Πανεπιστήμιο Ιωαννίνων,
Τμήμα Μηχανικών Η/Υ & Πληροφορικής

Δεύτερη Σειρά Ασκήσεων

Η προθεσμία για την δεύτερη σειρά ασκήσεων είναι την Δευτέρα 23 Δεκεμβρίου μέχρι το τέλος της ημέρας. Κάνετε turn-in τον κώδικα σας, με οδηγίες για το πώς τρέχει (αν χρειάζεται), και την αναφορά σας. Η αναφορά θα πρέπει να έχει λεπτομερείς παρατηρήσεις για τα αποτελέσματα σας. Για καθυστερημένες υποβολές ισχύει η πολιτική στην σελίδα του μαθήματος. Λεπτομέρειες για το turn-in στη σελίδα Ασκήσεις του μαθήματος. Αν δεν μπορείτε να κάνετε turnin στείλετε την εργασία σας μέσω email. Θα υπάρξει προφορική εξέταση της Άσκησης. Οι προγραμματιστικές ασκήσεις μπορούν να γίνουν σε ομάδες των δύο ατόμων.

Ερώτηση 1

Μία power-law κατανομή ορίζεται ως $P(X = x) = (a - 1)x^{-a}$, όπου a είναι ο εκθέτης της κατανομής. Σας δίνεται ένα σύνολο από παρατηρήσεις $Z = \{z_1, \dots, z_n\}$ που έχουν παραχθεί από μία power-law κατανομή. Χρησιμοποιήστε την Maximum Likelihood Estimation τεχνική που περιγράψαμε στην τάξη για να βρείτε τον εκθέτη της power-law κατανομής που ταιριάζει (fits) τα δεδομένα των παρατηρήσεων.

Ερώτηση 2 (Singular Value Decomposition)

Για την άσκηση αυτή θα πειραματιστείτε με την εφαρμογή του Singular Value Decomposition. Κατεβάσετε από το dataset “20 Newsgroups” του sklearn τα δεδομένα για τα topics 'talk.politics.mideast', και 'rec.sport.baseball' (υπάρχει παράδειγμα στο φροντιστήριο). Χρησιμοποιήσετε το TfidfVectorizer(stop_words='english',min_df=1) για να εξάγετε τα features, αφαιρέσετε τα κείμενα με μηδενικά features, και πάρετε τον τελικό πίνακα δεδομένων M.

Υπολογίσετε το SVD decomposition αυτού του πίνακα χρησιμοποιώντας την βιβλιοθήκη της Numpy, και κάνετε τα εξής:

1. Υπολογίσετε το μέσο tf-idf score της κάθε λέξης (την μέση τιμή των στηλών του M). Κάνετε ένα scatter plot αυτού του διανύσματος με το πρώτο δεξί singular vector (αυτό που αντιστοιχεί στο μεγαλύτερο singular value), και υπολογίσετε το pearson correlation coefficient μεταξύ των δύο διανυσμάτων. Τι παρατηρείτε?
   Το πρώτο δεξί singular vector είναι το διάνυσμα στο οποίο προβάλουμε τα κείμενα (τις γραμμές του M) για να πάρουμε το πρώτο αριστερό singular vector. Τι συμπεραίνετε για τις τιμές του πρώτου αριστερού singular vector?
2. Αναπαραστήσετε τα κείμενα σαν δισδιάστατα σημεία χρησιμοποιώντας τα δύο πρώτα αριστερά singular vectors (αυτά που αντιστοιχούν στα δύο μεγαλύτερα singular values) και κάνετε plot τα σημεία.
   Κάνετε το ίδιο plot δίνοντας διαφορετικό χρώμα στα σημεία που αντιστοιχούν σε κάθε topic. Τι παρατηρείτε? Τι συμπεραίνετε για τις τιμές του δεύτερου αριστερού singular vector?

3. Εφαρμόστε τον k-means στον πίνακα M και δημιουργήσετε τον πίνακα σύγχυσης με τα topics. Εφαρμόσετε μετά ξανά τον k-means στα δισδιάστατα σημεία από το προηγούμενο ερώτημα. Τι παρατηρείτε?

Παραδώσετε τον κώδικα σας και την αναφορά με τις παρατηρήσεις σας.

Ερώτηση 3 (Συστήματα συστάσεων)

Ο στόχος αυτής της άσκησης είναι να πειραματιστείτε με αλγορίθμους για συστήματα συστάσεων.

Θα χρησιμοποιήσετε ένα dataset με συνταγές από το Kaggle. Θα κατεβάσετε τα train δεδομένα. Για να αποκτήσετε πρόσβαση θα πρέπει να δημιουργήσετε ένα λογαριασμό. Χρησιμοποιήστε το email της σχολής για να κάνετε λογαριασμό γιατί θα το χρειαστείτε και στην επόμενη σειρά ασκήσεων.

Τα δεδομένα περιγράφονται στην σελίδα του Kaggle. Είναι σε json format, και θα χρησιμοποιήσετε την βιβλιοθήκη json της Python για να τα φορτώσετε. Περιέχουν λίστες από υλικά για διάφορες συνταγές και την κατηγορία του φαγητού (εθνικότητα).

Θα δουλέψετε με την κατηγορία “italian” που είναι η πιο δημοφιλής. Χρησιμοποιώντας τα δεδομένα θα παράγετε ένα δυαδικό πίνακα M, όπου οι γραμμές θα είναι οι συνταγές και οι στήλες τα υλικά. Μπορείτε να γράψετε δικό σας κώδικα γι αυτό ή να χρησιμοποιήσετε κάποια έτοιμη βιβλιοθήκη από το sklearn (π.χ., το CountVectorizer). Αν επιλέξετε τον δεύτερο τρόπο πιθανότατα θα χρειαστεί μια προ-επεξεργασία των δεδομένων ώστε να αφαιρέσετε τα σημεία στίξεως, και κάθε υλικό να γίνει μία μόνο λέξη.

Διαλέξτε τυχαία ένα σύνολο R απο 1000 συνταγές και αφαιρέσετε τυχαία ένα από τα υλικά από κάθε συνταγή στο R. Ο στόχος είναι να βρείτε αυτά τα υλικά χρησιμοποιώντας αλγόριθμους για συστήματα συστάσεων. Για κάθε συνταγή r στο R, και κάθε υλικό i που δεν υπάρχει ήδη στην συνταγή r θα υπολογίσετε ένα σκορ $s(r,i)$ και θα προτείνετε τα K υλικά με το μεγαλύτερο σκορ. Θα δοκιμάσετε τους παρακάτω αλγόριθμους για να υπολογίσετε το σκορ $s(r,i)$:

1. **Most Popular (MP):** Χρησιμοποιήστε την δημοφιλία του υλικού ως το σκορ (ανεξάρτητο από την συνταγή r).
2. **User-based Collaborative Filtering** (**UCF**): Για κάθε συνταγή r υπολογίστε το σύνολο $B_N(r)$ με τις N πιο όμοιες συνταγές. Για την ομοιότητα θα χρησιμοποιήσετε το Jaccard Similarity. Έστω $J(r,r')$ η ομοιότητα μεταξύ r και r'. Το σκορ $s(r,i)$ θα το υπολογίσετε ως:

$$s(r,i) = \frac{\sum_{r' \in B_N(r)} J(r,r') M[r',i]}{\sum_{r' \in B_N(r)} J(r,r')}$$

3. **Item-based Collaborative Filtering** (**ICF**): Για κάθε υλικό i που δεν χρησιμοποιεί η συνταγή r υπολογίστε το σύνολο $B_N(i)$ με τα N πιο όμοια υλικά. Για την ομοιότητα θα χρησιμοποιήσετε το Jaccard Similarity. Έστω $J(i,i')$ η ομοιότητα μεταξύ i και i'. Το σκορ $s(r,i)$ θα το υπολογίσετε ως:

$$s(r,i) = \frac{\sum_{i' \in B_N(i)} J(i,i') M[r,i']}{\sum_{i' \in B_N(i)} J(i,i')}$$

4. **Singular Value Decomposition** (**SVD**)**:** Εφαρμόστε το Singular Value Decomposition στον πίνακα M, και κρατήστε τα N μεγαλύτερα singular vectors για να πάρετε ένα rank-N πίνακα M_N. Στη συνέχεια χρησιμοποιήστε την τιμή $s(r,i) = M_N(r,i)$ για να διαλέξετε τα K υλικά με το μεγαλύτερο σκορ που θα προτείνετε για την συνταγή r.

Για την αξιολόγηση και σύγκριση των αλγορίθμων θα χρησιμοποιήσετε Precision@K δηλαδή το ποσοστό των συνταγών για τις οποίες ο αλγόριθμος βρίσκει το κρυμμένο υλικό μέσα στις K πρώτες συστάσεις.

Για να αποφασίσετε την τιμή της παραμέτρου N, για τους αλγορίθμους που χρησιμοποιούν την παράμετρο, θέσετε $K = 1$ και για διαφορετικές τιμές του N μεταξύ [1,100] υπολογίσετε το Precision@K. Κάνετε μια γραφική παράσταση με τις μετρήσεις σας. Επιλέξετε το N που σας δίνει το καλύτερο precision. Στη συνέχεια για το N που επιλέξατε για κάθε αλγόριθμο, υπολογίσετε το Precision@K για $K = 1,2,5,10$. Βάλετε όλα τα αποτελέσματα μαζί σε ένα πίνακα και σε μία γραφική παράσταση, και συγκρίνετε τους αλγορίθμους.

Επιλέξετε τυχαία 100 από τις συνταγές στο R και εξετάσετε χειρωνακτικά τα αποτελέσματα των αλγορίθμων (για $K = 1$). Σε κάποιες περιπτώσεις κάποια υλικά είναι ουσιαστικά τα ίδια (π.χ., διαφορετικοί τύποι από μακαρόνια, diced tomatoes αντί για tomatoes, ή minced garlic αντί για garlic). Πως αλλάζει το Precision@1 αν κάνετε αυτές τις διορθώσεις? Παραδώσετε ένα excel αρχείο (ή ένα dataframe) με την αξιολόγηση η οποία θα αποτελείται από τριάδες της μορφής: υλικό, σύσταση, αξιολόγηση (το τελευταίο θα το συμπληρώσετε χειρωνακτικά) .

Θα πρέπει να παραδώσετε τα ακόλουθα:

- Όλο τον κώδικα που θα γράψετε για την επεξεργασία των δεδομένων, την υλοποίηση των αλγορίθμων, και την αξιολόγηση.
- Μια αναφορά που θα περιγράφει τι κάνατε, θα συγκρίνει τους διαφορετικούς αλγορίθμους, και θα σχολιάζει τα αποτελέσματα.

Bonus I: Προσθέσετε μία ακόμη κατηγορία στα δεδομένα σας και κάνετε ξανά την παραπάνω διαδικασία (για $K = 1$). Εξετάσετε αν βελτιώνονται ή χειροτερεύουν τα αποτελέσματα.

Bonus II (για τους φαν της μαγειρικής): Επιλέξετε 100 νέες συνταγές τυχαία (από τις οποίες δεν έχουμε αφαιρέσει κάποιο υλικό). Εφαρμόσετε σε αυτές τον καλύτερο από τους αλγορίθμους που υλοποιήσατε για $K = 1$. Εξετάστε το υλικό που προτείνεται και αξιολογήστε αν είναι μια λογική σύσταση. Παραδώστε την αξιολόγηση, με ένα σχολιασμό για τις κρίσεις σας, και το Precision@1.

Σημειώσεις:

- Χρησιμοποιήστε τις συναρτήσεις της python για τον υπολογισμό αποστάσεων ή πράξεων με πίνακες, είναι πολύ πιο γρήγορες. Ο υπολογισμός των σκορ μπορεί να γίνει με πολλαπλασιασμούς πινάκων. Θα προσμετρηθεί θετικά αν χρησιμοποιήσετε αυτές τις λειτουργείες για να κάνετε τους υπολογισμούς.

Ερώτηση 4 (Clustering)

Στην ερώτηση αυτή θα εξασκηθείτε στην εφαρμογή αλγορίθμων ομαδοποίησης (clustering). Θα χρησιμοποιήσουμε τα ίδια δεδομένα όπως στην Ερώτηση 3. Επιλέξετε 3 κατηγορίες με περίπου το ίδιο μέγεθος και κάνετε ακριβώς την ίδια επεξεργασία όπως πριν για να πάρετε ένα δυαδικό πίνακα. Εφαρμόστε τους αλγόριθμους k-means, και agglomerative clustering για $k = 3$. (Μην πάρετε πολύ μεγάλες κατηγορίες για να μπορεί να τρέξει ο Agglomerative αλγόριθμος.) Δημιουργήστε τον πίνακα σύγχυσης και υπολογίστε precision και recall όπως στην τάξη. Τι παρατηρείτε?

Είναι πιθανό οι συνταγές να μην γίνονται clustered με τρόπο που συμφωνεί με τις κατηγορίες. Για τα δεδομένα που δημιουργήσατε στο προηγούμενο βήμα, χρησιμοποιήστε τον k-means αλγόριθμο και δημιουργήστε το silhouette plot για να αποφασίσετε για τον αριθμό των clusters. Για τον αριθμό που επιλέξατε, εξετάσετε τα clusters του k-means, και χρησιμοποιώντας τα κέντρα των clusters και τις κατηγορίες προσπαθήσετε να εξηγήσετε σε τι είδους συνταγές αντιστοιχεί το κάθε cluster.

Παραδώσετε τον κώδικα σας και μια αναφορά με λεπτομερή σχολιασμό και ανάλυση των αποτελεσμάτων σας.

Ερώτηση 5 (Bonus)

Συνδυάστε την δουλειά που κάνατε για τις Ερωτήσεις 3 και 4, και χρησιμοποιήσετε το clustering για να βελτιώσετε τον αλγόριθμο συστάσεων. Προτείνετε τον αλγόριθμο, υλοποιήσετε τον και συγκρίνετε με τον αντίστοιχο αλγόριθμο από την Ερώτηση 3.