

Εισαγωγή

Η δυνατότητα ηλεκτρονικής επεξεργασίας και έκδοσης κειμένων αποτέλεσε βασική εφαρμογή των εργαλείων της πληροφορικής στις Ανθρωπιστικές σπουδές και αποτέλεσε σημαντικό παράγοντα για την ανάδειξη, τη διαχείριση και την επιστημονική επεξεργασία των έργων του γραπτού λόγου.

Ένα από τα πλέον αποτελεσματικά εργαλεία για την επεξεργασία και διαχείριση των συλλογών ανθρωπιστικού περιεχομένου είναι το ηλεκτρονικό κείμενο, όρος που χρησιμοποιείται για να δηλώσει κυρίως την ψηφιακή κωδικοποίηση ενός κειμένου παρά την ψηφιακή απεικόνισή του.

Για τη δομημένη εμφάνιση και επεξεργασία των ηλεκτρονικών κειμένων αξιοποιούνται οι γλώσσες σήμανσης (markup languages) για την δημιουργία προτύπων κωδικοποίησης δεδομένων (και μεταδεδομένων) των κειμένων των Ανθρωπιστικών σπουδών και εργαλείων διαχείρισής τους στην ηλεκτρονική τους μορφή (ηλεκτρονικά κείμενα).

Κατηγορίες ψηφιακών εργαλείων επεξεργασίας χειρόγραφων κειμενικών τεκμηρίων:

- Transcription – Μεταγραφή / Παραγωγή ηλεκτρονικού κειμένου.
- Analysis – Κειμενική ανάλυση ηλεκτρονικού κειμένου
- Encoding & Markup - Κωδικοποίηση / σήμανση ηλεκτρονικού κειμένου.
- Visualisation - Οπτικοποίηση δεδομένων
- Edition – Έκδοση

Ψηφιακή μεταγραφή (Transcription)

Η διαδικασία της ψηφιοποίησης έχει ως αποτέλεσμα την μετατροπή του έντυπου κειμένου σε μορφή ψηφιακής εικόνας, το οποίο ονομάζεται ψηφιοποιημένο κείμενο (Digitized text).

Το ψηφιοποιημένο κείμενο είναι πολύτιμο, καθώς προσφέρει ταχύτητα, πληρότητα, ακρίβεια στη διάθεση του υλικού και πρόσβαση σε σπάνια βιβλία και χειρόγραφα μειώνοντας δραστικά το κόστος της έρευνας και διευρύνοντας σημαντικά το πεδίο της. Όμως δεν είναι αρκετό.

Εκτός από τις ψηφιακές φωτογραφίες κειμενικών τεκμηρίων, είναι απαραίτητη η μετατροπή του ψηφιοποιημένου κειμένου (Digitised text) σε ηλεκτρονική μορφή (ψηφιακό / ηλεκτρονικό κείμενο, digital text)), η οποία θα ξεπερνά τη δισδιάστατη εγγραφή πληροφοριών και θα επιτρέπει την ανάλυσή του. Η διαδικασία μετατροπής του ψηφιοποιημένου έντυπου κειμένου σε ηλεκτρονικό ονομάζεται Ψηφιακή μεταγραφή.

Ένα από τα πλέον αποτελεσματικά εργαλεία για την επεξεργασία και διαχείριση των συλλογών ανθρωπιστικού περιεχομένου είναι το ηλεκτρονικό κείμενο, όρος που χρησιμοποιείται για να δηλώσει κυρίως την ψηφιακή κωδικοποίηση ενός κειμένου παρά την ψηφιακή απεικόνισή του

Το ηλεκτρονικό κείμενο επιτρέπει την:

- Σήμανση και κωδικοποίηση των διαφόρων συστατικών του γραπτού λόγου.
- Οργάνωση, σύνθετη αναζήτηση και αναπαράσταση της κειμενικής πληροφορίας που εμπεριέχεται σε αυτό.
- Επεξεργασία και ανάλυση με ψηφιακά εργαλεία / λογισμικό για την εξαγωγή αποτελεσμάτων από το κειμενικό υλικό, που διαφορετικά θα απαιτούσαν πολύ ανθρώπινο χρόνο και μόχθο.
- Διασύνδεση και εμπλουτισμό του με δεδομένα εξωτερικών πηγών (Υπερκειμενικότητα – Πολυμεσικότητα – Διαδραστικότητα)

Δυστυχώς οι παραπάνω δυνατότητες δεν αξιοποιούνται σε μεγάλο βαθμό στην περίπτωση των ελληνικών προσπαθειών. Η απουσία συνεργασίας κι επικοινωνίας μεταξύ ερευνητών και η απουσία συντονισμού και οργάνωσης των επιμέρους

προσπαθειών οδηγούν σε ένα φαύλο κύκλο σπατάλησης κονδυλίων και παραγωγής μη χρηστικού ψηφιακού υλικού μακριά από το στόχο της αποτελεσματικής χρήσης ψηφιακών μεθόδων κι εργαλείων για την επεξεργασία, παρουσίαση και έκδοση κειμένων. Το μεγαλύτερο ποσοστό του ψηφιοποιημένου κειμενικού υλικού παραμένει διαθέσιμο είτε με τη μορφή σκαναρισμένων σελίδων είτε με τη μορφή ηλεκτρονικών απλών (μη σημασιολογικά κωδικοποιημένων) κειμένων (σε doc, text files ή σε HTML μορφή). Δηλαδή, το κειμενικό υλικό απλά παρέχεται μέσω ψηφιακών μέσων και δεν γίνεται αντικείμενο περαιτέρω επεξεργασίας, δεν εμπλουτίζεται και ως εκ τούτου αυτό που είναι ουσιαστικά το καινούριο είναι απλά η μετατροπή μετατροπής ενός έντυπου κειμένου από αναλογική σε ψηφιακή μορφή.

Η μετατροπή του έντυπου κειμένου σε ψηφιακό κείμενο είναι εφικτό να προκύψει με δύο τρόπους:

- Δακτυλογράφηση - Η χειρωνακτική εισαγωγή δεδομένων με πληκτρολόγηση έχει μεγάλο κόστος και είναι χρονοβόρα. Είναι απαραίτητη αν η ποιότητα του πρωτότυπου χειρόγραφου κειμενικού τεκμηρίου είναι χαμηλή. Προσφέρει μεγάλη ακρίβεια σε περιπτώσεις που αφορούν σε για τύπους δεδομένων στοιχείων (κατάλογοι, αριθμητικά σύνολα δεδομένων).
- Οπτική σάρωση και αναγνώριση χαρακτήρων χειρόγραφου (Handwritten Text Recognition - HTR)

Μεταγραφή με δακτυλογράφηση

Οι παπυρολόγοι αναλύουν, μεταγράφουν και επεξεργάζονται θραύσματα παπύρου προκειμένου να εμπλουτίσουν τη σύγχρονη ζωή με καλύτερη κατανόηση της γλωσσολογίας, του πολιτισμού και της λογοτεχνίας του αρχαίου κόσμου. Ένας συνήθης στόχος των Παπυρολόγων είναι να ταιριάξουν ένα άγνωστο κομμάτι σε ένα γνωστό χειρόγραφο. Αυτό είναι ιδιαίτερα δύσκολο όταν τα θραύσματα είναι κατεστραμμένα και περιέχουν μόνο περιορισμένες πληροφορίες λόγω της φθοράς.

Οι πάπυροι της Οξυρρύγχου Εκατοντάδες χιλιάδες θραύσματα παπύρων ανακτήθηκαν πριν από 100 χρόνια στην αιγυπτιακή πόλη της Οξυρρύγχου. Γραμμένα στην ελληνική γλώσσα και κατά την διάρκεια μιας περιόδου όπου η Αίγυπτος βρισκόταν υπό τον έλεγχο Ελλήνων και αργότερα των Ρωμαίων, τα κείμενα περιλαμβάνουν μια ποικιλία εγγράφων, από λογοτεχνικά έργα, επιστολές κ.α. Μέχρι σήμερα έχει μεταφραστεί μόνο το 2%.

Το Ancient Lives Project έχει σκοπό να δώσει την δυνατότητα σε εθελοντές να βοηθήσουν σε αυτήν την προσπάθεια μεταγραφής των Οξυρρύγγειων παπύρων με δακτυλογράφηση.

Το πρόγραμμα Ancient Lives είναι μια Crowdsourcing εφαρμογή του πανεπιστημίου της Οξφόρδης σε εξέλιξη που παρακινεί τους διαδικτυακούς επισκέπτες να συμμετέχουν ενεργά στην επιτάχυνση της προσπάθειας μεταγραφής όσο γίνεται περισσότερων κειμένων μέσα από τους πρωτότυπους πάπυρους της Οξυρρύγχου με την καθοδήγηση ακαδημαϊκών. Θα ακολουθήσει η διαδικασία ταυτοποίησης των παπύρων και ανακατασκευής του κειμένου.

Οι ερευνητές του προγράμματος έχουν ανεβάσει στη διαδικτυακή διεύθυνση www.ancientlives.org περισσότερα από 200.000 τμήματα καθώς και τα σχετικά εργαλεία που αναγνωρίζουν χαρακτήρες. Η διαδικτυακή πλατφόρμα Ancient Lives προσφέρει μια ειδική μηχανή αναπαραγωγής του περιεχομένου των παπύρων με όλα τα εργαλεία για να μαρκάρει και να αναγνωρίσει ο χρήστης τα γράμματα που αναγράφονται σε κομμάτι ενός παπύρου. Οι εθελοντές καλούνται να κοιτάξουν φωτογραφίες των πάπυρων, να κάνουν κλικ σε ένα χαρακτήρα της εικόνας και στη συνέχεια να και να αντιστοιχήσουν τους χαρακτήρες που βλέπουν στο χειρόγραφο με τα πλήκτρα ενός εικονικού πληκτρολόγιου στα ελληνικά. Επιπλέον, εργαλεία αναγνώρισης χαρακτήρων βοηθούν τους εθελοντές να ταιριάξουν τα σύμβολα με γράμματα. Κατόπιν, η μεταγραφή του εθελοντή περνάει αυτόματα σε ακαδημαϊκούς μελετητές για επαλήθευση. Μέχρι σήμερα, πάνω από 7 εκατομμύρια γράμματα έχουν αναγνωρισθεί από χρήστες όλου του κόσμου που έχουν καταγραφεί στη βάση δεδομένων της 'Αρχαίες Ζωές'.

Οπτική σάρωση και αναγνώριση χειρόγραφων χαρακτήρων

Οι τεχνικές OCR οι οποίες έχουν εφαρμοστεί με μεγάλη επιτυχία σε επεξεργασία σύγχρονων κειμένων.

Η εφαρμογή όμως των τεχνικών αυτών σε χειρόγραφα κείμενα, είναι εξαιρετικά δύσκολη και παραμένει μια ανακριβής διαδικασία και επομένως αναξιόπιστη επιλογή.

Οι παρακάτω ιδιαιτερότητες των χειρογράφων δυσχεραίνουν την διαδικασία εφαρμογής προτύπων ή εξαγωγής χαρακτηριστικών και δυσκολεύουν έτσι την αυτόματη επεξεργασία και αναγνώρισή τους:

- Φυσική φθορά λόγω παλαιότητας και κακή ποιότητα εγγράφων έλλειψης φροντίδας και συντήρησής τους.
- Κακή ποιότητα εκτύπωσης αν πρόκειται για τυπωμένα έγγραφα
- Άγνωστες και ασυνήθιστες γραμματοσειρές, πολύπλοκες δομές.
- Καλλιγραφία και αλληλοεπικαλύψεις /συνενώσεις (συνεχόμενη γραφή χωρίς κενό) μεταξύ γειτονικών χαρακτήρων, που καθιστούν πολύ δύσκολη την απομόνωσή τους και τον διαχωρισμό τους.
- Γραφή σε μη παράλληλες γραμμές ή λέξεις με κλίση.
- Ποικιλομορφία γραφικών χαρακτήρων και τρόπων γραφής ακόμα και σε κείμενα του ίδιου συγγραφέα.

Για τους παραπάνω λόγους, για ιστορικά έγγραφα τα οποία δεν είναι σε άριστη κατάσταση και σε όσες περιπτώσεις είναι εφικτό, συνήθως προηγείται της ψηφιοποίησης η συντήρηση των εγγράφων καθώς, όπως είναι λογικό, η αναγνώριση χαρακτήρων σε ένα έγγραφο τσακισμένο, ξεθωριασμένο και με πτυχώσεις θα έχει περισσότερα λάθη.

Σε περιπτώσεις κακής ποιότητας ψηφιοποίησης όπως χαμηλή ανάλυση, σκιές, χαμηλός βαθμός καθαρότητας, αντίθεσης και ομοιομορφίας στην φωτεινότητα του υπόβαθρου, έντονος θόρυβος, εμφάνιση περιθωρίου, παραμόρφωση, δυσδιάκριτες περιοχές κειμένου κ.α. κρίνεται σκόπιμη η χρήση λογισμικού επεξεργασίας εικόνας, ώστε να βελτιωθεί η αντίθεση και να μειωθούν οι αλλοιώσεις στα χρώματα της εικόνας.

Αναζήτηση και εντοπισμός λέξεων (*KeyWord Spotting*)

Η αναζήτηση και ο εντοπισμός λέξεων (Wordspotting) απευθείας επάνω στις εικόνες των ιστορικών εγγράφων, τόσο τυπωμένων όσο και χειρόγραφων, παρακάμπτοντας τη διαδικασία OCR, είναι μια εναλλακτική λύση η οποία αντιμετωπίζει τα προβλήματα που συναντούμε στα ιστορικά έγγραφα, όσον αφορά την πρόσβαση στο περιεχόμενο των κειμένων τους.

Οι εικόνες των λέξεων – κλειδιών που ενδιαφερόμαστε να εντοπίσουμε στα ιστορικά έγγραφα εντοπίζονται απευθείας επάνω στις εικόνες των εγγράφων, χωρίς να έχει εφαρμοστεί καμία διαδικασία OCR.

Οι μέθοδοι αναζήτησης λέξεων σε ιστορικά κείμενα χρησιμοποιούν εικόνες λέξεων-κλειδιών (query keyword images) ως πρότυπα, τις οποίες προσπαθούν να εντοπίσουν απευθείας επάνω στις εικόνες των ιστορικών εγγράφων.

Οι χαρακτηριστικότερες μέθοδοι εντοπισμού λέξεων σε ιστορικά κείμενα βασίζονται είτε εφαρμόζοντας κάποιου είδους κατάτμηση στα κείμενα είτε χρησιμοποιώντας τις εικόνες των κειμένων αυτούσιες χωρίς καμία κατάτμηση. Η μέθοδος της κατάτμησης προϋποθέτει την κατάτμηση των κειμένων σε γραμμές, λέξεις ή χαρακτήρες. Σταδιακά έχει αρχίσει να υπάρχει μεγάλο ενδιαφέρον για μεθόδους wordspotting οι οποίες είναι απαλλαγμένες από τη διαδικασία κατάτμησης σε οποιοδήποτε επίπεδο. Η επεξεργασία χειρόγραφων εγγράφων και συγκεκριμένα η αναζήτηση λέξεων σε τέτοιου είδους έγγραφα αποτελεί μεγάλη πρόκληση.

Αναγνώριση χειρόγραφων χαρακτήρων με μηχανική μάθηση (HTR)

Transcribus - Βασίζεται στην τεχνητή νοημοσύνη (νευρωνικά δίκτυα) και έχει φέρει την επανάσταση στην ψηφιοποίηση, αυτόματη μεταγραφή και αναγνώριση ιστορικών εγγράφων, τυπωμένων και χειρόγραφων. Οι χρήστες πρώτα καταχωρούν την μεταγραφή ενός μέρους του ψηφιοποιημένου κειμένου, κατόπιν το λογισμικό μαθαίνει να αναγνωρίζει τους χαρακτήρες και μετά ολοκληρώνει την μεταγραφή αυτόματα με εντυπωσιακή ακρίβεια, που ξεπερνά το 95% στα ιστορικά έγγραφα, ανεξάρτητα από τη γλώσσα ή το στυλ γραφής!

Στα βασικά στάδια της διαδικασίας είναι τα παρακάτω:

- **Δημιουργία λογαριασμού χρήστη και εγκατάσταση της εφαρμογής Transkribus.**
- **Αποθήκευση (upload) των αρχείων ψηφιοποιημένου κειμένου στον server.**
- **Αυτόματη κατάτμηση (Segmentation) του ψηφιοποιημένου κειμένου** - Σε αυτό το στάδιο γίνεται αυτόματη ανάλυση της δομής του κειμένου (π.χ. παράγραφος, τίτλος), προσδιορισμός των περιοχών της εικόνας που περιέχουν κείμενο (text regions), κατάτμηση του κειμένου σε γραμμές (text lines) και κατάτμηση των γραμμών του κειμένου (text lines) σε λέξεις (words). Με αυτόν τον τρόπο γίνεται εφικτή η άμεση σχέση ανάμεσα στην μεταγραφή και στο ψηφιοποιημένο κείμενο (εικόνα), είτε πρόκειται για έντυπο είτε για χειρόγραφο. Μόλις η αυτόματη κατάτμηση (segmentation) ολοκληρωθεί, τα αποτελέσματα θα εμφανιστούν στο πλαίσιο προβολής της σελίδας / εικόνας του ψηφιοποιημένου κειμένου. Διόρθωση τυχόν λαθών στον αυτόματο εντοπισμό των περιοχών (TR) και των γραμμών (BL) κειμένου μπορεί να γίνει χειροκίνητα.
- **Παραγωγή δεδομένων εκπαίδευσης ενός HTR μοντέλου** - Σε αυτό το στάδιο ο χρήστης καταχωρεί την μεταγραφή (Transcription) ενός αντιπροσωπευτικού μέρους του ψηφιοποιημένου κειμένου. Με αυτόν τον τρόπο παράγονται τα δεδομένα εκπαίδευσης για το μοντέλο HTR (Handwritten Text Recognition) αυτόματης αναγνώρισης χειρόγραφων χαρακτήρων. Η παραπάνω μεταγραφή πρέπει να είναι ακριβής και ανά γραμμή για κάθε εικόνα σελίδας, ενώ δεν είναι απαραίτητο να είναι πλήρης. Εάν υπάρχει μια δυσανάγνωστη λέξη είναι καλύτερα να απλώς διαγράψετε ολόκληρη τη γραμμή στην οποία εμφανίζεται αυτή η λέξη, ώστε η συγκεκριμένη γραμμή να μην χρησιμοποιηθεί στη συνέχεια για την εκπαίδευση του μοντέλου HTR. Σε περίπτωση που την μεταγραφή έχει αναλάβει ομάδα, υπάρχει η δυνατότητα η μεταγραφή να γίνει χρησιμοποιώντας την ελαφριά έκδοση του Transkribus που είναι εύκολη και απλή στη χρήση: <https://transkribus.eu/read>
- **Δημιουργία, εκπαίδευση και αξιολόγηση ενός HTR μοντέλου** - Για την εκπαίδευση ενός μοντέλου HTR, προτείνεται η αξιοποίηση μεταγραφής περίπου 5.000 λέξεων για ψηφιοποιημένο έντυπο κείμενο ή 15.000 λέξεων για ψηφιοποιημένο χειρόγραφο κείμενο. Όσο περισσότερα δεδομένα εκπαίδευσης, τόσο καλύτερα είναι τα αποτελέσματα. Επιπλέον, η

ποιότητα των ψηφιοποιημένων εικόνων είναι επίσης ανάλογη της προκύπτουσας ακρίβειας του μοντέλου HTR. Η λειτουργία εκπαίδευσης μοντέλου δεν περιλαμβάνεται αυτόματα στην τυποποιημένη πλατφόρμα Transkribus. Όταν είστε έτοιμοι να εκπαιδεύσετε ένα μοντέλο, επικοινωνήστε με την ομάδα του Transkribus (email@transkribus.eu) και ζητήστε έγκριση πρόσβασης. Μετά την ολοκλήρωση της διαδικασίας εκπαίδευσης του μοντέλου, ο χρήστης μπορεί να ελέγξει το ποσοστό της ακρίβειας των αποτελεσμάτων της αυτόματης μεταγραφής του μοντέλου, και επιπλέον να ελέγξει ποιες λέξεις το μοντέλο κατάφερε να μεταγράψει σωστά και ποιες λάθος. Τελικά ποσοστά λάθους της τάξεως του 10% θεωρούνται πολύ ικανοποιητικά για χρήση του μοντέλου για Keyword Spotting και αυτόματη μεταγραφή των υπόλοιπων σελίδων των ψηφιοποιημένων χειρόγραφων κειμένων του ίδιου γραφέα που είναι διαθέσιμα στην συλλογή του.

- **Παραγωγή αυτόματης μεταγραφής με χρήση ενός HTRμοντέλου** - Μετά την ολοκλήρωση της διαδικασίας εκπαίδευσης του μοντέλου HTR, ο χρήστης μπορεί να το χρησιμοποιήσει για την αυτόματη μεταγραφή των υπόλοιπων σελίδων των ψηφιοποιημένων χειρόγραφων κειμένων του ίδιου γραφέα που είναι διαθέσιμα στην συλλογή του.
- **Αποθήκευση και εξαγωγή του ηλεκτρονικού κειμένου σε διάφορες μορφές**
- **Αναζήτηση συγκεκριμένων λέξεων με τεχνολογία Keyword Spotting.**

Αυτόματη Αναγνώριση και ταυτοποίηση Γραφέα Χειρογράφων (AWI)

Automatic Writer Identification – AWI (Αυτόματη αναγνώριση του γραφέα)

Ένα σημαντικό ερευνητικό ζήτημα είναι η αναγνώριση και ταύτιση ενός αναγραφέα με μία επιγραφή, καθώς αποτελεί πολύτιμη πληροφορία για τη χρονολόγησή της, μία διαδικασία που παραδοσιακά στηρίζεται στη μελέτη και στην εμπειρία των επιγραφολόγων. Αξιοποιώντας ειδικές τεχνικές μέτρησης διαφόρων παραμέτρων των χαραγμένων γραμμάτων σε συνδυασμό με εξελιγμένα στατιστικά προγράμματα, αποδείχθηκε ότι είναι εφικτό να αποδοθεί σε αρχαίους αναγραφείς το σύνολο των επιλεγμένων προς μελέτη επιγραφών σε απόλυτη συμφωνία με τα συμπεράσματα κορυφαίων επιγραφολόγων.

Η ταυτοποίηση και χρονολόγηση των χειρογράφων είναι ζωτικής σημασίας για την αξιοποίηση του περιεχομένου τους.

Εργαλεία επεξεργασίας εικόνας που αξιοποιούν ειδικές τεχνικές μέτρησης διαφόρων παραμέτρων των γραμμάτων σε συνδυασμό με εξελιγμένα στατιστικά προγράμματα, επιτρέπουν τον αυτόματο προσδιορισμό και την ταυτοποίηση του γραφέα ενός αρχαίου χειρόγραφου και την έμμεση χρονολόγηση των αντίστοιχων κειμένων.

Τα βασικά στάδια της επεξεργασίας περιλαμβάνουν:

- Αυτόματη κατάτμηση της εικόνας για την εξαγωγή των γραμμάτων από το έγγραφο.
- Διαχωρισμό του σώματος κάθε γράμματος από το υπόβαθρό του και εξαγωγή του περιγράμματος κάθε γράμματος
- Ταξινόμηση των διαφορετικών περιγραμμάτων κάθε γράμματος σε ομάδες και καταχώρηση τους σε βάσεις δεδομένων για στατιστική ανάλυση και περαιτέρω επεξεργασία.

Αναλυτικότερα, οι μεθοδολογίες και οι αλγόριθμοι αναγνώρισης προτύπων για την σύγκριση των συμβόλων της αλφαβήτου από τα οποία αποτελούνται τα κείμενα αυτά για την αυτόματη ομαδοποίηση και κατάταξη ενός αριθμού κειμένων επί τη βάση του γραφέα τους, βασίζονται στην κωδικοποίηση και ομαδοποίηση ενός συνόλου ειδικών και διακριτών χαρακτηριστικών τεχνοτροπίας γραφής και γραφολογικής φύσης, όπως η καμπύλωση των γραμμάτων, η κλίση τους, η πίεση που έχει ασκηθεί στο υλικό αποτύπωσης της γραφής (χαρτί, πάπυρος κλπ.), το πόσο λεπτή είναι η γραφή στα διάφορα σημεία του γράμματος κλπ.

Αυτή η μέθοδος χρησιμοποιήθηκε και στο παλίμψηστο του Αρχιμήδη

Αποκατάσταση κατεστραμμένων κειμένων σε επιγραφές -"Pythia"

Ένα από τα κύρια ζητήματα που απασχολούν τους μελετητές των αρχαίων επιγραφών είναι η αποκατάσταση των κειμένων από επιγραφικά μνημεία που έχουν υποστεί καταστροφές. Το Πανεπιστήμιο της Οξφόρδης σε συνεργασία με την ιδιωτική εταιρεία DeepMind Technologies ανέπτυξαν την "Pythia", ένα σύστημα αποκατάστασης των αρχαίων κειμένων με εντυπωσιακά αποτελέσματα. Το ψηφιακό εργαλείο έχει σχεδιαστεί, ώστε να συμπληρώνει γράμματα ή και ολόκληρες λέξεις που η πάροδος του χρόνου έχει αλλοιώσει πάνω στον λίθο, προτείνοντας εναλλακτικές συμπληρώσεις.

Ένα ψηφιακό εργαλείο που το οποίο βασίζεται τεχνητή νοημοσύνη (AI) και την επιγραφική επιστήμη για την ανάγνωση μισοκατεστραμμένων επιγραφών με αρχαία ελληνικά κείμενα. Ένα πλήρως αυτοματοποιημένο σύστημα αποκατάστασης αρχαίου κειμένου που μαντεύει το χαμένο κείμενο από μερικώς κατεστραμμένες αρχαίες επιγραφές. Η Πυθία θα μπορούσε επίσης να εφαρμοστεί σε όλους τους κλάδους που ασχολούνται με τα αρχαία κείμενα (φιλολογία, παπυρολογία, κωδικολογία) σε οποιαδήποτε γλώσσα (αρχαία ή σύγχρονη).

Έχει σχεδιαστεί έτσι ώστε να συμπληρώνει γράμματα ή και ολόκληρες λέξεις που η πάροδος του χρόνου έχει αλλοιώσει πάνω σε επιγράψες, προτείνοντας εναλλακτικές συμπληρώσεις. Συμπληρώνει τις λέξεις που λείπουν, κάνοντας εναλλακτικές προτάσεις, πράγμα που μπορεί να βοηθήσει σημαντικά στο μέλλον τους επιστήμονες να αναδημιουργήσουν αποσπασματικά αρχαία ελληνικά κείμενα τα οποία είναι χαραγμένα σε αρχαίες επιγραφές οι οποίες έχουν διαβρωθεί ή καταστραφεί με το πέρασμα των αιώνων, με αποτέλεσμα να είναι πια δυσανάγνωστες ή ένα μέρος του κειμένου τους έχει χαθεί για πάντα, και τελικά να τις διαβάσουν.

Στα πλαίσια του project η DeepMind:

- Αρχικά μετέτρεψε μια μεγάλη ψηφιακή συλλογή αρχαίων ελληνικών επιγραφών σε κείμενο
- Έπειτα, κατασκεύασε και εκπάιδευσε ένα νευρωνικό δίκτυο (είδος αλγόριθμου τεχνητής νοημοσύνης) να αναγνωρίζει μοτίβα σε επιγραφές και να μαντεύει τις λέξεις ή τα γράμματα της αρχαίας ελληνικής γλώσσας που λείπουν σε επιγραφές.

Όταν βρεθεί μπροστά σε μια νέα ελλιπή επιγραφή, η «Πυθία» προτείνει έως 20 διαφορετικές λέξεις ή γράμματα και αφήνει τους ειδικούς να επιλέξουν αυτήν που θεωρούν καλύτερη λύση. Μαθαίνει να αναγνωρίζει μοτίβα σε επιγραφές χρησιμοποιώντας νευρωνικά δίκτυα (Neural Networks) και όταν παρουσιάζεται μια ελλιπής επιγραφή, η Πυθία διατυπώνει μέχρι 20 υποθέσεις, δηλαδή προτείνει έως και 20 διαφορετικά γράμματα ή λέξεις, αφήνοντας στους εμπειρογνώμονες να κάνουν την τελική επιλογή.

Η Πυθία παίρνει μια ακολουθία κατεστραμμένου κειμένου ως εισαγωγή και εκπαιδεύεται για να προβλέψει αλληλουχίες χαρακτήρων που περιλαμβάνουν υποθετικές αποκαταστάσεις αρχαίων ελληνικών επιγραφών (κείμενα γραμμένα στο ελληνικό αλφάβητο που χρονολογούνται μεταξύ του 7ου αιώνα π.Χ. και του 5ου αιώνα μ.Χ.). Η αρχιτεκτονική λειτουργεί τόσο σε επίπεδο χαρακτήρων όσο και σε επίπεδο λέξεων, με αποτέλεσμα την αποτελεσματική διαχείριση μακροπρόθεσμων πληροφοριών πλαισίου και την αποτελεσματική αντιμετώπιση ελλιπών αντιγράφων λέξεων

Η DeepMind δοκίμασε το Pythia χρησιμοποιώντας ήδη διορθωμένα κείμενα και με τον καιρό και χάρη στο machine learning το Pythia θα βελτιώνεται και ίσως μια μέρα καταφέρει να αποκαταστήσει κάποιο κείμενο με ποσοστό επιτυχίας 100%.

Σε μια συγκριτική δοκιμή, όπου τόσο η «Πυθία» όσο και οι ειδικοί κλήθηκαν να καλύψουν τα κενά σε 2.949 επιγραφές, οι τελευταίοι είχαν ποσοστό σφαλμάτων 57,3%, ενώ η τεχνητή νοημοσύνη σχεδόν το μισό (30,1%). Επιπλέον, ενώ οι άνθρωποι χρειάστηκαν δύο ώρες για μόνο 50 επιγραφές, η «Πυθία» έκανε τις προτάσεις της για το σύνολο των επιγραφών μέσα σε μερικά δευτερόλεπτα!

Αυτό δείχνει τις δυνατότητες της τεχνητής νοημοσύνης στο πεδίο της ψηφιακής επιγραφικής συμβάλλοντας στην αποκατάσταση των επιγραφών.

Η «Πυθία» θα μπορούσε να αξιοποιηθεί και σε συναφή πεδία, όπως η αρχαία φιλολογία, η παπυρολογία, η κωδικολογία κ.ά.