13 \neq 0.19$. Αυτό δεν είναι τυχαίο. Στα Ρώσικα, όπως και σε άλλες γλώσσες, τα φωνήεντα τείνουν να εναλλάσσονται με σύμφωνα. Οι μετρήσεις του Μάρκοφ για τις υπόλοιπες πιθανότητες επίσης επιβεβαίωσαν ότι τα γράμματα δεν εμφανίζονται ανεξάρτητα το ένα από τα άλλα στις λέξεις, αλλά κάθε γράμμα επηρεάζει το επόμενό του.

Η ενασχόληση του Μάρκοφ με τον Ευγένιο Όνεγκιν δεν ήταν ουρανοκατέβατη. Ο Μάρκοφ είχε ξεκινήσει, μερικά χρόνια πριν, να ερευνά φαινόμενα όπου ενδεχόμενα συμβαίνουν στη σειρά το ένα μετά το άλλο, και η εμφάνιση ενός συγκεκριμένου ενδεχομένου επηρεάζει το επόμενο ενδεχόμενο. Ας πάρουμε ως παράδειγμα την πρόβλεψη του καιρού. Έστω ότι έχουμε τρεις δυνατές προγνώσεις: ήλιος, συννεφιά, βροχή. Αν μια μέρα έχει ήλιο, η πιθανότητα την επόμενη μέρα να έχει συννεφιά είναι διαφορετική από την πιθανότητα την επόμενη ημέρα να έχει βροχή. Ομοίως, αν μια μέρα έχει συννεφιά, η πιθανότητα την επόμενη ημέρα να έχει βροχή είναι διαφορετική από την πιθανότητα την επόμενη ημέρα να έχει ήλιο. Γενικότερα, ξέρουμε ότι ο καιρός που θα κάνει αύριο σχετίζεται με τον καιρό που κάνει σήμερα—δεν έχουμε δύο ανεξάρτητα γεγονότα.

Ένα τέτοιο σύστημα μπορούμε να το απεικονίσουμε με ένα διάγραμμα που δείχνει τις καταστάσεις και τις μεταβάσεις μεταξύ τους· στην πραγματικότητα ένα γράφο. Στην παρακάτω εικόνα μπορείτε να δείτε έναν τέτοιο γράφο. Ο γράφος αυτός απεικονίζει τις εβδομαδιαίες καταστάσεις και μετακινήσεις του δείκτη τιμών Dow Jones στο χρηματιστήριο της Νέας Υόρκης, από τη δημιουγία του δείκτη, το 1885, μέχρι μια μέρα του Σεπτέμβρη 2021. Διακρίνουμε τρεις καταστάσεις: άνοδος του δείκτη τιμών στην εβδομάδα σε ποσοστό άνω του $0,5\%$ (κόμβος U), πτώση του δείκτη τιμών στην εβδομάδα σε ποσοστό μεγαλύτερο του $-0,5\%$ (κόμβος D), και στασιμότητα (κόμβος S), όπου ο δείκτης τιμών στο τέλος της εβδομάδας έκλεισε με διαφορά μεταξύ $-0,5\%$ και $0,5\%$ σε σχέση με την τιμή ανοίγματος στην αρχή της εβδομάδας.



Αν μελετήσουμε ιστορικά τις κινήσεις του χρηματιστηρίου, προκύπτει ότι:

- Αν είμαστε σε εβδομάδα ανόδου, η επόμενη εβδομάδα με πιθανότητα 42% θα είναι εβδομάδα ανόδου, με πιθανότητα 23% θα είναι εβδομάδα στασιμότητας, και με πιθανότητα 35% θα είναι εβδομάδα καθόδου.
- Αν είμαστε σε εβδομάδα καθόδου, η επόμενη εβδομάδα με πιθανότητα 32% θα είναι εβδομάδα καθόδου, με πιθανότητα 45% θα είναι εβδομάδα ανόδου, και με πιθανότητα 23% θα είναι εβδομάδα στασιμότητας.
- Αν είμαστε σε εβδομάδα στασιμότητας, η επόμενη εβδομάδα με πιθανότητα 26% θα είναι εβδομάδα στασιμότητας, με πιθανότητα επίσης 26% θα είναι εβδομάδα καθόδου, και με πιθανότητα 48% θα είναι εβδομάδα ανόδου.

Αν λοιπόν βρισκόμαστε σε μία εβδομάδα και θέλουμε να προβλέψουμε την κίνηση του χρηματιστηρίου την επόμενη εβδομάδα, μπορούμε να χρησιμοποιήσουμε τον παραπάνω γράφο και να ακολουθήσουμε το σύνδεσμο με την μεγαλύτερη πιθανότητα από τον κόμβο που αντιστοιχεί στην τρέχουσα εβδομάδα. Αν θέλουμε να προβλέψουμε την κίνηση του χρηματιστηρίου δύο εβδομάδες από σήμερα, θα ακολουθήσουμε ένα μονοπάτι δύο συνδέσμων, ακολουθώντας κάθε φορά το σύνδεσμο με το μεγαλύτερο βάρος (δηλαδή πιθανότητα). Μπορούμε να προχωρήσουμε έτσι πιο βαθιά στο μέλλον, ακολουθώντας μακρύτερα μονοπάτια που αντιστοιχούν σε αλυσίδες γεγονότων.

Ένα τέτοιο σύστημα το οποίο περιγράφει ένα σύνολο καταστάσεων και τις πιθανότητες μετάπτωσης από τη μία κατάσταση σε κάθε άλλη κατάσταση ονομάζεται *αλυσίδα Μάρκοφ* (Markov chain). Οι αλυσίδες Μάρκοφ έχουν πλήθος εφαρμογών, από την αναγνώριση γονιδίων στο DNA μέχρι τις μηχανές αναζήτησης. Εμείς εδώ θα δούμε πώς μπορούν να χρησιμοποιηθούν για να αποδώσουμε την πατρότητα ενός κειμένου.

Μία αλυσίδα Μάρκοφ την αναπαριστούμε με έναν πίνακα, ο οποίος ονομάζεται πίνακας μεταπτώσεων (transition matrix). Ο πίνακας αυτός έχει τόσες γραμμές και τόσες στήλες όσες οι καταστάσεις, και οι τιμές στα κελιά του πίνακα είναι οι πιθανότητες μετάβασης από μία κατάσταση σε μία άλλη κατάσταση. Αν το δούμε διαφορετικά, ο πίνακας μετάπτωσης δεν είναι άλλος από τον πίνακα γειτνίασης του γράφου που περιγράφει τις καταστάσεις και τις μεταπτώσεις του συστήματος. Ο γράφος είναι πλήρης, άρα όλες οι θέσεις του πίνακα έχουν έναν αριθμό και όλοι οι αριθμοί είναι μεταξύ του 0 και του 1. Το άθροισμα σε κάθε γραμμή και είναι πάντα 1, αφού κάθε γραμμή περιγράφει το σύνολο των μεταπτώσεων από την κατάσταση της γραμμής σε οποιαδήποτε άλλη κατάσταση. Ο παρακάτω πίνακας μεταπτώσεων T αντιστοιχεί στο παράδειγμα του Dow Jones:

	U	D	S
U	0,42	0,35	0,23
D	0,45	0,32	0,23
S	0,48	0,26	0,26

Βασικό χαρακτηριστικό των αλυσίδων Μάρκοφ είναι ότι η πιθανότητα μετάβασης σε μία κατάσταση εξαρτάται *μόνο από την προηγούμενη κατάσταση* στην οποία βρίσκεται στο σύστημα. Άρα οι αλυσίδες Μάρκοφ δεν μπορούν να μοντελοποίησουν συστήματα στα οποία, για παράδειγμα, αυτό που θα συμβεί αύριο εξαρτάται και από το σήμερα και από το χτες. Ταυτόχρονα όμως, η απλότητα της αναπαράστασης του συστήματος μέσω ενός πίνακα μας δίνει τη δυνατότητα να μελετήσουμε την εξέλιξή του πολύ εύκολα.

Πράγματι, αν θέλουμε να δούμε την πιθανότητα να βρισκόμαστε σε μία κατάσταση μετά από δύο μεταπτώσεις, αρκεί να πολλαπλασιάσουμε τον πίνακα μεταπτώσεων με τον εαυτό του, να τον υψώσουμε δηλαδή στη δεύτερη δύναμη. Έστω ότι είμαστε στην κατάσταση D. Τότε η πιθανότητα την επόμενη εβδομάδα να είμαστε σε κάθε μία από τις καταστάσεις U, D, S, δίνεται από τη δεύτερη γραμμή του πίνακα. Αν θέλουμε να βρούμε την πιθανότητα μετά από δύο εβδομάδες να βρισκόμαστε από

την κατάσταση D στην κατάσταση U, αυτή προκύπτει αν υπολογίσουμε:

$$T_{D,U} \times T_{U,U} + T_{D,D} \times T_{D,U} + T_{D,S} \times T_{S,U}$$

Αυτή η παράσταση δεν είναι παρά η μαθηματική απεικόνιση των μονοπατιών με δύο συνδέσμους από την κατάσταση D στην κατάσταση U. Παρατηρούμε ότι είναι το άθροισμα των γινομένων των στοιχείων της γραμμής D με τη στήλη U, δηλαδή το στοιχείο $\mathrm{T}_{U,D}$ του τετραγώνου του πίνακα. Αυτό ακριβώς υπολογίζεται από το τετράγωνο του πίνακα. Με τον ίδιο τρόπο πάμε από το τετράγωνο στην τρίτη δύναμη. Αν θέλουμε να δούμε την πιθανότητα να βρισκόμαστε σε μία κατάσταση μετά από τρεις μεταπτώσεις, αρκεί να τον υψώσουμε στην τρίτη δύναμη, κ.ο.κ. Μάλιστα, όσο πιο μακριά προβλέπουμε στο μέλλον, σε τόσο μεγαλύτερη δύναμη υψώνουμε τον πίνακα, έως σε κάποια στιγμή οι τιμές του πίνακα σταματούν να αλλάζουν και όλες οι γραμμές του είναι ίσες. Οι τιμές του πίνακα τότε μας δίνουν την μακροπρόθεσμη πρόβλεψη για το φαινόμενο που εξετάζουμε, η οποία μάλιστα δεν εξαρτάται από την κατάσταση που ξεκινήσαμε.

Ένας τέτοιος πίνακας, όπου όλα τα στοιχεία του είναι μεταξύ του 0 και του 1 και όπου το άθροισμα σε κάθε γραμμή είναι ίσο με 1 ονομάζεται *στοχαστικός πίνακας*. Ένας τέτοιος στοχαστικός πίνακας χρησιμοποιείται στον αλγόριθμο PageRank για τον υπολογισμό της σημαντικότητας των σελίδων στο παγκόσμιο ιστό. Κάτω από αυτό το πρίσμα, ο αλγόριθμος PageRank είναι μια εφαρμογή των αλυσίδων Μάρκοφ.

Ας επιστρέψουμε τώρα στο πρόβλημα που θέλουμε να αντιμετωπίσουμε, αυτό του προσδιορισμού της πατρότητας ενός κειμένου. Θα το κάνουμε αυτό λαμβάνοντας υπόψη μόνο την πιθανότητα ένα γράμμα να ακολουθεί ένα άλλο γράμμα.

Ξεκινάμε φτιάχνοντας για κάθε κείμενο ένα πίνακα με 26 γραμμές και 26 στήλες, όσοι οι χαρακτήρες του αλφαβήτου μας. Στο κελί i, j θα βρίσκεται το πλήθος των μεταπτώσεων μέσα στο κείμενό μας από το γράμμα i του αλφαβήτου στο γράμμα j του αλφαβήτου. Αν έχουμε $w_1, w_2, ... w_n$ συγγραφείς, για κάθε κείμενο t ενός συγγραφέα w θα συμβολίζουμε τον πίνακα αυτό με Q^{wt}.

Από τους πίνακες Q^{wt} κάθε κειμένου κάθε συγγραφέα, μπορούμε να υπολογίσουμε ένα συνολικό πίνακα με το πλήθος των μεταπτώσεων για όλα τα κείμενα του εν λόγω συγγραφέα, απλώς παίρνοντας το άθροισμα όλων των πινάκων του. Αυτόν τον πίνακα θα τον συμβολίσουμε με Q^w και κάθε στοιχείο του προκύπτει ως εξής:

$$Q^w_{i,j} = \sum_t Q^{wt}_{i,j}$$

όπου το t διατρέχει όλα τα κείμενα του συγγραφέα w.

Αν τώρα διαιρέσουμε κάθε στοιχείο του πίνακα Q^w με το άθροισμα των στοιχείων της γραμμής του, θα πάρουμε έναν πίνακα μεταπτώσεων που θα μας δίνει την πιθανότητα να μεταβούμε από το γράμμα i στο γράμμα j, για το έργο του συγγραφέα w. Αυτον τον πίνακα μεταπτώσεων λοιπόν θα τον συμβολίζουμε με T^w και κάθε στοιχείο του προκύπτει ως εξής:

$$T^w_{i,j} = \frac{Q^w_{i,j}}{Q^w_i}$$

όπου το Q^w_i είναι το άθροισμα των στοιχείων κάθε γραμμής i του πίνακα Q^w:

$$Q^w_i = \sum_j Q^w_{i,j}$$

με το j να διατρέχει το σύνολο των γραμμάτων του αλφαβήτου.

Αφού έχουμε τον πίνακα μεταπτώσεων T^w για τα κείμενα του συγγραφέα w, αν μας δώσει κάποιος ένα οποιοδήποτε κείμενο, μπορούμε να υπολογίσουμε την πιθανότητα να έχει προκύψει από μια αλυσίδα Μάρκοφ που ακολουθεί τον πίνακα T^w. Παίρνουμε το κείμενο που μας δίνεται γράμμα-γράμμα και βρίσκουμε την πιθανότητα αυτής της μετάπτωσης στα κείμενα του συγγραφέα από τον T^w. Για παράδειγμα, έστω ότι το κείμενό μας είναι το:

$\hat{t} = c_1 c_2 \dots c_n$

όπου c_i είναι οι χαρακτήρες του κειμένου στη σειρά. Αν απεικονίσουμε τις μεταπτώσεις ως γράφο, όπως κάναμε με το χρηματιστήριο, τότε η πιθανότητα να παραχθεί αυτό το κείμενο από τον συγγραφέα είναι ίση με την πιθανότητα του μονοπατιού μέσω των κόμβων $c_1, c_2, \dots, c_n$ του γράφου. Και αυτή η πιθανότητα είναι ίση με το γινόμενο των βαρών των συνδέσμων $c_1 \to c_2$, $c_2 \to c_3$, μέχρι $c_{n-1} \to c_n$. Αυτή η πιθανότητα με τη σειρά της είναι ίση με το γινόμενο των αντίστοιχων στοιχείων του πίνακα μετάπτωσης. Αν με $|c_i|$ συμβολίσουμε τη θέση του χαρακτήρα c_i στο αλφάβητο, θα έχουμε:

$$p = T^w_{|c_1|,|c_2|} \times T^w_{|c_2|,|c_3|} \times \cdots \times T^w_{|c_{n-1}|,|c_n|}$$

Για παράδειγμα, αν το $\hat{t}$ είναι «berserker», τότε:

$$\begin{aligned} p &= T^w_{|b|,|e|} \times T^w_{|e|,|r|} \times T^w_{|r|,|s|} \times T^w_{|s|,|e|} \times T^w_{|e|,|r|} \times T^w_{|r|,|k|} \times T^w_{|k|,|e|} \times T^w_{|e|,|r|} \\ &= T^w_{2,5} \times T^w_{5,18} \times T^w_{18,19} \times T^w_{19,5} \times T^w_{5,18} \times T^w_{18,11} \times T^w_{11,5} \times T^w_{5,18} \end{aligned}$$

Είναι πιο πρακτικό να δουλεύουμε με αθροίσματα πιθανοτήτων παρά με γινόμενα, το οποίο μπορούμε να επιτύχουμε παίρνοντας το λογάριθμο της πιθανότητας:

$$\begin{aligned} \lg p &= \lg T^w_{2,5} + \lg T^w_{5,18} + \lg T^w_{18,19} + \lg T^w_{19,5} + \lg T^w_{5,18} + \lg T^w_{18,11} + \lg T^w_{11,5} + \lg T^w_{5,18} \\ &= \lg T^w_{2,5} + 3 \times \lg T^w_{5,18} + \lg T^w_{18,19} + \lg T^w_{19,5} + \lg T^w_{18,11} + \lg T^w_{11,5} \end{aligned}$$

Παρατηρούμε ότι κάθε όρος του αθροίσματος εμφανίζεται τόσες φορές όσες εμφανίζεται η αντίστοιχη μετάπτωση στο κείμενο που εξετάζουμε (τρεις φορές το $T^w_{5,18}$).

Αν $Q^{\hat{t}}$ είναι ο πίνακας που περιλαμβάνει το πλήθος των μεταπτώσεων μεταξύ δύο χαρακτήρων στο κείμενο $\hat{t}$, τότε γενικεύοντας έχουμε:

$$\lg p = \sum_{i,j} Q^{\hat{t}}_{i,j} \lg T^w_{i,j} = \sum_{i,j} Q^{\hat{t}}_{i,j} \lg \left(\frac{Q^w_{i,j}}{Q^w_i} \right)$$

Επειδή η πιθανότητα είναι μεταξύ 0 και 1, αλλάζουμε το πρόσημο της παράστασης ώστε να έχουμε θετικό αποτέλεσμα και θα το συμβολίζουμε με $\Lambda(w, \hat{t})$:

$$\Lambda(w, \hat{t}) = -\lg p = -\sum_{i,j} Q^{\hat{t}}_{i,j} \lg T^w_{i,j} = -\sum_{i,j} Q^{\hat{t}}_{i,j} \lg \left(\frac{Q^w_{i,j}}{Q^w_i} \right)$$

Το $\Lambda(w, \hat{t})$, σύμφωνα με τα όσα είπαμε, είναι ένα μέτρο που μας δείχνει πόσο πιθανό είναι το κείμενο $\hat{t}$ να έχει γραφτεί από τον συγγραφέα w. Έτσι, μας δίνει μια μέθοδο για την απόδοση της πατρότητας ενός κειμένου:

- Ξεκινώντας από ένα σύνολο κειμένων των οποίων γνωρίζουμε τους συγγραφείς, υπολογίζουμε τους πίνακες μετάπτωσής τους.
- Για κάθε κείμενο $\hat{t}$ που θέλουμε να βρούμε το συγγραφέα του w, υπολογίζουμε το $\Lambda(w, \hat{t})$.
- Αποδίδουμε το κείμενο στον συγγραφέα με το μικρότερο $\Lambda(w, \hat{t})$.

Αυτή τη μέθοδο θα πρέπει να υλοποιήσετε στην εργασία σας.

Απαιτήσεις Προγράμματος

Κάθε φοιτητής θα εργαστεί σε αποθετήριο στο GitHub. Για να αξιολογηθεί μια εργασία θα πρέπει να πληροί τις παρακάτω προϋποθέσεις:

- Για την υποβολή της εργασίας θα χρησιμοποιηθεί το ιδιωτικό αποθετήριο του φοιτητή που δημιουργήθηκε για τις ανάγκες του μαθήματος και του έχει αποδοθεί. Το αποθετήριο αυτό έχει όνομα του τύπου `username-algo-assignments`, όπου `username` είναι το όνομα του φοιτητή στο GitHub. Για παράδειγμα, το σχετικό αποθετήριο του διδάσκοντα θα ονομαζόταν `louridas-algo-assignments` και θα ήταν προσβάσιμο στο https://github.com/dmst-algorithms-course/louridas-algo-assignments. *Τυχόν άλλα αποθετήρια απλώς θα αγνοηθούν.*
- Μέσα στο αποθετήριο αυτό θα πρέπει να δημιουργηθεί ένας κατάλογος `assignment-2021-4`.
- Μέσα στον παραπάνω κατάλογο το πρόγραμμα θα πρέπει να αποθηκευτεί με το όνομα `writer_id.py`.

- Μέσα στον ίδιο κατάλογο θα πρέπει να αποθηκευτεί ένα αρχείο με όνομα `transition_matrices.json`, τα περιεχόμενα του οποίου εξηγούνται παρακάτω.
- Δεν επιτρέπεται η χρήση έτοιμων βιβλιοθηκών γράφων ή τυχόν έτοιμων υλοποιήσεων των αλγορίθμων, ή τμημάτων αυτών, εκτός αν αναφέρεται ρητά ότι επιτρέπεται.
- Επιτρέπεται η χρήση δομών δεδομένων της Python όπως στοίβες, λεξικά, σύνολα, κ.λπ.
- Επιτρέπεται η χρήση των παρακάτω βιβλιοθηκών ή τμημάτων τους όπως ορίζεται:
  - `argparse`
  - `defaultdict` από τη βιβλιοθήκη `collections`
  - `glob`
  - `json`
  - `log2` από τη βιβλιοθήκη `math`
  - `os`
  - `re`
- Το πρόγραμμα θα πρέπει να είναι γραμμένο σε Python 3.

Το πρόγραμμα θα καλείται ως εξής (όπου `python` η κατάλληλη εντολή στο εκάστοτε σύστημα):

```
python writer_id.py [-p | -i] input transition_matrices
```

Η σημασία των παραμέτρων είναι η εξής:

- `-p`, `--preprocess`: το πρόγραμμα θα διαβάζει όλα τα αρχεία που θα βρει στον κατάλογο `input` και θα δημιουργεί ένα αρχείο JSON το οποίο θα περιέχει τις μετρήσεις των μεταπτώσεων για όλους τους συγγραφείς των οποίων έργα θα βρει στον κατάλογο `input`. Το αρχείο JSON θα αποθηκεύεται στο αρχείο που θα δίνεται από την παράμετρο `transition_matrices`.
- `-i`, `--id`: το πρόγραμμα θα διαβάζει το αρχείο που θα δίνεται από την παράμετρο `input` και το αρχείο που δίνεται από την παράμετρο `transition_matrices` και θα εμφανίζει το ονόματα του πρώτων δέκα συγγραφέων που θα προτείνει, σε αύξουσα σειρά $\Lambda(w, \hat{t})$. Το αρχείο που δίνεται από την παράμετρο `transition_matrices` θα έχει δημιουργηθεί προηγούμενως καλώντας το πρόγραμμα με την παράμερο `-p`.

Για να χρησιμοποιήσετε το πρόγραμμά σας θα χρειαστείτε ένα σύνολο κειμένων συγγραφέων. Τα κείμενα αυτά μπορείτε να τα βρείτε στα αρχεία train.7z.001, train.7z.002, train.7z.003, train.7z.004. Τα αρχεία αυτά, αφού τα μεταφορτώσετε στον υπολογιστή σας, μπορείτε να τα αποσυμπίεσετε με το εργαλείο `7z`:

```
7z x train.7z.001
```

Η παραπάνω εντολή θα αποσυμπιέσει και τα τέσσερα αρχεία (στην πραγματικότητα είναι ένα αρχείο σπασμένο στα τέσσερα για λόγους ευκολίας αποθήκευσης στο GitHub) και θα τοποθετήσει τα περιεχόμενά του σε έναν κατάλογο `train`. Θα δείτε ότι μέσα στον κατάλογο `train` υπάρχουν αρχεία κειμένου. Το όνομα κάθε αρχείου αποτελείται από τον συγγραφέα και το βιβλίο που έχει γράψει.

Έχοντας ένα σύνολο αρχείων για εκμάθηση, θα καλέσετε το πρόγραμμά σας με την παράμετρο `-p`, οπότε θα δημιουργήσετε ένα αρχείο με όνομα `transition_matrices.json` με πίνακες με τον αριθμό των μεταπτώσεων γραμμάτων για όλους τους συγγραφείς.

Προσοχή: κανονικά ως πίνακα μεταπτώσεων (transition matrix) ονομάζουμε, όπως είδαμε παραπάνω, έναν πίνακα όπου κάθε γραμμή δίνει τις πιθανότητες των μεταπτώσεων. Το αρχείο `transition_matrices.json` θα περιέχει όμως τους ακέραιους αριθμούς των μετρήσεων των μεταπτώσεων, δηλαδή θα περιέχει τους πίνακες Q_i^w, έναν για κάθε συγγραφέα w.

Προκειμένου να καταγράψετε τις μεταπτώσεις των χαρακτήσεων σε κάθε κείμενο, θα πρέπει να το σπάσετε σε λέξεις, να αφαιρέσετε σημεία στίξης και αριθμητικά ψηφία, και να κρατήσετε μόνο τις λέξεις που αποτελούνται από πεζά γράμματα του αλφαβήτου.

Αφού επεξεργαστείτε τα δεδομένα εκμάθησης, θα μπορείτε πλέον να καλείτε το πρόγραμμά σας με την παράμερο `-i` δίνοντάς του το αρχείο με τις μεταπτώσεις και ένα κείμενο του οποίου την πατρότητα θέλετε να βρείτε. Τέτοια κείμενα μπορείτε να βρείτε στα αρχεία test.7z.001, test.7z.002, test.7z.003. Τα αρχεία αυτά, αντίστοιχα με τα αρχεία εκμάθησης, αφού τα μεταφορτώσετε στον υπολογιστή σας, μπορείτε να τα αποσυμπίεσετε με το εργαλείο `7z`:

```
7z x test.7z.001
```

Προφανώς, τα αρχεία ελέγχου περιέχουν στο όνομά τους το συγγραφέα—αυτό γίνεται για δική σας διευκόλυνση, για να ξέρετε ποια θα πρέπει να είναι η σωστή απάντηση. Η μέθοδος που περιγράψαμε φυσικά δεν λαμβάνει υπόψη το όνομα του αρχείου. Ούτε λαμβάνει υπόψη ότι το κείμενο μπορεί να περιέχει το όνομα του συγγραφέα. Τα κείμενα εκμάθησης και τα κείμενα ελέγχου προέρχονται από το Project Gutenberg, συγκεντρωμένα από τον Matthew D. Sholefield, βλ. Gutenberg Dataset.

Παραδείγματα

Παράδειγμα 1

Αν ο χρήστης του προγράμματος δώσει:

```
python writer_id.py -p train transition_matrices.json
```

το πρόγραμμά σας θα διαβάζει τα κείμενα που θα βρει στον κατάλογο και θα δημιουργήσει το αρχείο `transition_matrices.json` με τις μετρήσεις μεταπτώσεων των συγγραφέων. Το αρχείο σας θα πρέπει να είναι όπως αυτό και

όπως αναφέρθηκε παραπάνω θα πρέπει να το συμπεριλάβετε στο αποθετήριό σας μαζί με το πρόγραμμά σας. Το αρχείο σας θα πρέπει να είναι ταξινομημένο ανά συγγραφείς και γράμματα, όπως αυτό εδώ.

Παράδειγμα 2

Αν ο χρήστης του προγράμματος δώσει:

```
python writer_id.py -i test/Charles\ Dickens___Great\ \
Expectations.txt transition_matrices.json
```

το πρόγραμμά σας θα πρέπει να εμφανίζει:

```
Charles Dickens
Frank Richard Stockton
Charlotte Mary Yonge
Jerome Klapka Jerome
Thomas Hardy
William Somerset Maugham
Bram Stoker
P G Wodehouse
Sir Arthur Conan Doyle
Winston Churchill
```

Παράδειγμα 3

Αν ο χρήστης του προγράμματος δώσει:

```
python writer_id.py -i test/Charles\ Darwin___On\ the\ Origin\ of\ \
Species\ by\ Means\ of\ Natural\ Selection\ or\ the\ Preservation\ \
of\ Favoured\ Races\ in\ the\ Struggle\ for\ Life.\ \
\(2nd\ edition\).txt transition_matrices.json
```

το πρόγραμμά σας θα πρέπει να εμφανίζει:

```
Charles Darwin
Alfred Russel Wallace
Herbert Spencer
Thomas Henry Huxley
Bertrand Russell
John Stuart Mill
Samuel Taylor Coleridge
Percival Lowell
John Morley
Edmund Burke
```

Το πρόγραμμα βρίσκει όχι μόνο τη σωστή απάντηση, αλλά στις επόμενες επιλογές συμπεριλαμβάνονται συγγραφείς που σχετίζονται με τον Δαρβίνο.

Παράδειγμα 4

Αν ο χρήστης του προγράμματος δώσει:

```
python writer_id.py -i test/T\ S\ Eliot___Poems.txt \
transition_matrices.json
```

το πρόγραμμά σας θα πρέπει να εμφανίζει:

```
William Blake
T S Eliot
William Penn
Ezra Pound
James Joyce
Elizabeth Barrett Browning
Aldous Huxley
O Henry
Robert Browning
P B Shelley
```

Αυτή τη φορά η σωστή απάντηση είναι η δεύτερη επιλογή, αλλά το πρόγραμμα προτιμάει ως πιθανούς συγγραφείς ποιητές.

Αν δοκιμάσετε το πρόγραμμά σας για όλα τα κείμενα ελέγχου, θα πρέπει το ποσοστό επιτυχίας σας να είναι γύρω στο 63%. Αυτό δεν είναι άσχημο, αν σκεφτείτε ότι η μέθοδος στην ουσία της είναι πολύ απλή, και σίγουρα δεν καταλαβαίνει τίποτε από τα κείμενα που διαβάζει.

Λίγα Επιπλέον Στοιχεία

Ας θυμηθούμε λίγο τον ορισμό της εντροπίας. Έστω ότι έχουμε ένα σύνολο γεγονότων, $X = \{x_1, x_2, \ldots, x_n\}$ που αποτελείται από n διαφορετικά ενδεχόμενα, κάθε ένα από τα οποία εμφανίζεται με πιθανότητα $p(x_i)$, τότε η εντροπία του συστήματος ορίζεται ως:

$$H(X) = -p(x_1)\lg p(x_1) - p(x_2)\lg p(x_2) - \cdots - p(x_n)\lg p(x_n)$$

ή

$$H(X) = -\sum_{i=1}^{n} p(x_i)\lg p(x_i)$$

Αν πάρουμε το $\Lambda(w, \hat{t}\,)$ και το διαιρέσουμε με το μήκος του κειμένου $\hat{t}$ μείον 1, το οποίο θα συμβολίσουμε με $\|\hat{t}\| - 1$, έχουμε:

$$\frac{\Lambda(w, \hat{t})}{\|\hat{t}\| - 1} = -\sum_{i,j} \frac{Q^{\hat{t}}_{i,j}}{\|\hat{t}\| - 1} \lg T^{w}_{i,j} = -\sum_{i,j} \frac{Q^{\hat{t}}_{i,j}}{\|\hat{t}\| - 1} \lg \left(\frac{Q^{w}_{i,j}}{Q^{w}_{i}} \right)$$

Τώρα όμως, η τιμή $\|\hat{t}\| - 1$ δεν είναι άλλη από το σύνολο των μεταπτώσεων στο κείμενο $\hat{t}$. Οπότε το κλάσμα $\frac{Q^{\hat{t}}_{i,j}}{\|\hat{t}\|-1}$ είναι η πιθανότητα στο κείμενο $\hat{t}$ να εμφανιστεί η μετάπτωση από τον i χαρακτήρα του αλφαβήτου στο j χαρακτήρα του αλφαβήτου. Το κλάσμα $\frac{Q^{w}_{i,j}}{Q^{w}_{i}}$ είναι η πιθανότητα στο σύνολο των κειμένων του συγγραφέα w, αν βρισκόμαστε στον χαρακτήρα i του αλφαβήτου ο επόμενος χαρακτήρας να είναι ο χαρακτήρας j του αλφαβήτου. Αν συμβολίσουμε μην πρώτη πιθανότητα με $p(i,j)$ και τη δεύτερη πιθανότητα με $q(i,j)$ έχουμε:

$$\frac{\Lambda(w,\hat{t}\,)}{\|\hat{t}\| - 1} = -\sum_{x\in(i,j)} p(x)\lg q(x)$$

Βλέπουμε λοιπόν ότι η τιμή που προσπαθούμε να ελαχιστοποιήσουμε μοιάζει πολύ με τον ορισμό της εντροπίας. Πράγματι, αν έχουμε δύο κατανομές πιθανοτήτων, $p(x)$ και $q(x)$ πάνω στο ίδιο σύνολο ενδεχομένων, τότε η *διεντροπία* (cross-entropy) της κατανομής $q(x)$ ως προς την κατανομή $p(x)$ ορίζεται ως:

$$H(p,q) = -\sum_{x} p(x)\lg(x)$$

Η διεντροπία είναι μια μετρική που χρησιμοποιείται συχνά στον κλάδο της Μηχανικής Μάθησης προκειμένου να μετρήσουμε τη διαφορά μεταξύ της εξόδου ενός συστήματος, που μπορεί να είναι λάθος, και της σωστής, επιθυμητής εξόδου του συστήματος. Στην περίπτωσή μας, η πατρότητα του κειμένου προκύπτει από την ελαχιστοποίηση της διεντροπίας μεταξύ δύο κατανομών πιθανοτήτων για τις μεταπτώσεις από κάθε χαρακτήρα i σε κάθε χαρακτήρα j. Η πρώτη κατανομή πιθανοτήτων $p(x)$ προκύπτει από το κείμενο που εξετάζουμε και η δεύτερη κατανομή πιθανοτήτων $q(x)$ προκύπτει από το σύνολο των κειμένων του συγγραφέα που εξετάζουμε.

Για Περισσότερες Πληροφορίες

Η μέθοδος για την απόδοση της πατρότητας που πραγματεύεται αυτή η άσκηση παρουσιάστηκε από τον Khmlelev (2000) και Khmelev και Tweedie (2001).

Για μια εισαγωγή στις αλυσίδες Μάρκοφ, ανατρέξτε στο άρθρο του Hayes (2013)· στην ιστορική αναδρομή θα δείτε τη σχέση των αλυσίδων Μάρκοφ με τη θεολογία και τον ντετερμινισμό.

Το άρθρο του Μάρκοφ για τον Ευγένιο Όνεγκιν δημοσιεύθηκε στα ρωσικά το 1913, υπάρχει όπως μεταφρασμένο στα αγγλικά (Markov 1913, 2006).

Hayes, Brian. 2013. “First Links in the Markov Chain.” *American Scientist* 101 (2): 92–97.

Khmelev, D. V. 2000. "Disputed Authorship Resolution Through Using Relative Empirical Entropy for Markov Chains of Letters in Human Language Texts." *Journal of Quantitative Linguistics* 7 (3): 201–7.

Khmelev, Dmitri. V., and Fiona J. Tweedie. 2001. "Using Markov Chains for Identification of Writers." *Literary and Linguistic Computing* 16 (3): 299–307.

Markov, A. A. 1913. "An Example of Statistical Investigation of the Text *Eugene Onegin* Concerning the Connection of Samples in Chains." *Bulletin of the Imperial Academy of Sciences of St. Petersbhurg* 7 (3): 53–162.

———. 2006. "An Example of Statistical Investigation of the Text *Eugene Onegin* Concerning the Connection of Samples in Chains." *Science in Context* 19 (4): 591–600.

Καλή Επιτυχία!