Σύστημα Απάντησης Ερωτήσεων (ΣΑΕ)

ΠΑΡΟΥΣΙΑΣΗ ΕΡΓΑΣΙΑΣ

ΑΛΕΞΑΝΔΡΟΣ ΠΛΕΣΣΙΑΣ
2022201702012
PCST17012@UOP.GR

Περιεχόμενα Παρουσίασης

- ΠΕΡΙΓΡΑΦΗ
- ΤΕΧΝΗΤΉ ΝΟΗΜΟΣΎΝΗ – ARTIFICIAL INTELLIGENCE (AI)
- ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ – NATURAL LANGUAGE PROCESSING (NLP)
- ΣΥΣΤΗΜΑΤΑ ΑΠΑΝΤΗΣΗΣ ΕΡΩΤΗΣΕΩΝ – QUESTION ANSWERING (QA)
- ΕΡΓΑΛΕΙΑ/ΓΛΩΣΣΑ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΥ ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΗΘΗΚΑΝ
- ΤΥΠΟΙ & ΔΟΜΗ ΜΙΑΣ ΕΡΩΤΗΣΗΣ (ΑΓΓΛΙΚΗ ΓΛΩΣΣΑ)
- ΠΩΣ ΧΡΗΣΙΜΟΠΟΙΩ ΤΟ ΣΥΣΤΗΜΑ ΑΠΑΝΤΗΣΗΣ ΕΡΩΤΗΣΕΩΝ (ΣΑΕ)
- Η ΚΑΤΑΣΚΕΥΗ - ΕΞΗΓΗΣΗ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ ΑΠΑΝΤΗΣΗΣ ΕΡΩΤΗΣΕΩΝ (ΣΑΕ)
- ΠΩΣ ΧΡΗΣΙΜΟΠΟΙΕΙΤΕ ΤΟ ΣΥΣΤΗΜΑ ΑΠΑΝΤΗΣΗΣ ΕΡΩΤΗΣΕΩΝ (ΣΑΕ) ΜΕ ΔΕΔΟΜΕΝΑ ΤΟΥ ΧΡΗΣΤΗ
- ΣΥΜΠΕΡΑΣΜΑΤΑ - ΜΕΛΛΟΝΤΙΚΕΣ ΚΑΤΕΥΘΥΝΣΕΙΣ

ΠΕΡΙΓΡΑΦΗ

ΕΡΓΑΣΙΑΣ

Θέμα εργασίας

Η εργασία μου αναφέρεται στην δημιουργία ενός συστήματος το οποίο δύναται να απαντήσει σε ερώτησες.

Τα συστήματα απάντησης ερωτήσεων ανέκαθεν ήταν ένας βασικός κλάδος της IR και του NLP η οποία έχει να κάνει με συστήματα αυτόματης απάντησης σε ερωτήσεις η οποίες είναι σε ανθρώπινη γλώσσα.

Αρχικά ξεκίνησα με την εύρεση ενός corpus το οποίο περιέχει συλλογές από κείμενα και ζευγάρια ερωτήσεων – απαντήσεων, έπειτα δημιούργησα έναν ταξινομητή ερωτήσεων και τέλος την απάντηση των ερωτήσεων ανάλογα με την κατηγορία της με χρήση Python, NLP και γλωσσολογικών – γραμματικών κανόνων.

ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ – ARTIFICIAL INTELLIGENCE (AI)

ΟΡΙΣΜΟΣ
ΓΙΑΤΙ ΕΙΝΑΙ ΣΗΜΑΝΤΙΚΗ ?
ΠΩΣ ΑΚΡΙΒΩΣ ΛΕΙΤΟΥΡΓΕΙ ?

Ορισμός

Ο όρος τεχνητή νοημοσύνη αναφέρεται στον κλάδο της πληροφορικής ο οποίος ασχολείται με τη σχεδίαση και την υλοποίηση υπολογιστικών συστημάτων που μιμούνται στοιχεία της ανθρώπινης συμπεριφοράς τα οποία υπονοούν έστω και στοιχειώδη ευφυΐα: μάθηση, προσαρμοστικότητα, εξαγωγή συμπερασμάτων, κατανόηση από συμφραζόμενα, επίλυση προβλημάτων κλπ.

Ο Τζον Μακάρθι όρισε τον τομέα αυτόν ως «επιστήμη και μεθοδολογία της δημιουργίας νοούντων μηχανών».

ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ – ARTIFICIAL INTELLIGENCE (AI)

ΟΡΙΣΜΟΣ
ΓΙΑΤΙ ΕΙΝΑΙ ΣΗΜΑΝΤΙΚΗ ?
ΠΩΣ ΑΚΡΙΒΩΣ ΛΕΙΤΟΥΡΓΕΙ ?

Γιατί είναι σημαντική ? (1/2)

- **Η AI αυτοματοποιεί την επαναληπτική μάθηση και την ανακάλυψη μέσω δεδομένων**. Αντί των αυτοματοποιημένων χειροκίνητων έργων, το AI εκτελεί συχνά, μεγάλου όγκου μηχανογραφημένα έργα, αξιόπιστα και χωρίς κόπο. Για αυτόν τον τύπο της αυτοματοποίησης, η ανθρώπινη έρευνα εξακολουθεί να είναι καίριας σημασίας για να εγκατασταθεί το σύστημα και να τεθούν οι κατάλληλες ερωτήσεις.

- **Η AI προσθέτει ευφυΐα στα υπάρχοντα προϊόντα**. Στις περισσότερες περιπτώσεις, το AI δεν πωλείται ως μεμονωμένη εφαρμογή. Πιθανότερο είναι οι ικανότητες του AI να βελτιώνουν τα προϊόντα που ήδη χρησιμοποιείτε, όπως η πρόσθεση της ψηφιακής βοηθού **Siri** ως λειτουργία στα προϊόντα Apple νέας γενιάς.

- **Η AI προσαρμόζεται μέσω προοδευτικών αλγορίθμων εκμάθησης (learning algorithms) ώστε να αφεθούν τα δεδομένα να κάνουν τον προγραμματισμό**. Το AI βρίσκει δομή και κανονικότητες στα δεδομένα, οπότε ο αλγόριθμος αποκτά μια δεξιότητα: Ο αλγόριθμος ταξινομεί ή κατηγοριοποιεί. Έτσι, όπως ο αλγόριθμος μπορεί να διδάξει τον εαυτό του τον τρόπο να παίζει σκάκι, μπορεί και να τον διδάξει ποιο προϊόν να συστήσει στη συνέχεια διαδικτυακά στον πελατη. Και τα μοντέλα προσαρμόζονται όταν τους δίνονται νέα δεδομένα.

Γιατί είναι σημαντική ? (2/2)

- **Η AI αναλύει περισσότερα και βαθύτερα δεδομένα με χρήση neural networks (νευρωνικά δίκτυα) που διαθέτουν πολλά κρυφά επίπεδα**. Η κατασκευή ενός συστήματος με πέντε κρυφά επίπεδα ήταν σχεδόν αδύνατη λίγα χρόνια νωρίτερα. Όλα αυτά έχουν αλλάξει με την απίστευτη ισχύ των υπολογιστών και το μέγεθος των δεδομένων. Χρειάζεσαι **μια πληθώρα δεδομένων για να εκπαιδεύσεις μοντέλα μάθησης σε βάθος, επειδή μαθαίνουν απευθείας από τα δεδομένα**. Με **όσο περισσότερα δεδομένα** τα τροφοδοτείς, **τόσο πιο ακριβή γίνονται**.

- **Η AI επιτυγχάνει απίστευτη ακρίβεια μέσω Deep neural networks** – κάτι που προηγουμένως ήταν αδύνατο. Για παράδειγμα, οι αλληλεπιδράσεις σας με το **Alexa, το Google Search και το Google Photos** βασίζονται στη στο Deep learning – και συνεχίζονται να γίνονται πιο ακριβείς όσο περισσότερο τα χρησιμοποιείτε.

- **Η AI αξιοποιεί στο έπακρο τα δεδομένα**. Όταν οι αλγόριθμοι είναι αυτο-εκπαιδευόμενοι, **τα ίδια τα δεδομένα μπορούν να καταστούν πνευματική ιδιοκτησία**. Οι απαντήσεις βρίσκονται στα δεδομένα, απλά, πρέπει να εφαρμόσετε μεθόδους τεχνητής νοημοσύνης για να τις ανακτήσετε. Επειδή ο ρόλος των δεδομένων είναι πλέον πιο σημαντικός από ποτέ άλλοτε, μπορούν να δημιουργήσουν ένα ανταγωνιστικό πλεονέκτημα.

ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ – ARTIFICIAL INTELLIGENCE (AI)

ΟΡΙΣΜΟΣ
ΓΙΑΤΙ ΕΙΝΑΙ ΣΗΜΑΝΤΙΚΗ ?
ΠΩΣ ΑΚΡΙΒΩΣ ΛΕΙΤΟΥΡΓΕΙ ?

Πώς ακριβώς λειτουργεί ? (1/2)

- **Machine learning:** αυτοματοποιεί την κατασκευή αναλυτικών μοντέλων. Χρησιμοποιεί μεθόδους από τα νευρωνικά δίκτυα (neural networks), τη στατιστική και την επιχειρησιακή έρευνα (operational research) για την **εύρεση κρυφών γνώσεων εντός των δεδομένων** χωρίς να έχει προγραμματιστεί εμφανώς για το πού να εξετάσει ή τι να συμπεράνει.
- **Neural network:** είναι ένας τύπος μηχανικής μάθησης που αποτελείται από αλληλοσυνδεόμενες μονάδες (όπως οι νευρώνες) που επεξεργάζονται τις πληροφορίες ανταποκρινόμενο σε εξωτερικές εισαγωγές δεδομένων, προωθώντας πληροφορίες μεταξύ κάθε μονάδας. Η διαδικασία απαιτεί **πολλαπλές διελεύσεις στα δεδομένα προκειμένου να βρεθούν συνδέσεις και να γίνει εξαγωγή νοήματος από ακαθόριστα δεδομένα**.
- **Deep learning:** Χρησιμοποιεί τεράστια neural networks με πολλά επίπεδα μονάδων επεξεργασίας, αξιοποιώντας τις εξελίξεις στην υπολογιστική ισχύ και τις βελτιωμένες τεχνικές εκπαίδευσης για την μάθηση πολύπλοκων μορφών σε μεγάλες ποσότητες δεδομένων. Οι κοινές εφαρμογές της περιλαμβάνουν την αναγνώριση εικόνας και ομιλίας.

Πώς ακριβώς λειτουργεί ? (1/2)

- **Cognitive computing:** είναι ένα υπο-επίπεδο του AI που στοχεύει σε μια φυσική, ανθρωπόμορφη αλληλεπίδραση με μηχανές. Κατά τη χρήση τεχνικών AI και cognitive computing, ο απώτατος στόχος είναι η προσομοίωση ανθρώπινων αλληλεπιδράσεων από μια μηχανή μέσω της ικανότητας να ερμηνευτούν εικόνες και ομιλία – και να υπάρξει κανονική απάντηση από την μηχανή η οποία έχει ειρμό.

- **Computer vision :** βασίζεται στην αναγνώριση μορφών (pattern recognition) και στο Deep learning ώστε να αναγνωρίζεται τι υπάρχει σε μια εικόνα ή ένα βίντεο. Όταν οι μηχανές μπορούν να επεξεργαστούν, να αναλύσουν και **να κατανοήσουν εικόνες μπορούν να συλλάβουν εικόνες ή βίντεο σε πραγματικό χρόνο και να ερμηνεύσουν τα περιβάλλοντά τους.**

- **Natural Language Processing (NLP):** είναι η ικανότητα των υπολογιστών να αναλύουν, να κατανοούν και να παράγουν ομιλούμενη γλώσσα, συμπεριλαμβανομένης της ομιλίας. Το επόμενο στάδιο στην ΕΦΓ είναι η φυσική γλωσσική αλληλεπίδραση, η οποία επιτρέπει στους ανθρώπους **να επικοινωνούν με υπολογιστές χρησιμοποιώντας την κανονική, καθημερινή γλώσσα για την εκτέλεση καθηκόντων.**

Εμείς σε αυτήν την εργασία θα αναλύσουμε περαιτέρω την Επεξεργασία Φυσικής Γλώσσας που σχετίζεται με το θέμα της εργασίας στην συνέχεια.

ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ – NATURAL LANGUAGE PROCESSING (NLP)

ΕΙΣΑΓΩΓΗ

ΕΡΓΑΣΙΕΣ ΤΗΣ ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ

Εισαγωγή (1/2)

Η Επεξεργασία της Φυσικής Γλώσσας είναι ένα **πεδίο της επιστήμης των υπολογιστών και της γλωσσολογίας που ασχολείται με τις αλληλεπιδράσεις μεταξύ των υπολογιστών και της φυσικής γλώσσας**. Οι τεχνολογίες που βασίζονται στον τομέα της Επεξεργασίας της Φυσικής Γλώσσας αυξάνονται ραγδαία καθώς παρατηρείται όλο και περισσότερο η ανάγκη για τη χρησιμοποίησή τους.

Με την παροχή όλο και περισσότερων διεπαφών μεταξύ ανθρώπου-μηχανής και την εξελιγμένη πρόσβαση σε αποθηκευμένες πληροφορίες, η επεξεργασία της φυσικής γλώσσας **παίζει πλέον ένα σημαντικό ρόλο στην πολύγλωσση κοινωνία της πληροφορίας.**

Οι προκλήσεις της Επεξεργασία της Φυσικής Γλώσσας συχνά εμπλέκονται με κατανόηση της φυσικής γλώσσας, παραγωγή/γέννηση φυσικής γλώσσας, συστήματα διαλόγου ή κάποιο συνδυασμό των παραπάνω.

Η Επεξεργασία της Φυσικής Γλώσσας **έχει να κάνει με τη δημιουργία πραγματικών εφαρμογών χρησιμοποιώντας τεχνικές NLP**. Σε ένα πρακτικό πλαίσιο είναι ανάλογο με τη διδασκαλία μιας γλώσσας σε ένα παιδί.

Εισαγωγή (2/2)

Μετά την έλευση των μεγάλων δεδομένων (big data), η μεγάλη πρόκληση είναι ότι χρειαζόμαστε περισσότερους ανθρώπους που είναι καλοί όχι μόνο με τα δομημένα δεδομένα **αλλά και με τα ημι-δομημένα ή τα μη-δομημένα δεδομένα.** Οι εταιρείες συλλέγουν όλα αυτά τα διαφορετικά είδη δεδομένων για κ**αλύτερη στόχευση των πελατών και ουσιαστικές γνώσεις**.

Είμαστε στην εποχή των πληροφοριών! Δεν μπορούμε ούτε να φανταστούμε τη ζωή μας χωρίς την Google. Χρησιμοποιούμε **Speech Engines** (όπως την Siri, Cortana, Google Voice, Alexa) για τα περισσότερα βασικά πράγματα. Χρησιμοποιούμε **φίλτρα ανεπιθύμητης αλληλογραφίας** (Spam classifiers) για το φιλτράρισμα ανεπιθύμητων ηλεκτρονικών μηνυμάτων. Χρειαζόμαστε τον **ορθογραφικό έλεγχο για τα έγγραφα του Word**. Υπάρχουν πολλά παραδείγματα πραγματικών εφαρμογών NLP γύρω μας.

ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ – NATURAL LANGUAGE PROCESSING (NLP)

ΕΙΣΑΓΩΓΗ

ΕΡΓΑΣΙΕΣ ΤΗΣ ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ

Εργασίες της Επεξεργασία της Φυσικής Γλώσσας

Μερικές από τις πιο συνηθισμένες εργασίες, όπως η κατανόηση λέξεων, φράσεων και ο σχηματισμός γραμματικών προτάσεων που είναι γραμματικά σωστές, είναι πολύ φυσικές για τον άνθρωπο για τον υπολογιστή δεν είναι. Στην Επεξεργασία της Φυσικής Γλώσσας, μερικές από αυτές τις εργασίες είναι οι εξής:

χώρισμα του κειμένου σε λεκτικές μονάδες (**tokenization**), αναγνώριση μερών του λόγου (part of speech tagging - **POS**), σε συντακτική ανάλυση (**parsing**), σε μηχανική μετάφραση (**machine translation**), αναγνώριση ομιλίας (speech recognition), σε αναγνώριση ονομαστικών οντοτήτων (Named Entity Recognition- **NER**), σε παραγωγή φυσικής γλώσσας (**Natural Language generation**), σε κατανόηση της φυσικής γλώσσας (**Natural Language understanding**), απάντηση ερωτήσεων (questions answering), εξαγωγή σχέσεων (**relationship extraction**), ανάλυση συνθήματος (**sentiment analysis**), διαίρεση του γραπτού κείμενου σε μονάδες με ουσιαστικό νόημα (**topic segmentation**), η αποσαφήνιση της έννοιας μια λέξης (**word sense disambiguation**), δημιουργίας περίληψης αυτόματα (automatic summarization), εύρεση όλων των εκφράσεων που αναφέρονται σε μια ίδια οντότητα (**coreference resolution**), αναγνώριση φωνής (**speech recognition**), και κείμενο σε ομιλία (**text-to-speech**)

Τα περισσότερα από αυτά είναι ακόμα δύσκολες προκλήσεις για τους υπολογιστές αλλά σε αυτή την εργασία **θα επικεντρωθούμε στην απάντηση ερωτήσεων (questions answering) και σε όποιες άλλες εργασίες απαιτούνται για την επίτευξη αυτού.**

ΣΥΣΤΗΜΑΤΑ ΑΠΑΝΤΗΣΗΣ ΕΡΩΤΗΣΕΩΝ – QUESTION ANSWERING (QA)

ΟΡΙΣΜΟΣ & ΕΙΔΗ ΕΡΩΤΗΣΕΩΝ
ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΕΡΩΤΗΣΕΩΝ

Ορισμός & Είδη ερωτήσεων

Η απάντηση ερωτήσεων (QA) είναι ένας κλάδος του τομέα της πληροφορικής στους τομείς της ανάκτησης πληροφοριών και της επεξεργασίας φυσικής γλώσσας (NLP), η οποία ασχολείται με την κατασκευή συστημάτων που **απαντούν αυτόματα σε ερωτήσεις που θέτουν οι άνθρωποι σε μια φυσική γλώσσα.**

- **Η ερωτήσεις κλειστού πεδίου** ασχολείται με ερωτήσεις κάτω από έναν συγκεκριμένο τομέα (για παράδειγμα ιατρική) και μπορεί να εκμεταλλευτεί ειδικές γνώσεις για το συγκεκριμένο τομέα, οι οποίες συχνά τυποποιούνται στις οντολογίες. Εναλλακτικά, ο τομέας κλειστού χώρου μπορεί να αναφέρεται σε μια κατάσταση όπου μόνο **περιορισμένο είδος ερωτήσεων είναι αποδεκτές, όπως ερωτήσεις που ζητούν περιγραφικές και όχι διαδικαστικές πληροφορίες.**

- Η **ερώτησης ανοιχτού τομέα** ασχολείται με ερωτήσεις για σχεδόν οτιδήποτε και μπορεί να βασιστεί μόνο σε γενικές οντολογίες και την γνώση του κόσμου. Από την άλλη πλευρά, αυτά τα συστήματα έχουν συνήθως πολύ περισσότερα διαθέσιμα δεδομένα από τα οποία μπορούν να εξάγουν την απάντηση.

Στα πλαίσια της εργασίας να ασχοληθούμε με τις ερωτήσεις ανοιχτού τομέα.

ΣΥΣΤΗΜΑΤΑ ΑΠΑΝΤΗΣΗΣ ΕΡΩΤΗΣΕΩΝ – QUESTION ANSWERING (QA)

ΟΡΙΣΜΟΣ & ΕΙΔΗ ΕΡΩΤΗΣΕΩΝ
ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΕΡΩΤΗΣΕΩΝ

Κατηγοριοποίηση ερωτήσεων

Ένα σύστημα απάντησης ερωτήσεων εξαρτάται πολύ από **ένα καλό corpus**. Επομένως, έχει νόημα να έχουμε ένα **μεγάλο μέγεθος συλλογής** γενικά προσδίδει στην καλή επίδοση του συστήματος, εκτός εάν ο τομέας του ερωτήματος είναι ανεξάρτητος της συλλογή.

Έπειτα ακολουθεί η **εξαγωγή λέξεων-κλειδιών για τον προσδιορισμό του τύπου της ερώτησης**, σε ορισμένες περιπτώσεις, υπάρχουν σαφείς λέξεις που υποδεικνύουν τον τύπο ερωτήματος άμεσα πχ. "Who), "Where" ή "How many", αυτές οι λέξεις λένε στο σύστημα ότι οι απαντήσεις πρέπει να είναι **τύπου "Πρόσωπο", "Θέση" ή "Αριθμός"** αντίστοιχα. *Για παράδειγμα, η λέξη "When" υποδεικνύει ότι η απάντηση πρέπει να είναι τύπου "Date", οπότε έπειτα από NER Parsing επιλέγω μια οντότητα τύπου "DATE".*

Δυστυχώς, μερικές ερωτηματικές λέξεις όπως οι "Which", What" ή "How" δεν δίνουν σαφείς τύπους απαντήσεων. Κάθε μία από αυτές τις λέξεις **μπορεί να αντιπροσωπεύει περισσότερους από έναν τύπους**. Σε καταστάσεις όπως αυτό, **πρέπει να εξεταστούν και άλλες λέξεις στο ερώτημα**. Το πρώτο πράγμα που πρέπει να κάνετε είναι να βρείτε τις λέξεις που μπορούν να υποδηλώνουν το νόημα της ερώτησης.

Τέλος, βρίσκω το σύνολο από προτάσεις που περιέχουν τις λέξεις που υποδηλώνουν το νόημα της ερώτησης και **βρισκω αυτή που περιέχει την απάντηση και έπειτα την φράση της πρότασης που είναι η απάντηση**.

ΕΡΓΑΛΕΙΑ/ΓΛΩΣΣΑ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΥ ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΗΘΗΚΑΝ

ΓΛΩΣΣΑ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΥ PYTHON
ΠΛΑΤΦΟΡΜΑ NLTK
ΒΙΒΛΙΟΘΗΚΗ STANFORD CORENLP

Γλώσσα προγραμματισμού Python (1/2)

Η Python είναι μια **interpreted και αντικειμενοστρεφής γλώσσα προγραμματισμού υψηλού επιπέδου** με δυναμική σημασιολογία. Η απλή και εύκολη σύνταξη της Python δίνει έμφαση στην αναγνωσιμότητα και επομένως μειώνει το κόστος συντήρησης του προγράμματος.

Η Python υποστηρίζει modules και πακέτα, τα οποία ενθαρρύνουν τη διαμόρφωση του προγράμματος και την επαναχρησιμοποίηση του κώδικα.

Ο interpreter της Python και η εκτεταμένη τυποποιημένη βιβλιοθήκη της είναι διαθέσιμα είτε σαν πηγή είτε σε δυαδική μορφή για χρήση σε όλες τις μεγάλες πλατφόρμες και μπορούν να διανεμηθούν ελεύθερα.

Δεδομένου ότι δεν υπάρχει κανένα βήμα για τη μεταγλώττιση, **ο κύκλος επεξεργασίας-δοκιμής-εντοπισμού σφαλμάτων είναι απίστευτα γρήγορος**. Από την άλλη πλευρά, συχνά ο ταχύτερος τρόπος για τον εντοπισμό σφαλμάτων σε ένα πρόγραμμα είναι να προσθέσετε μερικές εκτυπώσεις στον κώδικα: **ο γρήγορος κύκλος επεξεργασίας-δοκιμής-εντοπισμού καθιστά αυτή την απλή προσέγγιση πολύ αποτελεσματική**.

Γλώσσα προγραμματισμού Python (2/2)

Στην Python υπάρχει το ρητό: **Η Python 2.x είναι κληρονομιά, η Python 3.x είναι το παρόν και το μέλλον της γλώσσας**. Ο Guido van Rossum (ο αρχικός δημιουργός της γλώσσας Python) αποφάσισε να σταματήσει η υποστήριξη στην Python 2.x και να υποστηρίζεται πλέον μόνο η Python 3.x.

Εκτός αυτού, πολλές πτυχές της βασικής γλώσσας έχουν προσαρμοστεί ώστε να **είναι πιο εύκολο για τους νεοεισερχόμενους** να μάθουν και να είναι πιο συνεπείς με την υπόλοιπη γλώσσα.

Ο κυριότερος λόγος όμως για την επιλογή της έκδοσης Python 3.5 είναι το NLTK μια βιβλιοθήκη την οποία θα αναλύσουμε στην συνέχεια η οποία παρότι υποστηρίζει τις εκδόσεις Python 2.7, 3.4, 3.5, 3.6 ή 3.7 από προσωπική εμπειρία και έπειτα από πολλές δοκιμές για την ώρα **μόνο στην έκδοση Python 3.5 δεν παρουσιάζει κανένα πρόβλημα και είναι μια από τις πιο δημοφιλείς βιβλιοθήκες της κοινότητας του NLP** (Επεξεργασίας της Φυσικής Γλώσσας).

ΕΡΓΑΛΕΙΑ/ΓΛΩΣΣΑ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΥ ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΗΘΗΚΑΝ

ΓΛΩΣΣΑ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΥ PYTHON
ΠΛΑΤΦΟΡΜΑ NLTK
ΒΙΒΛΙΟΘΗΚΗ STANFORD CORENLP

Πλατφόρμα NLTK (1/2)

Το **NLTK** (Natural Language ToolKit) είναι μια ανοικτού κώδικά βιβλιοθήκη συναρτήσεων, που έχει αναπτυχθεί σε γλώσσα Python, με στόχο την επεξεργασία της φυσικής γλώσσας και ανάπτυξη ανάλογων εφαρμογών.

Η ανάπτυξή του ξεκίνησε το 2001, σαν ένα μέρος του μαθήματος της Υπολογιστικής Γλωσσολογίας του τμήματος Υπολογιστών και Επιστήμης της Πληροφορίας στο πανεπιστήμιο της Πενσυλβανία, κυρίως για την Αγγλική γλώσσα.

Περιλαμβάνει components, δομές δεδομένων και interfaces και συνοδεύεται από οδηγούς, αναφορές και τεχνικές οδηγίες που εξηγούν την ακριβή λειτουργία του αλλά και καθοδηγούν για τη χρήση και την επέκτασή του.

Βασικά του χαρακτηριστικά είναι ότι ενώ παρέχει ένα ευρύ σύνολο από λειτουργίες, παραμένει ένα εργαλείο που ασχολείται με το πεδίο της επεξεργασίας της φυσικής γλώσσας (NLP) και δεν μετατρέπεται σε σύστημα.

Πλατφόρμα NLTK (2/2)

Το NLTK προορίζεται για να υποστηρίξει την έρευνα και τη διδασκαλία της Επεξεργασίας της Φυσικής Γλώσσας. Συνδέεται στενά με τους τομείς της **γλωσσολογίας,** της **γνωστικής επιστήμης,** της **τεχνητής νοημοσύνης** (AI)**,** της **ανάκτησης πληροφοριών** (IR) και της **μηχανικής μάθησης** (Machine Learning).

Ωστόσο, όταν πρόκειται για την ευκολία χρήσης και την εξήγηση των εννοιών, τα αποτελέσματα του NLTK είναι πολύ θετικά. Το NLTK είναι επίσης **πολύ καλό για την εκμάθηση επειδή η καμπύλη μάθησης της Python (στην οποία είναι γραμμένο το NLTK) είναι πολύ γρήγορη**.

Το **NLTK έχει ενσωματώσει τις περισσότερες εργασίες της Επεξεργασίας της Φυσικής Γλώσσας,** είναι πολύ κομψό και εύκολο στη χρήση. Για όλους αυτούς τους λόγους, το NLTK έχει γίνει μια από τις πιο δημοφιλείς βιβλιοθήκες της κοινότητας και για αυτό την επέλεξα για την ανάπτυξη της εργασίας μου.

ΕΡΓΑΛΕΙΑ/ΓΛΩΣΣΑ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΥ ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΗΘΗΚΑΝ

ΓΛΩΣΣΑ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΥ PYTHON
ΠΛΑΤΦΟΡΜΑ NLTK
ΒΙΒΛΙΟΘΗΚΗ STANFORD CORENLP

Βιβλιοθήκη Stanford CoreNLP (1/3)

Το Stanford CoreNLP παρέχει ένα σύνολο τεχνολογικών εργαλείων που έχουν να κάνουν με την ανθρώπινη γλώσσα. Μπορεί να μας δώσει την ρίζα μιας λέξης, τα μέρη το λόγου (πχ. ρήμα, ουσιαστικό κοκ.), αναγνωρίζει οντότητες όπως εταιρείες, ανθρώπους, ημερομηνίες, ώρες και άλλα. Επισημάνει τη δομή μιας πρότασης σε φράσεις και βρίσκει τις συντακτικές εξαρτήσεις.

Ο στόχος του είναι να κάνει πολύ εύκολη την εφαρμογή την γλωσσολογικής ανάλυση ενός κειμένου. Το CoreNLP έχει σχεδιαστεί για να είναι ιδιαίτερα ευέλικτο και επεκτάσιμο. **Το Stanford CoreNLP ενσωματώνει πολλά εργαλεία NLP, συμπεριλαμβανομένου του tagger για τα τμήμα ομιλίας (POS), την αναγνώρισή ονομαστικών οντοτήτων (NER), του parser, του συστήματος ανάλυσης coreference, της ανάλυσης συναίσθημα και την εξαγωγής πληροφοριών.**

Η βασική διανομή παρέχει αρχεία μοντέλων για την ανάλυση καλά επεξεργασμένων αγγλικών, αλλά είναι συμβατός και με μοντέλα για άλλες γλώσσες όπως αραβικά, τα κινέζικα, τα γαλλικά, τα γερμανικά και τα ισπανικά. **Στην εργασία μου όλες οι είσοδοι θα είναι στα Αγγλικά οπότε θα χρησιμοποιήσω το Αγγλικό πακέτο του Stanford CoreNLP**. Η χρήση του θα γίνει μέσω ενός ειδικού **wrapper που έχει η NLTK βιβλιοθήκη** και χρειάζεται Java 1.8+ έκδοση για να τρέξει (το CoreNLP) μέσω NLTK που τρέχει σε Python

Βιβλιοθήκη Stanford CoreNLP (2/3)

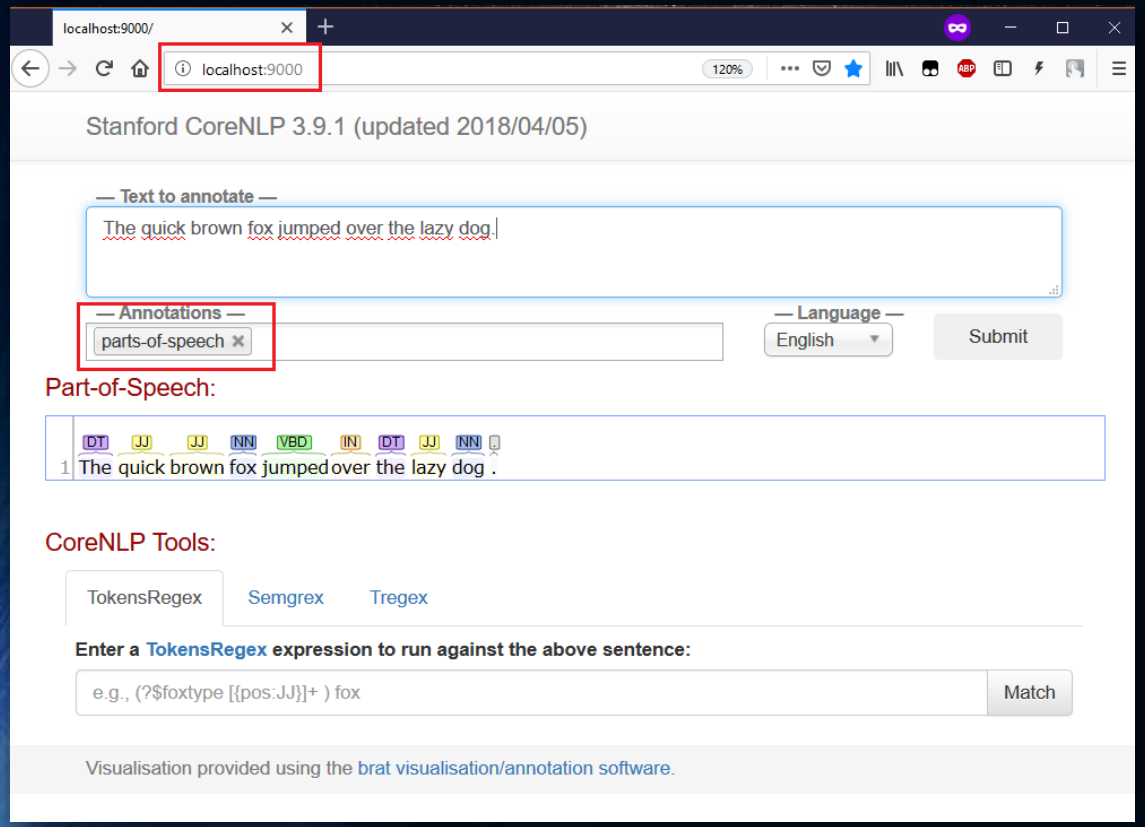
Για την χρήση του CoreNLP αρχικά θα χρειαστεί να κατεβάσουμε την τελευταία έκδοση του CoreNLP, για την εργασία μου χρησιμοποίησα την **έκδοση 3.9.2**. Έπειτα πρέπει να κατεβάσουμε μια έκδοση της γλώσσας **Java μεγαλύτερη από την 1.8** από την επίσημη σελίδα, για να μπορέσουμε αργότερα να τρέξουμε τον Server του CoreNLP. Τέλος, για να **λειτουργήσει ο Server** ανοίγουμε το Command Prompt και τρέχουμε την εντολή:

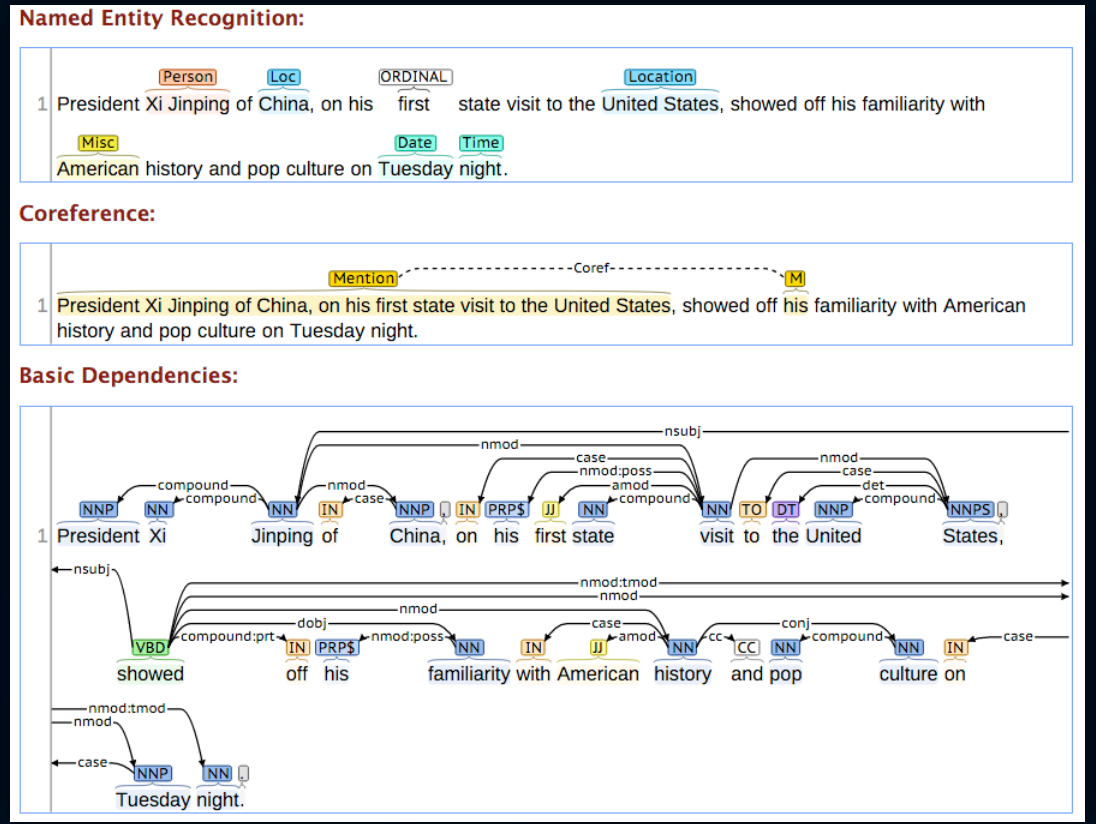
```
‘java -mx6g -cp "*" edu.stanford.nlp.pipeline.StanfordCoreNLPServer -port 9000 -timeout 20000 -annotators pos -annotators ner’
```

στον φάκελο του CoreNLP, ουσιαστικά η εντολή αυτή καλεί την Java με μέγιστη(-mx) δέσμευση μνήμης τα 6GB γενικά χρειαζόμαστε αρκετή μνήμη για την επεξεργασία μεγάλων προτάσεων ώστε να μην μας ‘σκάσει’ το πρόγραμμα, και εκκινεί τον StanfordCoreNLPServer στην θύρα (port) 9000 με μέγιστο όριο τα 20000 χιλιοστά του δευτερολέπτου (milliseconds) για κάθε request και την χρήση annotators για POS και NER.

Αξίζει να αναφερθεί ότι υπάρχει και άλλος **τρόπος να χρησιμοποιήσουμε απευθείας τον parser του Stanford με πιο απλό τρόπο ο οποίος όμως στις επόμενες εκδόσεις του NLTK θα καταργηθεί**. Η λύση του Server αν και είναι λίγο χρονοβόρα μας λύνει τα χέρια διότι αφενός θα δουλεύει και για τις επόμενες εκδόσεις του NLTK αλλά κάνει και πολύ πιο εύκολο η χρήση του μέσω του Browser μας για testing (Βλέπε επόμενη διαφάνεια).

Βιβλιοθήκη Stanford CoreNLP (1/3)





Παραδείγματα τρεξίματος του Stanford CoreNLP

ΤΥΠΟΙ & ΔΟΜΗ ΜΙΑΣ ΕΡΩΤΗΣΗΣ (ΑΓΓΛΙΚΗ ΓΛΩΣΣΑ)

ΤΥΠΟΙ ΕΡΩΤΗΣΕΩΝ
ΔΟΜΗ ΕΡΩΤΗΣΗΣ

Τύποι ερωτήσεων

Κλειστές ερωτήσεις (Closed Questions) - Οι Ναι / Όχι ερωτήσεις καλούνται επίσης κλειστές ερωτήσεις επειδή υπάρχουν **μόνο δύο πιθανές απαντήσεις: Ναι ή Όχι**. Όταν σχηματίζεται ένα Ναι / Όχι ερώτημα, πρέπει να περιλαμβάνει ένα από αυτά τα ρήματα ή κλίσεις αυτών: be , do, have ή ένα μεταβατικό ρήμα. Είναι αδύνατο να υπάρχει ένα Ναι / Όχι ερώτημα χωρίς ένα από αυτά τα ρήματα.

Πχ. Ερώτηση: Are these islands Greek? Απάντηση: Yes. ή Yes, they are. ή Yes, these islands are Greek.

Ανοιχτές ερωτήσεις (Open Questions) - Οι Wh-ερωτήσεις, είναι όλες οι ερωτήσεις οι οποίες ξεκινούν με λέξεις (της Αγγλικής γλώσσας) που αρχίζουν με -Wh και ονομάζονται ανοικτές ερωτήσεις επειδή ο **αριθμός των πιθανών απαντήσεων είναι απεριόριστος**. Οι -wh λέξεις που υποστηρίζονται από το πρόγραμμα είναι οι εξής: **when , how long, how many , how much , how old , who , whom , whose, where, what , how, why και which** . Χρησιμοποιούμε τις λέξεις when, how long, how many, how much και how old όταν **ρωτάμε για κάτι σχετικό με ημερομηνία , ώρα ή κάποιο αριθμό**. Χρησιμοποιούμε τις λέξεις who, whom και whose όταν **ρωτάμε κάτι σχετικό με την ταυτότητα κάποιου**. Χρησιμοποιούμε τη λέξη where όταν **ρωτάμε κάτι σχετικό με κάποια τοποθεσία, πόλη θέση ή χώρα**. Χρησιμοποιούμε τις λέξεις what, how, why και which όταν ρωτάμε για κάτι που χρειάζεται περεταίρω διερεύνηση ή μέθοδο ή διεργασία ή εξήγηση.

Π.χ. Ερώτηση: When does Anna arrive? Απάντηση: She arrives at 10:30

ΤΥΠΟΙ & ΔΟΜΗ ΜΙΑΣ ΕΡΩΤΗΣΗΣ (ΑΓΓΛΙΚΗ ΓΛΩΣΣΑ)

ΤΥΠΟΙ ΕΡΩΤΗΣΕΩΝ
ΔΟΜΗ ΕΡΩΤΗΣΗΣ

Δομή ερώτησης

Δομή ερώτησης κλειστού τύπου:

Ρήμα (προηγούμενης διαφάνειας) + **υποκείμενο** (πχ. you, he, his, she κ.α.) + **ρήμα** + (αντικείμενο) + ? .

*Π.χ. **Does** **he** **live** in New York? ή **Have** **you** **seen** that film?*

Δομή ερώτησης ανοιχτού τύπου:

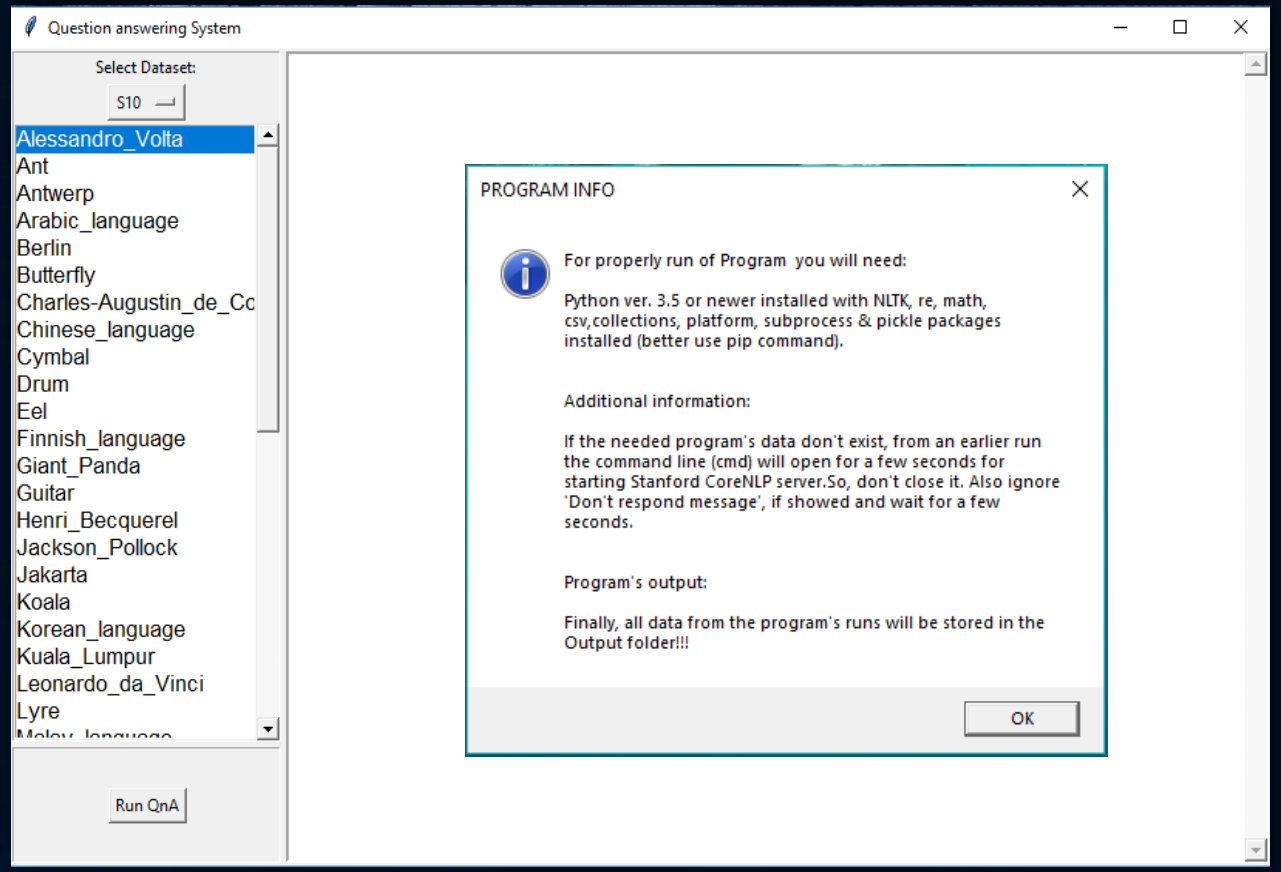
wh λέξη (προηγούμενης διαφάνειας) + **βοηθητικό ρήμα** + **υποκείμενο** (πχ. you, he, his, she κ.α.) + **ρήμα** + ? .

*Π.χ. **Who** **is** **coming** to the party next week? ή **How** **long** **has** **your** teacher **worked** at this school?*

ΠΩΣ ΧΡΗΣΙΜΟΠΟΙΩ ΤΟ ΣΥΣΤΗΜΑ ΑΠΑΝΤΗΣΗΣ ΕΡΩΤΗΣΕΩΝ (ΣΑΕ)

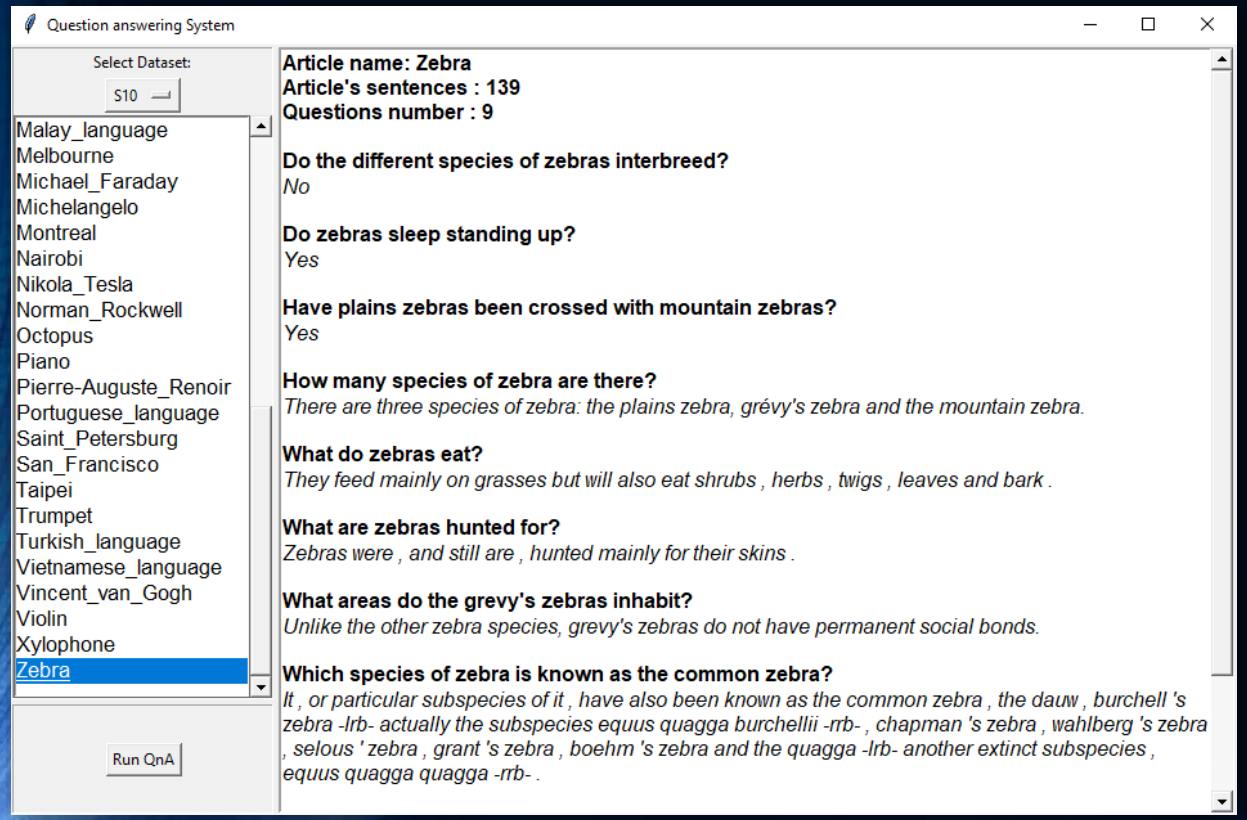
ΤΡΕΞΙΜΟ ΠΡΟΓΡΑΜΜΑΤΟΣ

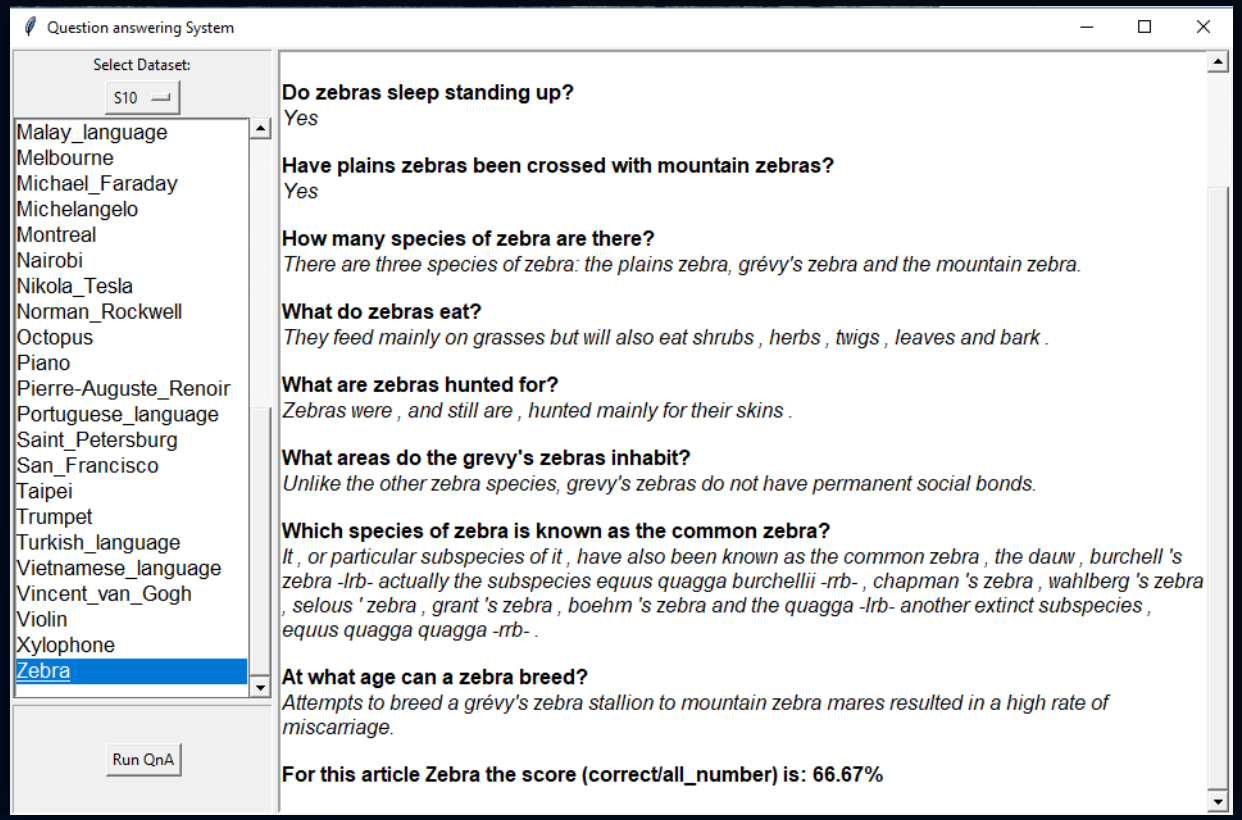
Τρέξιμο προγράμματος (1/2)



Ανοίγουμε τον Φάκελο **'QnA System (postgraduate project)'** και εκτελούμε το αρχείο **'QnA_System START.bat'** το οποίο εκκινεί το γραφικό περιβάλλον του προγράμματος.

Τρέξιμο προγράμματος (1/2)





Η ΚΑΤΑΣΚΕΥΗ - ΕΞΗΓΗΣΗ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ ΑΠΑΝΤΗΣΗΣ ΕΡΩΤΗΣΕΩΝ (ΣΑΕ)

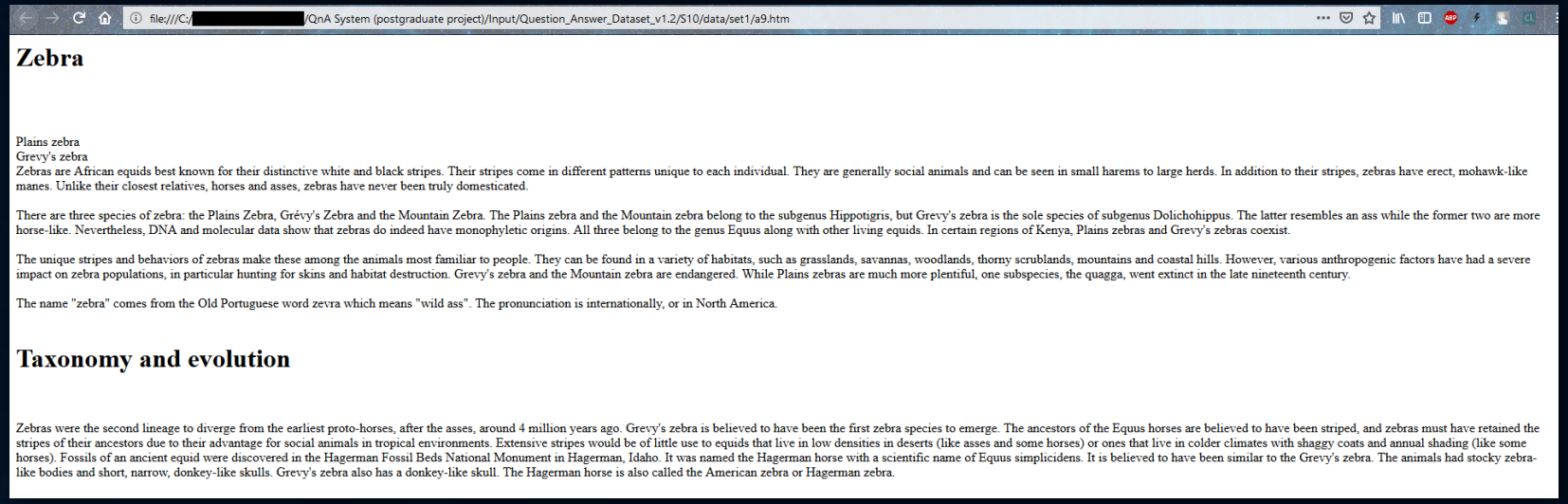
ΕΙΣΟΔΟΙ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ
ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΤΗΣ ΕΙΣΟΔΟΥ
ΑΠΑΝΤΗΣΗ ΕΡΩΤΗΣΕΩΝ - ΔΥΑΔΙΚΕΣ ΕΡΩΤΗΣΕΙΣ
ΑΠΑΝΤΗΣΗ ΕΡΩΤΗΣΕΩΝ - ΠΟΥ ΕΧΟΥΝ ΝΑ ΚΑΝΟΥΝ ΜΕ ΟΝΤΟΤΗΤΕΣ
ΑΠΑΝΤΗΣΗ ΕΡΩΤΗΣΕΩΝ - ΟΛΕΣ ΟΙ ΥΠΟΛΟΙΠΕΣ ΕΡΩΤΗΣΕΙΣ
ΕΞΟΔΟΣ ΤΟΥ ΠΡΟΓΡΑΜΜΑΤΟΣ

Είσοδοι του Συστήματος (1/3)

Στον Φάκελο '**Input/Question_Answer_Dataset_v1.2**' θα βρούμε μια ομάδα από **Datasets** τα οποία έχουν δημιουργηθεί από τον Noah Smith του Πανεπιστημίου του Carnegie Mellon και της Rebecca Hwa του Πανεπιστημίου του Pittsburgh. Ουσιαστικά πρόκειται για **3 datasets το S08, το S09 και το S10** τα οποία δημιουργήθηκαν την Άνοιξη του 2008, 2009 και 2010 αντίστοιχα και το καθένα έχει ένα διαφορετικό σύνολο από Άρθρα τα οποία ο χρήστης μπορεί να επιλέξει.

Ο κάθε φάκελος του κάθε **Dataset** αποτελείτε από έναν φάκελο, τον 'data' και ένα αρχείο, το '**question_answer_pairs.txt**'. Στον φάκελο data εχουμε τα sets τα οποία περιέχουν τα αρχεία με τα άρθρα της Wikipedia σε μορφή ηλεκτρονικής σελίδας (.htm), κειμένου (.txt) και καθαρού κειμένου (.txt.clean) και το αρχείο 'question_answer_pairs.txt' που περιέχει τις ερωτήσεις και τις απαντήσεις

Είσοδοι του Συστήματος (2/3)



file:///C:/[redacted]/QnA System (postgraduate project)/Input/Question_Answer_Dataset_v1.2/S10/data/set1/a9.htm

Zebra

Plains zebra
Grevy's zebra
Zebras are African equids best known for their distinctive white and black stripes. Their stripes come in different patterns unique to each individual. They are generally social animals and can be seen in small harems to large herds. In addition to their stripes, zebras have erect, mohawk-like manes. Unlike their closest relatives, horses and asses, zebras have never been truly domesticated.

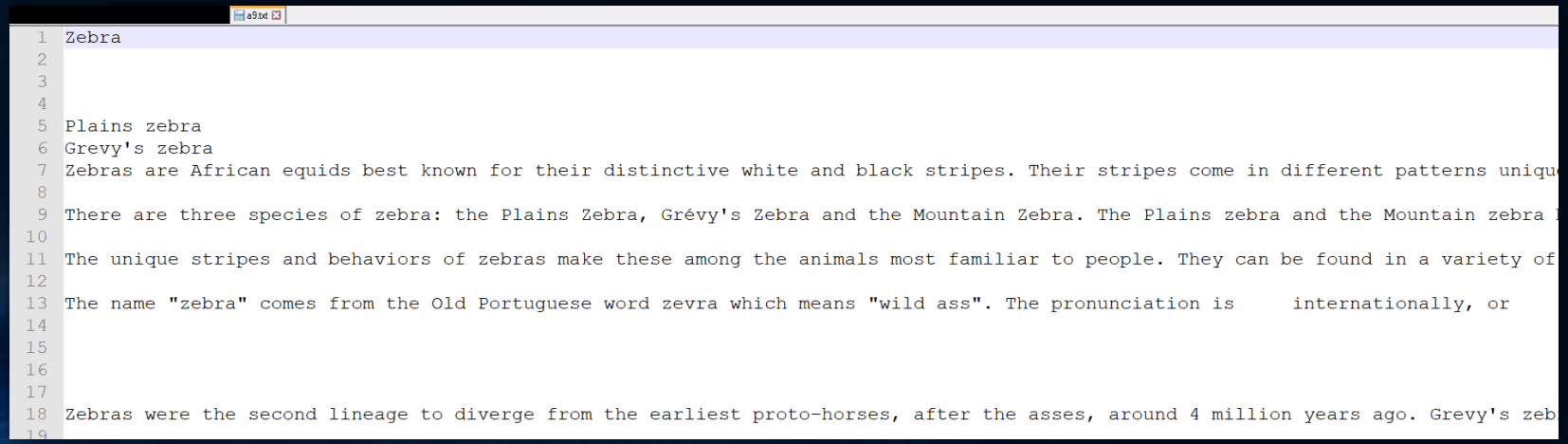
There are three species of zebra: the Plains Zebra, Grévy's Zebra and the Mountain Zebra. The Plains zebra and the Mountain zebra belong to the subgenus Hippotigris, but Grevy's zebra is the sole species of subgenus Dolichohippus. The latter resembles an ass while the former two are more horse-like. Nevertheless, DNA and molecular data show that zebras do indeed have monophyletic origins. All three belong to the genus Equus along with other living equids. In certain regions of Kenya, Plains zebras and Grevy's zebras coexist.

The unique stripes and behaviors of zebras make these among the animals most familiar to people. They can be found in a variety of habitats, such as grasslands, savannas, woodlands, thorny scrublands, mountains and coastal hills. However, various anthropogenic factors have had a severe impact on zebra populations, in particular hunting for skins and habitat destruction. Grevy's zebra and the Mountain zebra are endangered. While Plains zebras are much more plentiful, one subspecies, the quagga, went extinct in the late nineteenth century.

The name "zebra" comes from the Old Portuguese word zevra which means "wild ass". The pronunciation is internationally, or in North America.

Taxonomy and evolution

Zebras were the second lineage to diverge from the earliest proto-horses, after the asses, around 4 million years ago. Grevy's zebra is believed to have been the first zebra species to emerge. The ancestors of the Equus horses are believed to have been striped, and zebras must have retained the stripes of their ancestors due to their advantage for social animals in tropical environments. Extensive stripes would be of little use to equids that live in low densities in deserts (like asses and some horses) or ones that live in colder climates with shaggy coats and annual shading (like some horses). Fossils of an ancient equid were discovered in the Hagerman Fossil Beds National Monument in Hagerman, Idaho. It was named the Hagerman horse with a scientific name of Equus simplicidens. It is believed to have been similar to the Grevy's zebra. The animals had stocky zebra-like bodies and short, narrow, donkey-like skulls. Grevy's zebra also has a donkey-like skull. The Hagerman horse is also called the American zebra or Hagerman zebra.



[redacted] a9.txt

```
1  Zebra
2
3
4
5  Plains zebra
6  Grevy's zebra
7  Zebras are African equids best known for their distinctive white and black stripes. Their stripes come in different patterns uniqu
8
9  There are three species of zebra: the Plains Zebra, Grévy's Zebra and the Mountain Zebra. The Plains zebra and the Mountain zebra
10
11 The unique stripes and behaviors of zebras make these among the animals most familiar to people. They can be found in a variety of
12
13 The name "zebra" comes from the Old Portuguese word zevra which means "wild ass". The pronunciation is     internationally, or
14
15
16
17
18 Zebras were the second lineage to diverge from the earliest proto-horses, after the asses, around 4 million years ago. Grevy's zeb
19
```

Είσοδοι του Συστήματος (3/3)

```
1161 Zebra|Do the different species of zebras interbreed?|No|easy|easy|data/set1/a9
1162 Zebra|Do zebras sleep standing up?|yes|easy|easy|data/set1/a9
1163 Zebra|Do zebras sleep standing up?|Yes|easy|easy|data/set1/a9
1164 Zebra|Have plains zebras been crossed with mountain zebras?|yes|easy|easy|data/set1/a9
1165 Zebra|Have plains zebras been crossed with mountain zebras?|Yes|easy|easy|data/set1/a9
1166 Zebra|How many species of zebra are there?|three|medium|medium|data/set1/a9
1167 Zebra|How many species of zebra are there?|Three|medium|medium|data/set1/a9
1168 Zebra|What do zebras eat?|mainly grass|medium|hard|data/set1/a9
1169 Zebra|What do zebras eat?|Grasses, shrubs, herbs, twigs, leaves, and bark|medium|hard|data/set1/a9
1170 Zebra|What are zebras hunted for?|mainly for their skins|medium|medium|data/set1/a9
1171 Zebra|What are zebras hunted for?|Skins|medium|medium|data/set1/a9
1172 Zebra|What areas do the Grevy's Zebras inhabit?|semi-arid grasslands of Ethiopia and northern Kenya|hard|hard|da
1173 Zebra|What areas do the Grevy's Zebras inhabit?||hard||data/set1/a9
1174 Zebra|Which species of zebra is known as the common zebra?|Plains Zebra (Equus quagga, formerly Equus burchelli)
1175 Zebra|Which species of zebra is known as the common zebra?|Plains Zebra|hard|medium|data/set1/a9
1176 Zebra|At what age can a zebra breed?|five or six|hard|medium|data/set1/a9
1177 Zebra|At what age can a zebra breed?|5 or 6|hard|hard|data/set1/a9
```

Η ΚΑΤΑΣΚΕΥΗ - ΕΞΗΓΗΣΗ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ ΑΠΑΝΤΗΣΗΣ ΕΡΩΤΗΣΕΩΝ (ΣΑΕ)

ΕΙΣΟΔΟΙ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ
ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΤΗΣ ΕΙΣΟΔΟΥ
ΑΠΑΝΤΗΣΗ ΕΡΩΤΗΣΕΩΝ - ΔΥΑΔΙΚΕΣ ΕΡΩΤΗΣΕΙΣ
ΑΠΑΝΤΗΣΗ ΕΡΩΤΗΣΕΩΝ - ΠΟΥ ΕΧΟΥΝ ΝΑ ΚΑΝΟΥΝ ΜΕ ΟΝΤΟΤΗΤΕΣ
ΑΠΑΝΤΗΣΗ ΕΡΩΤΗΣΕΩΝ - ΟΛΕΣ ΟΙ ΥΠΟΛΟΙΠΕΣ ΕΡΩΤΗΣΕΙΣ
ΕΞΟΔΟΣ ΤΟΥ ΠΡΟΓΡΑΜΜΑΤΟΣ

Προεπεξεργασία της εισόδου

Αφού **ο χρήστης έχει επιλέξει ένα συγκεκριμένο άρθρο**, το συστήμα ελέγχει εάν έχει εκτελεστεί ήδη, από παλιότερη εκτελεστή και διαβάζει τα απαραίτητα δεδομένα αλλιώς ξεκινάει η απαραίτητη προ-επεξεργασία δηλαδή η δημιουργία των **συντακτικών δένδρων** για κάθε μια πρόταση του αρχικού κείμενου του άρθρου και των ερωτήσεων με την χρήση του **CoreNLPPOSTagger**. Έπειτα **βρισκω όλες τις οντότητες (NERs)** της κάθε πρότασης του κειμένου και τις αποθηκεύω για μελλοντική χρήση με τη βοήθεια του πάλι του CoreNLP και του **CoreNLPNERTagger**.

Αποθηκεύω τα εξής αρχεία:

- context_parsed_trees.txt: pickle (δυαδικό αρχείο) με την λίστα των συντακτικών δέντρων της κάθε πρότασης του άρθρου

- context_sentences_ners.txt: pickle (δυαδικό αρχείο) με την λίστα των οντοτήτων της κάθε πρότασης του άρθρου, όπου κάθε πρόταση είναι μια λίστα από δυάδες (λέξη, NER ετικέτα).

- questions_parsed_trees.txt: pickle (δυαδικό αρχείο) με την λίστα των συντακτικών δέντρων της κάθε ερωτήσεις του άρθρου

Είναι προτιμότερο η προ-επεξεργασία να γίνει μια και καλή πριν αρχίσει η διαδικασία της απάντησης των ερωτήσεων και η αποθήκευση των δεδομένων της (προ-επεξεργασίας) ώστε ναι κερδίζουμε χρόνο στις επανεκτελέσεις.

Η ΚΑΤΑΣΚΕΥΗ - ΕΞΗΓΗΣΗ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ ΑΠΑΝΤΗΣΗΣ ΕΡΩΤΗΣΕΩΝ (ΣΑΕ)

ΕΙΣΟΔΟΙ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ
ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΤΗΣ ΕΙΣΟΔΟΥ
ΑΠΑΝΤΗΣΗ ΕΡΩΤΗΣΕΩΝ - ΔΥΑΔΙΚΕΣ ΕΡΩΤΗΣΕΙΣ
ΑΠΑΝΤΗΣΗ ΕΡΩΤΗΣΕΩΝ - ΠΟΥ ΕΧΟΥΝ ΝΑ ΚΑΝΟΥΝ ΜΕ ΟΝΤΟΤΗΤΕΣ
ΑΠΑΝΤΗΣΗ ΕΡΩΤΗΣΕΩΝ - ΟΛΕΣ ΟΙ ΥΠΟΛΟΙΠΕΣ ΕΡΩΤΗΣΕΙΣ
ΕΞΟΔΟΣ ΤΟΥ ΠΡΟΓΡΑΜΜΑΤΟΣ

Απάντηση ερωτήσεων - Δυαδικές ερωτήσεις

Δυαδικές ερωτήσεις (Ναι/Όχι) είναι οι ερωτήσεις στις οποίες **απαντούμε με Ναι ή Όχι**. Εχουμε δυο (2) μεγάλες υποκατηγορίες τις **Καταφατικές** και της **Αρνητικές** ερωτήσεις. Γίνεται έλεγχος για το εάν η ερώτηση είναι Καταφατική ή Αρνητική. Αυτή είναι μια σημαντική εργασία διότι στην περίπτωση της Αρνητικής ερώτησης θα χρειαστεί αντιστροφή της απάντησης της ερώτησης.

Η διεργασία της απάντησης έχει να κάνει με την **μετατροπή της ερωτήσεις σε δήλωση** αυτό επιτυγχάνεται με το φόρτωμα του αρχείου 'Negative Yes No Words (15).txt', της **αφαίρεσης του βοηθητικού ρήματος,** την **αφαίρεση του χαρακτήρα του ερωτηματικού,** την εφαρμογή **tokenization** (δηλαδή σπάσιμο της πρότασης σε λίστα λέξεων), την **αφαίρεση των** και τέλος την εφαρμογή του **SnowballStemmer. Όπως** παρατηρείτε και εδώ εφαρμόζουμε linguistic – γλωσσολογική προσέγγιση.

Μόλις η ερώτηση γίνει δήλωση εφαρμόζουμε **πανόμοια επεξεργασία και στις προτάσεις του άρθρου** και ψάχνουμε για το **εάν όλα τα tokens της δήλωσης είναι υποσύνολο των tokens κάποιας πρότασης του άρθρου τότε υπάρχει απάντηση και είναι Ναι αλλιώς είναι Όχι**. Τέλος **εάν χρειάζεται κάνω αντιστροφή απάντησης και επιστρέφω την απάντηση (Ναι ή Όχι) μαζί με ένα μήνυμα** (την δήλωση δηλαδή τροποποιημένη με Yes στην αρχή της ή No ανάλογα με την απάντηση).

Η ΚΑΤΑΣΚΕΥΗ - ΕΞΗΓΗΣΗ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ ΑΠΑΝΤΗΣΗΣ ΕΡΩΤΗΣΕΩΝ (ΣΑΕ)

ΕΙΣΟΔΟΙ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ
ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΤΗΣ ΕΙΣΟΔΟΥ
ΑΠΑΝΤΗΣΗ ΕΡΩΤΗΣΕΩΝ - ΔΥΑΔΙΚΕΣ ΕΡΩΤΗΣΕΙΣ
ΑΠΑΝΤΗΣΗ ΕΡΩΤΗΣΕΩΝ - ΠΟΥ ΕΧΟΥΝ ΝΑ ΚΑΝΟΥΝ ΜΕ ΟΝΤΟΤΗΤΕΣ
ΑΠΑΝΤΗΣΗ ΕΡΩΤΗΣΕΩΝ - ΟΛΕΣ ΟΙ ΥΠΟΛΟΙΠΕΣ ΕΡΩΤΗΣΕΙΣ
ΕΞΟΔΟΣ ΤΟΥ ΠΡΟΓΡΑΜΜΑΤΟΣ

Απάντηση ερωτήσεων - Που έχουν να κάνουν με οντότητες (1/2)

Οι ερωτήσεις που έχουν να κάνουν με οντότητες (NERs) είναι αυτές που έχουν σαν απάντηση τους μια ημερομηνία, μια ώρα, έναν αριθμό, ένα πρόσωπο, μια τοποθεσία, μια πόλη ή μια χώρα είναι δηλαδή οι ερωτήσεις που έχουν να κάνουν με τις ερωτήσεις που περιχέουν τις λέξεις when, how long, how many, how much, how old, who, whom, whose και where.

Για της ερωτήσεις που έχουν να κάνουν με τις λέξεις **when, how long, how many, how much και how old εχουμε τις παραμέτρους DATE, TIME, NUMBER και ORDINAL**, με τις λέξεις **who, whom και whose εχουμε την παράμετρο PERSON** και για την λέξη **where εχουμε τις παραμέτρους LOCATION, CITY, STATE_OR_PROVINCE και COUNTRY.** Οι λίστες με τις ετικέτες που δίνονται σαν παράμετρος έχουν να κάνουν με το τι είδους οντότητα είναι η απάντηση της εκάστοτε ερώτηση.

Στη συνάρτηση wh_familiar πρώτα ψάχνω για ταυτίσεις - ταιριάσματα ανάμεσα στην λίστα context_sentences_ners και τα NER tags που ζητάει η ερώτηση και τα αποθηκεύω σε μια λίστα αυτή με τις πιθανές απατήσεις ανεξαρτήτου ερώτησης. **Δηλαδή εάν η ερώτηση μου έχει την λέξη who ψάχνω για οντότητες PERSON στην λίστα context_sentences_ners που έχουν σαν tag το PERSON** και αποθηκεύω το/τα ζευγάρια αριθμό πρότασης, λέξη/εις (που αντιστοιχούν στην οντότητα που ψάχνω) .

Π.χ. Εάν ψάχνω για NER tag = PERSON μπορεί να έχω αυτά τα πιθανά αποτελέσματα: [[4, (George, Bush)] , [10, (Kennedy)]] δηλαδή στην 4η πρόταση του άρθρου να έχω George, Bush & στην 10η πρόταση του άρθρου την Kennedy σαν πιθανές απαντήσεις.

Απάντηση ερωτήσεων - Που έχουν να κάνουν με οντότητες(2/2)

Οπότε τα σενάρια των ταυτίσεων – πιθανών απαντήσεων που επεξεργάζομαι είναι ο: να υπάρχει 1 ταύτιση, από >1 ταυτίσεις και καμία δηλαδή ο ταυτίσεις:

Εάν η ταύτιση είναι **μονό 1 τότε αυτή είναι και η απάντηση** και την επιστρέφω ενώ **εάν έχω μηδέν (0) ταυτίσεις τότε επιστρέφω None** .

Εάν όμως έχω από >1 ταυτίσεις τότε πρέπει **επεξεργαστώ την ερώτηση** δηλαδή να εφαρμόσω την αφαίρεση του χαρακτήρα του ερωτηματικού, την εφαρμογή tokenization, την αφαίρεση των stopwords και τέλος την εφαρμογή του SnowballStemmer και **έπειτα να εξετάσω τις προτάσεις των πιθανών απαντήσεων μια μέχρι να βρω το καθαρό υποσύνολο** της ερωτήσεις σε μια από αυτές των πιθανών απαντήσεων – προτάσεων. Αυτή η διαδικασία **παίρνει το συντακτικό δένδρο της κάθε πιθανής πρότασης και μέσω τις συνάρτησης tree_search επιστρέφει τα υποδένδρα τύπου S** δηλαδή τις υποπροτάσεις της προτάσεις, πολλές φορές μια πρόταση μπορεί να έχει πολλές υποπροτάσεις. Αφού γίνει αυτό στο επόμενο βήμα γίνεται έλεγχος στα φύλλα της υποπροτάσεις της κάθε πιθανής πρότασης ώστε να **βρεθεί το υποσύνολο της ερωτήσεις σε μια από αυτές** (γίνεται όμοια επεξεργασία με αυτή της ερωτώσης) και **επιστέφω το ζευγάρι οντότητα και την σχετική υποπρόταση εάν δεν βρεθεί επιστρέφω None.**

Η ΚΑΤΑΣΚΕΥΗ - ΕΞΗΓΗΣΗ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ ΑΠΑΝΤΗΣΗΣ ΕΡΩΤΗΣΕΩΝ (ΣΑΕ)

ΕΙΣΟΔΟΙ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ
ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΤΗΣ ΕΙΣΟΔΟΥ
ΑΠΑΝΤΗΣΗ ΕΡΩΤΗΣΕΩΝ - ΔΥΑΔΙΚΕΣ ΕΡΩΤΗΣΕΙΣ
ΑΠΑΝΤΗΣΗ ΕΡΩΤΗΣΕΩΝ - ΠΟΥ ΕΧΟΥΝ ΝΑ ΚΑΝΟΥΝ ΜΕ ΟΝΤΟΤΗΤΕΣ
ΑΠΑΝΤΗΣΗ ΕΡΩΤΗΣΕΩΝ - ΟΛΕΣ ΟΙ ΥΠΟΛΟΙΠΕΣ ΕΡΩΤΗΣΕΙΣ
ΕΞΟΔΟΣ ΤΟΥ ΠΡΟΓΡΑΜΜΑΤΟΣ

Απάντηση ερωτήσεων - Όλες οι υπόλοιπες Ερωτήσεις (1/2)

Όποια ερώτηση δεν ανήκει στις προηγούμενες περιπτώσεις ή εάν κάποια από τις προηγούμενες συναρτήσεις δεν βρει την απάντηση (return None)τότε καλείτε η συνάρτηση **wh_questions**.

Αρχικά βρίσκει όλα τα ρήματα της ερώτησης και όλους τους πιθανούς χρόνους τους που μπορεί να έχουν και γίνεται μια αναζήτηση αυτών στις προτάσεις του άρθρου. Ο λόγος που επιλέγω τα ρήμα είναι διότι τα ρήματα από γραμματικής άποψης περιγράφουν μια δράση ή κατάσταση και είναι το βασικό στοιχείο μιας πρότασης, Υποκείμενο - Ρήμα - Αντικείμενο - όχι πάντα με αυτή την σειρά αλλά πρέπει να υπάρχουν και τα τρία είδη μαζί**. Εάν βρω μια ταύτιση την επιστρέφω αλλιώς καλώ την συνάρτηση find_best_answer** για να βρω την καλύτερη πιθανή απάντηση.

Αρχικά **βρίσκω τα ρήμα** της ερώτησης ψάχνοντας στο συντακτικό της δένδρο για **φύλλα με ετικέτες VB, VBD, VBG, VBN, VBP ή VBZ** και κρατάω μόνο αυτά μου δεν ανήκουν στο αρχείο 'Yes No Words (18).txt' το οποίο περιέχει τα βοηθητικά ρήματα που έχουν οι ερωτήσεις.

Για κάθε ρήμα της ερώτησης, ψάχνω για όλες τους διαθέσιμους χρόνους του στο περιεχόμενο του αρχείου 'verbs (8567).txt' το οποίο διαθέτει όλους τους χρόνους για **8567 ρήματα είναι ένα είδος thesaurus**. Έπειτα, ψάχνω στις υποπροτάσεις των προτάσεων του άρθρου για υποπροτάσεις **που περιέχουν οποιαδήποτε ρήμα** της ερώτησης και τις **αποθηκεύω σε μια λίστα**.

Απάντηση ερωτήσεων - Όλες οι υπόλοιπες Ερωτήσεις (2/2)

Εάν η λίστα αυτή **έχει μέγεθος 1 τότε η υποπρόταση αυτή είναι και η απάντηση και επιστέφεται** αλλιώς εάν η λίστα έχει **μέγεθός >1 τότε καλώ την find_best_answer** με παραμέτρους την ερώτηση και τις πιθανές υποπροτάσεις ενώ εάν η λίστα έχει **μέγεθος 0 τότε καλώ την find_best_answer** με παραμέτρους την ερώτηση και όλες τις προτάσεις του άρθρου

Η συνάρτηση **find_best_answer** αυτό που κάνει είναι να υπολογίζει το cosine similarity ανάμεσα στην **διάνυσμα της ερώτησης και τα διανύσματα των παραμέτρων της** είτε αυτά είναι πιθανές απαντήσεις είτε όλες τις προτάσεις του άρθρου ανάλογα με την κλήση της, η **πρόταση με το μεγαλύτερο score – βαθμολογία επιστρέφεται απάντηση σαν.**

Τέλος, η συνάρτηση **wh_questions** επιστρέφει είτε την μια και καλύτερη **απάντηση** είτε την απάντηση της find_best_answer μαζί με **ένα σχόλιο** το οποίο δεν εμφανίζεται τον χρήστη αλλά γράφεται μόνο στο αρχείο με της απαντήσεις (για λόγους μελλοντικής ανάπτυξης του προγράμματος)

Η ΚΑΤΑΣΚΕΥΗ - ΕΞΗΓΗΣΗ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ ΑΠΑΝΤΗΣΗΣ ΕΡΩΤΗΣΕΩΝ (ΣΑΕ)

ΕΙΣΟΔΟΙ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ
ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΤΗΣ ΕΙΣΟΔΟΥ
ΑΠΑΝΤΗΣΗ ΕΡΩΤΗΣΕΩΝ - ΔΥΑΔΙΚΕΣ ΕΡΩΤΗΣΕΙΣ
ΑΠΑΝΤΗΣΗ ΕΡΩΤΗΣΕΩΝ - ΠΟΥ ΕΧΟΥΝ ΝΑ ΚΑΝΟΥΝ ΜΕ ΟΝΤΟΤΗΤΕΣ
ΑΠΑΝΤΗΣΗ ΕΡΩΤΗΣΕΩΝ - ΟΛΕΣ ΟΙ ΥΠΟΛΟΙΠΕΣ ΕΡΩΤΗΣΕΙΣ
ΕΞΟΔΟΣ ΤΟΥ ΠΡΟΓΡΑΜΜΑΤΟΣ

Έξοδος του προγράμματος

Αφού ο χρήστης έχει επιλέξει ένα κείμενο στο **δεξιό μέρος της οθόνης εμφανίζονται** (βλέπε τρέξιμο προγράμματός) **όλες οι ερωτήσει και απαντήσεις** και το ποσοστό επιτυχίας. Το πρόγραμμα όμως έχει σχεδιαστή έτσι ώστε να αποθηκεύει τις απαντήσεις και το ποσοστό επιτυχίας – ακρίβειας στα αρχεία '**system_answers.txt**' και '**system_answers_eval.txt**' αντίστοιχα για μελλοντική εξέλιξη του προγράμματος και αξιολόγηση του συστήματος.

ArticleTitle	Question	Answer	Answer'sComment
Zebra	do the different species of zebras interbreed?	No	Not the different species of zebras interbreed.
Zebra	do zebras sleep standing up?	Yes	Zebras sleep standing up.
Zebra	have plains zebras been crossed with mountain zebras?	Yes	Plains zebras been crossed with mountain zebras.
Zebra	how many species of zebra are there?	There are three species of zebra: the plains zebra, grévy's zebra and the mountain zebra.	Found in
Zebra	what do zebras eat?	They feed mainly on grasses but will also eat shrubs , herbs , twigs , leaves and bark .	Found in wh_questions
Zebra	what are zebras hunted for?	Zebras were , and still are , hunted mainly for their skins .	Found in wh_questions -> Best possible
Zebra	what areas do the grevy's zebras inhabit?	Unlike the other zebra species, grevy's zebras do not have permanent social bonds.	Found in wh
Zebra	which species of zebra is known as the common zebra?	It , or particular subspecies of it , have also been known as the common zebra , the	
Zebra	at what age can a zebra breed?	Attempts to breed a grévy's zebra stallion to mountain zebra mares resulted in a high rate of miscarriage.	

```
For this article Zebra the score (correct/all_number) is: 66.67%
```

ΠΩΣ ΧΡΗΣΙΜΟΠΟΙΕΙΤΕ ΤΟ ΣΥΣΤΗΜΑ ΑΠΑΝΤΗΣΗΣ ΕΡΩΤΗΣΕΩΝ (ΣΑΕ) ΜΕ ΔΕΔΟΜΕΝΑ ΤΟΥ ΧΡΗΣΤΗ

ΤΡΟΠΟΠΟΙΗΣΗ ΤΟΥ ΑΡΧΕΙΟΥ ΕΙΣΟΔΟΥ ΑΠΟ ΤΟΝ ΧΡΗΣΤΗ

Τροποποίηση του αρχείου εισόδου από τον χρήστη

Ο χρήστης που επιθυμεί να τρέξει το πρόγραμμα με δικά του δεδομένα δεν έχει παρά να πάει στον φάκελο Input/Question_Answer_Dataset_v1.2 να επιλέξει οποιαδήποτε από τους τρεις φακέλου (S08, S09 & S10) θέλει δεν παίζει ρόλο και να ανοίξει με έναν κειμενογράφο το αρχείο '**question_answer_pairs.txt**' και να **προσθέσει στο τέλος ή σε όποιο σημείο θέλει μια σειρά για κάθε ερώτηση του με βάση την κεφαλίδα που υπάρχει ήδη:**

'ArticleTitle | Question | Answer | DifficultyFromQuestioner | DifficultyFromAnswerer | ArticleFile'

δηλαδή ένα όνομα που θέλει να δώσει ο χρήστης π.χ. MyRun1 ακολουθούμενο από την **Ερώτηση** την **Απάντησης** εάν την ξέρει (ώστε να εφαρμόσει benchmark) αλλιώς NULL έπειτα να γράψει **NULL, NULL** και τέλος την **θέση του κειμένου**. Όλα πρέπει να χωρίζονται με τον χαρακτήρα '|'αυτό επαναλαμβάνεται για κάθε ερώτηση και επίσης μπορεί να προστεθούν παραπάνω από ένα αρχεία κειμένου και ArticleTitle αντίστοιχα.

Για να **εκτελέσει** και να δει τα αποτελέσματα του ο χρήστης πρέπει να εκτελέσει το πρόγραμμα να **επιλέξει Dataset S08, S09 ή S10 ανάλογα με το αρχείο το οποίο τροποποίησε πιο πριν** και να δει τα αποτελέσματα στο δεξιό μέρος της οθόνης και στα αρχεία που παράχθηκαν.

ΣΥΜΠΕΡΑΣΜΑΤΑ - ΜΕΛΛΟΝΤΙΚΕΣ ΚΑΤΕΥΘΥΝΣΕΙΣ

ΣΥΜΠΕΡΑΣΜΑΤΑ
ΜΕΛΛΟΝΤΙΚΕΣ ΚΑΤΕΥΘΥΝΣΕΙΣ

Συμπεράσματα

Λάβουμε υπόψιν ότι το σύστημα μου προκύπτει από έναν **συνδυασμό γλωσσολογίας-γραμματικής και NLP τεχνολογιών**, δύναται να απαντήσει ερωτήσεις **ανοιχτού τύπου ερωτήσεις,** η οποίες **δεν απαιτούν την συλλογή πρώτα ερωτήσεων και απαντήσεων** (και μόνο για συγκεκριμένο domain), **χωρίς να χρησιμοποιηθεί ένα σύστημα που θα χρειάζεται εκπαίδευση** και με **γρήγορη απάντηση** (χωρίς την προεπεξεργασία) μπορεί να θεωρηθεί καλό λόγο την οικουμενικότητας του και σίγουρα όχι τέλειο δηλαδή να αγγίζει το 83% οπότε είναι καλό για μια γρήγορη απάντηση, αξίζει να αναφερθεί ότι ο **έλεγχος για τις απαντήσεις ήταν αυστηρός και υπολογίστηκε μόνο με τις 100% όμοιες απαντήσεις σαν σωστές (δεν υπήρξε καθόλου ελαστικότητα)**.

Έπειτα από δοκιμή της 'εξαιρετικής' βιβλιοθήκης Keras (https://keras.io/) η οποία ασχολείται με υψηλού επιπέδου Νευρωνικά Δίκτυα (Neural Networks – NN) και είναι γραμμένη σε Python και της βιβλιοθήκης TensorFlow (https://www.tensorflow.org μια πλατφόρμα ανοιχτού κώδικα για μηχανική μάθηση και την **χρήση RNN και LSTM** συστημάτων πάνω στο **bAbI Project** παρατήρησα ότι αν έκανα εκπαίδευση με το 70% του Dataset και testing στο άλλο 30% να βλέπω **ποσοστά επιτυχίας 70% με 90% και σε άγνωστα δεδομένα να αγγίζει το 50%**. **Αυτός ήταν και ο λόγος που δεν επέλεξα την κατεύθυνση των Νευρωνικών Δικτύων όπου θα κατέληγα σε ένα μεγάλο ποσοστό αλλά επι τις ουσίας πλασματικό – μη ρεαλιστικό.**

ΣΥΜΠΕΡΑΣΜΑΤΑ - ΜΕΛΛΟΝΤΙΚΕΣ ΚΑΤΕΥΘΥΝΣΕΙΣ

ΣΥΜΠΕΡΑΣΜΑΤΑ
ΜΕΛΛΟΝΤΙΚΕΣ ΚΑΤΕΥΘΥΝΣΕΙΣ

Μελλοντικές κατευθύνσεις

Η **αύξηση της διαθέσιμης μνήμης RAM** που χρειάζεται το σύστημα, λόγο της έλλειψης της διαθέσιμης μνήμης RAM **απέκλεισα κάποια άρθρα τα οποία θα ήθελα να συμπεριλάβω αλλά δεν μπορούσα διότι είχαν μεγάλες πρότασης** και το CoreNLP ήθελε πάνω από τα 6 GB που είχα διαθέσιμα (βλέπε διακόπτη –mx6g) για να μην καταρρεύσει (το CoreNLP) ακόμα **λόγο αυτής της έλλειψης δεν υπήρχε η υποστήριξη των Dependency Parse tree & Coref** στο σύστημα μου.

Η χρήση **Dependency Parse Tree & Coref** τα οποίο υποστηρίζονται από το CoreNLP, το οποίο βρίσκει τις εξαρτήσεις που υπάρχουν ανάμεσα στις προτάσεις του κειμένου μέσω των ρημάτων και θα διεύρυνε – μεγάλωνε την ακρίβεια του συστήματος ειδικά στην κατηγορία όλες οι υπόλοιπες ερωτήσεις (από την οποία έχω χάσει το μεγαλύτερο ποσοστό της ακρίβειας μου). **Θα διόρθωνε το τωρινό σύστημα το οποίο βρίσκει την απάντηση της ερώτησης εάν η πρόταση που περιέχει την απάντηση περιέχει το ρήμα της** ερώτησης πράγμα που δεν ισχύει πάντα.

Τέλος, η **αγορά ενός domain και τι στήσιμο ενός Server** ο οποίος θα έτρεχε το πρόγραμμα **online** σε συνδυασμό με ένα interface φιλικό προς τον χρήστη και **η υποστήριξη της Ελληνικής γλώσσας (όπως η Speech Engine MAIC της MLS)** και γενικά η συλλογή extra feedback.

ΕΡΩΤΗΣΕΙΣ - ΑΠΟΡΙΕΣ - ΠΑΡΑΤΗΡΗΣΕΙΣ