ΠΑΠΑΔΟΠΟΥΛΟΣ ΝΙΚΟΛΑΟΣ 4140

ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ
ΔΕΥΤΕΡΟ ΣΕΤ ΑΣΚΗΣΕΩΝ

ΑΝΑΦΟΡΑ ΔΕΥΤΕΡΗΣ ΑΣΚΗΣΗΣ

ΒΗΜΑ 1

Το αρχείο που προκύπτει από την δικιά μου εξόρυξη και επεξεργασία δεδομένων για τις εταιρείες και τους χρήστες του yelp dataset με μια πρώτη ματιά μοιάζει με το ιδανικό αποτέλεσμα που μας δίνεται. Αν σκρολάρει λίγο κάποιος στο αρχείο my_pruned_data.csv και το συγκρίνει με το αρχείο pruned_data.csv θα προσέξει ότι εμφανίζονται και στα δύο αρχεία οι ίδιοι κωδικοί χρηστών με την ίδια σειρά στην αρχή, ενώ η μορφολογία (user, business, rating) είναι ακριβώς ίδια. Παρόλα αυτά, το δικό μου αρχείο είναι κατά περίπου 2000kb μεγαλύτερο σε μέγεθος. Αυτό μάλλον συμβαίνει επειδή στο τελικό μου αποτέλεσμα πιθανόν να υπάρχουν χρήστες που δεν έχουν τουλάχιστον 15 reviews για businesses του συνόλου, και αντιστοίχως μπορεί κάποιες από τις επιχειρήσεις του συνόλου να μην έχουν reviews από τουλάχιστον 15 χρήστες του συνόλου. Πριν εντάξω έναν χρήστη στο σύνολο, κάνω μεν έλεγχο για το αν έχει κάνει τουλάχιστον 15 reviews σε επιχειρήσεις του Toronto, αυτό όμως δεν εξασφαλίζει ότι τουλάχιστον 15 από αυτές θα βρεθούν στο σύνολο εν τέλει. Κάτι ανάλογο ισχύει από πλευράς επιχειρήσεων.

ΒΗΜΑ 3

Για τον αλγόριθμο UCF τυπώνω ανά σειρά το rating που αφαιρέθηκε και την αντίστοιχη πρόβλεψη του αλγορίθμου. Από αυτό παρατηρείται είναι ότι, για k = 10 γίνονται αρκετά ακριβείς προβλέψεις τις περισσότερες φορές. Βέβαια, υπάρχουν και προβλέψεις που απέχουν κατά 2 ή και παραπάνω μονάδες από το πραγματικό rating, το οποίο είναι αναμενόμενο, αφού πάντα θα υπάρχουν εξαιρέσεις και περιπτώσεις στις οποίες ο χρήστης έκανε ένα rating που δεν είναι συνεπές με τα δεδομένα που έχουμε για αυτόν.

ΒΗΜΑ 4

Παρατηρώ ακριβώς τα ίδια για τον αλγόριθμο ICF με k = 5.

ΒΗΜΑ 5

Ο αλγόριθμος SVD (k = 100), με τα αποτελέσματα που τυπώνω, παρουσιάζει απογοητευτικά αποτελέσματα. Οι περισσότερες προβλέψεις απέχουν πολύ από την πραγματικότητα, ενώ παρουσιάζονται αρκετές τιμές στο 0, το οποίο σίγουρα δε βοηθάει την ακρίβεια, εφόσον όλα τα ratings είναι τουλάχιστον 1. Μάλιστα, όταν έτρεξα τον αλγόριθμο με τα μικρότερα k που δοκίμασα στους προηγούμενους, τα αποτελέσματα ήταν ακόμα χειρότερα. Άρα, ο συγκεκριμένος αλγόριθμος, ακόμα και με περισσότερο χρόνο εκτέλεσης για μεγαλύτερα k επιτυγχάνει χειρότερα αποτελέσματα από τους δύο προηγούμενους.

ΒΗΜΑ 6

Από τις γραφικές παραστάσεις των σφαλμάτων των τριών μεθόδων φαίνεται μία μεγάλη ομοιότητα στην αποτελεσματικότητα μεταξύ των UCF και ICF και τονίζεται η ανακρίβεια της μεθόδου SVD. Οι γραφικές των UCF και ICF έχουν σχεδόν ίδια μορφή. Η πτώση είναι πολύ απότομη για τις μικρές τιμές του k, ενώ το σφάλμα φαίνεται να σταθεροποιείται, για την UCF μετά την τιμή 50 και για την ICF μετά την τιμή 20. Αξίζει να γίνει εκ νέου γραφική για τις δύο αυτές μεθόδους εστιάζοντας στις μικρότερες τιμές του k. Σε αυτήν φαίνεται ότι η ICF ξεκινάει με λιγάκι μεγαλύτερο σφάλμα, αλλά όσο το k πλησιάζει στην τιμή 10, η διαφορά μικραίνει. Μπορεί κανείς να πει ότι αν θέλουμε να συνδιάσουμε ταχύτητα και ακρίβεια με τον καλύτερο δυνατό τρόπο, γι αυτές τις δύο μεθόδους η τιμή k = 10 είναι ιδανική. Αν δωθεί μεγαλύτερη σημασία στην ακρίβεια, ίσως να αξίζει η τιμή k = 50 για την UCF και k = 20 για την ICF. Μεγαλύτερες τιμές δεν έχουν νόημα, καθώς επιφέρουν μηδαμινή βελτίωση σφάλματος. Οι καλύτερερες τιμές για τον SVD είναι από 100 και πάνω (η καμπύλη φαίνεται να έχει λιγάκι πτωτική πορεία ακόμα και μετά το 100), όμως πάντα ο συγκεκριμένος αλγόριθμος θα έχει ανακριβή αποτελέσματα για τη συγκεκριμένη εφαρμογή, καθώς το σφάλμα δεν φαίνεται να μπορεί να πέσει ούτε κάτω από το 2.

Στο γράφημα με όλες τις μεθόδους μαζί (και με τα baselines), θα ήθελα να τονίσω το γεγονός ότι οι UCF και ICF φαίνεται να ταυτίζονται με τα δύο baselines για k >= 10, το οποίο επιβεβαιώνει τα συμπεράσματα της παραπάνω παραγράφου.