Code how to use document loaders

- load euviroument variables, like Open AI API key
- για PDF's θα πρέπει να ανεβάσουμε τον PDF loader
- Να φορτώσουμε τα data καλώντας την μέθοδο φόρτωσης

We load a bunch of PDF's in the folder in workspace what the ween by that. 2.21 win video 1

Document splitting
Split documents in to smaller chunks.
Δεν είναι εύκολη διαδικασία, είναι κάποιες λεπτομέρειες που έχουν μεγάλη επίπτωση.
Το splitting γίνεται μετά την φόρτωση των data, στο document format. και ΠΡΙΝ πάει στο vector store.
‼ Το πως θα σπάσεις τα κείμενα σε chunks είναι ιδιαίτερα σημαντικό. Πολλές φορές αν δεν μπορέσεις να κάνεις καλό χωρισμό στα chunks, μπορεί να καταλήξει με ένα μέρος μιας πρότασης σε ένα chunk και με το άλλο μέρος της πρότασης σε ένα άλλο chunk. Και μετά όταν προσπαθήσεις να απαντήσεις σε ερωτήσεις το chunk δεν θα περιέχει την σωστή πληροφορία και δεν θα μπορεί να απαντήσει την ερώτηση σωστά.

Example Splitter
Μια σωστή μέθοδος για να σπας τα chunks είναι να είναι ενός συγκεκριμένου αριθμού, όπου ένα κομμάτι του τέλους του ενός chunk να είναι η αρχή του άλλου chunk (overlap)
Αυτό βοηθάει στο να δημιουργηθεί μια έννοια συνέχειας καθώς υπάρχει το ίδιο κομμάτι περιεχομένου στο τέλος ενός κομματιού και στην αρχή του άλλου.

Στο langChain υπάρχουν πολλοί τύποι splitters που ο κάθε ένας έχει διαφορετική λεπτομέρεια ή χαρακτηριστικά. Επίσης μπορεί να υπάρχει ποικιλία στον τρόπο που μετράνε το μήκος των κομματιών.
Επειδής υπάρχουν κάποια που χρησιμοποιούν μικρότερα μοντέλα για να προσδιορίσουν πότε μπορεί να τελειώσει μια πρόταση.
Άλλα ένας άλλος παράγοντας στο πως να χωρίσεις τα

chunks είναι τα metadata. Η διατήρηση των ίδιων metadata σε όλα τα chunks, αλλά και η προσθήκη νέων κομματιών metadata όταν είναι σχετικό είναι κάτι που κοιτάνε κάποιοι document splitting.

Η διαίρεση των κομματιών μπορεί συχνά να είναι ειδική ανάλογα με το τύπο εγγράφων με τον οποίο εργαζόμαστε. Τέλος έχουμε splitters που διαχωρίζουν με διαφορετικούς seperators για διαφορετικές γλώσσες προγραμματισμού όπως Python, Ruby, C κλπ
Και αυτοί λαμβάνουν υπόψη για την κάθε διαφορετική γλώσσα, τους διαφορετικούς seperators που έχουν.

Metadata
Είναι σαν μια επικεφαλίδα ή μια κατηγορία που περιγράφει & παρέχει πληροφορίες για το σύνολο δεδομένων.
Σε ένα κείμενο που έχει χωριθεί σε chunk τα metadata αυτά λειτουργούν ως κοινά χαρακτηριστικά που διατηρούν συνέπεια σε όλα τα κομμάτια.

Μπορούμε να συνδιάσουμε
τεχνικές splitter;