ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικόν και Καποδιστριακόν
Πανεπιστήμιον Αθηνών
ΙΔΡΥΘΕΝ ΤΟ 1837

Σχολή Οικονομικών και Πολιτικών Επιστημών
Τμήμα Επικοινωνίας και Μέσων Μαζικής Ενημέρωσης

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΨΗΦΙΑΚΑ ΜΕΣΑ ΕΠΙΚΟΙΝΩΝΙΑΣ ΚΑΙ ΠΕΡΙΒΑΛΛΟΝΤΑ ΑΛΛΗΛΕΠΙΔΡΑΣΗΣ»

Μελέτη και υλοποίηση περιβάλλοντος παραγωγής συστάσεων άρθρων σε έναν ενημερωτικό ιστοχώρο

ΤΟΥΜΑΝΙΔΗΣ ΗΛΙΑΣ

Διπλωματική εργασία που κατατίθεται ως μέρος των απαιτήσεων του Προγράμματος Μεταπτυχιακών Σπουδών «Ψηφιακά Μέσα Επικοινωνίας και Περιβάλλοντα Αλληλεπίδρασης»

Επιβλέπων: Κωνσταντίνος Μουρλάς

Αθήνα, Φεβρουάριος, 2021

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή της εργασίας μου κ. Κωνσταντίνο Μουρλά Αναπληρωτής Καθηγητής στο Τμήμα Επικοινωνίας και Μέσων Μαζικής Ενημέρωσης, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών, για την πολύτιμη καθοδήγησή του και τις συμβουλές του καθόλη την διάρκεια συγγραφής της διπλωματικής μου εργασίας.

Με την καθοδήγηση του κύριου Μουρλά και την συμπαράσταση σου, κατάφερα να δημιουργήσω έναν αλγόριθμο παραγωγής συστάσεων άρθρων, κάτι το οποίο μου φάνταζε ως ακατόρθωτο στις αρχές των μεταπτυχιακών σπουδών μου στο πρόγραμμα ‘’Ψηφιακών Μέσων και Περιβαλλόντων Επικοινωνίας’’.

Επιπλέον, θα ήθελα να ευχαριστήσω την οικογένεια μου, η οποία με κράτησε αισιόδοξο όσες στιγμές πίστευα ότι δεν θα τα καταφέρω και μου έδωσε δύναμη να συνεχίσω, ακόμα και όταν εγώ δεν πίστευα ότι μπορούσα.

Τέλος, ένα μεγάλο ευχαριστώ και στους φίλους μου, οι οποίοι μου προσέφεραν στιγμές γέλιου στις στιγμές που αμφέβαλλα.

Σύνοψη

Η μηχανική μάθηση έχει σημειώσει τρομακτικά άλματα προόδου τις τελευταίες δεκαετίες και πλέον έχει καταφέρει να θεωρείται αναπόσπαστο κομμάτι της καθημερινότητας μας. Για του λόγου του αληθές, η μηχανική μάθηση έχει ξεπεράσει πλέον τα καθαρά όρια της τεχνολογίας και θεωρείται βασικό γρανάζι σε πολλούς τομείς. Ένας από αυτούς τους τομείς πλέον, είναι και η δημοσιογραφία.

Πιο συγκεκριμένα, όπως κάθε ιστοσελίδα που ασχολείται με το εμπόριο θέλει να προσφέρει την καλύτερη δυνατή υπηρεσία στους πελάτες της, έτσι και οι σύγχρονες, διαδικτυακές, ειδησεογραφικές ιστοσελίδες, επιθυμούν να κάνουν το περιεχόμενο τους όσο πιο ελκυστικό γίνεται, προκειμένου να αναπτύξουν μια σχέση εμπιστοσύνης με τους χρήστες τους, με απώτερο σκοπό την παραμονή τους στην ιστοσελίδα και βέβαια την επιστροφή τους.

Για να πετύχουν τον συγκεκριμένο στόχο, η εξατομίκευση περιεχομένου είναι από τα πιο σημαντικά εργαλεία. Η εξατομίκευση περιεχομένου είναι γνωστή παγκοσμίως πλέον, καθώς γίγαντες της τεχνολογίας και των media, όπως είναι η Amazon, το Netflix, και το Spotify, χρησιμοποιούν ειδικούς αλγόριθμους. Αυτοί οι αλγόριθμοι αναλύουν τις προτιμήσεις και την πλοήγηση των χρηστών και δημιουργούν ένα προσωποποιημένο περιεχόμενο.

Σκοπός αυτής της εργασίας, είναι να εφαρμόσουμε την ιδέα της εξατομίκευσης περιεχομένου σε μια ειδησεογραφική ιστοσελίδα. Προκειμένου να το καταφέρουμε αυτό, δημιουργήθηκε ένας αλγόριθμος παραγωγής συστάσεων, ο οποίος είναι βασισμένος στο υβριδικό φιλτράρισμα. Με τον τρόπο αυτό, και αυτό εξορύχθησαν παραπάνω από 1000 άρθρα, αναλύθηκα τα στοιχεία τους, προκειμένου να καθοριστούν οι μεταξύ τους συσχετίσεις. Στη συνέχεια, χρήστες βαθμολόγησαν συγκεκριμένα άρθρα, με σκοπό να αναλυθούν οι προσωπικές εκτιμήσεις άρθρων.

Έχοντας πλέον, αποτελέσματα που ήταν βασισμένα στο περιεχόμενο των άρθρων, όσο και στις προσωπικές εκτιμήσεις, τα συνδυάσαμε και δημιουργήσαμε έναν υβριδικό αλγόριθμο. Ο συγκεκριμένος κώδικας κατάφερε τελικά να παρουσιάσει στους χρήστες συγκεκριμένες προτάσεις άρθρων, οι οποίες ήταν βασισμένες στις προσωπικές τους προτιμήσεις, αλλά και στα στοιχεία των άρθρων τα οποία έχουν διαβάσει.

Λέξεις κλειδιά: *μηχανική μάθηση, συστήματα παραγωγής συστάσεων, φιλτράρισμα βάσει περιεχομένου, συνεργατικό φιλτράρισμα, υβριδικό φιλτράρισμα.*

Abstract

Machine learning has made tremendous leaps of progress in recent decades and has now come to be considered an integral part of our daily lives. In fact, machine learning has now exceeded the pure limits of technology and is considered a key cog in many fields. One of these areas is journalism.

More specifically, just as any commercial website wants to offer the best possible service to its customers, today's modern, online, news websites, also want to make their content as attractive as possible, in order to develop a relationship of trust with their users. The ultimate goal is for the users to stay on the site and of course return to them.

To achieve this goal, content personalization is one of the most important tools. Content personalization is now known worldwide, as tech and media giants such as Amazon, Netflix, and Spotify use special algorithms. These algorithms analyze users' preferences and navigation and create personalized content.

The purpose of this paper is to implement the idea of personalizing content on a news website. In order to achieve this, a recommendation system was created, which is based on hybrid filtering. More specifically, after mining more than 1000 articles, their data were analyzed in order to determine the correlations between them. After that, users then rated specific articles in order to analyze their preferences regarding certain news articles.

After concluding on the results that were based on the content of the articles, as well as on personal assessments, we combined them and created a hybrid algorithm. This code finally managed to present users with specific article suggestions, which were based on their personal preferences, but also on the details of the articles they have read.

Περιεχόμενα

Κατάλογος εικόνων

Πρόλογος

Η εξέλιξη της μηχανικής μάθησης έχει κάνει τεράστια άλματα προόδου, ειδικά τα τελευταία χρόνια. Αιτία αυτής της ανοδικής πορείας ήταν η θέληση των ερευνητών να μετατρέψουν το τομέα της ανίχνευσης μοτίβων σε μια μηχανική διαδικασία αναπαραγωγής της ανθρώπινης κρίσης (Roberge, 2020). Για να επιτύχουν τον σκοπό αυτό, δημιούργησαν μηχανές οι οποίες είναι ικανές να μάθουν (learning machines). Αυτές οι μηχανές εκτελούν μια συγκεκριμένη δραστηριότητα σε ένα στενό πλαίσιο, το οποίο ωστόσο μπορεί να έχει σωρεία δεδομένων που εισάγονται. Αυτές οι δραστηριότητες ποίκιλαν, ειδικά στην αρχή της μηχανικής μάθησης, από τον εντοπισμό πυραύλων στον αέρα, μέχρι τον εντοπισμό συγκεκριμένων στοιχειών μέσα από χιλιάδες πληροφορίες.

Πλέον, οι τομείς με τους οποίους ασχολείται η μηχανικοί μάθηση έχουν διευρυνθεί κατά πολύ. Δεν είναι τυχαίο πως σημαντικό ρόλο στην ανάπτυξη και εξέλιξη της μηχανικής έχουν παίξει τον τελευταίο καιρό εταιρείες που ασχολούνται στον τομέα της ψυχαγωγίας. Netflix και Spotify χρησιμοποιούν την μηχανική μάθηση και ειδικούς αλγόριθμους προκειμένου να δημιουργήσουν εξατομικευμένο περιεχόμενο για τους χρήστες τους. Η έννοια αυτή – της εξατομίκευσης περιεχομένου έχει κερδίσει αρκετούς θαυμαστές πλέον. Στόχος της εξατομίκευσης είναι ο κάθε χρήστης να βλέπει διαφορετικό περιεχόμενο ανάλογα με τις προτιμήσεις του όταν περιηγείται σε μια ιστοσελίδα. Με τον τρόπο αυτό, δηλαδή βλέποντας περιεχόμενο που τον ενδιαφέρει, είναι πολύ περισσότερο πιθανό να παραμείνει στην ιστοσελίδα ή και να επιστρέφει σε αυτή τακτικά.

Φυσικά, αυτή είναι μια λειτουργία η οποία μπορεί να φανεί εξαιρετικά χρήσιμη στην δημοσιογραφία. Σκοπός αυτής της εργασίας είναι ακριβώς αυτό. Να εξακριβωθεί κατά πόσο είναι δυνατόν να δημιουργήσουμε εξατομικευμένο περιεχόμενο σε μια ιστοσελίδα ειδησεογραφικού περιεχομένου.

Προκειμένου να το επιτύχουμε την εξατομίκευση, δημιουργήσαμε έναν αλγόριθμο ο οποίος ανάλογα με τα χαρακτηριστικά των άρθρων που διαβάζουν οι χρήστες, αλλά και βασιζόμενος στις προσωπικές τους προτιμήσεις, είναι ικανός να προτείνει σε κάθε χρήστη ξεχωριστά, άρθρα τα οποία θα τον ενδιαφέρουν. Αυτό το καταφέραμε αφού εξορύξαμε περισσότερα από 1000 άρθρα, τα οποία αναλύθηκαν έτσι ώστε να εξάγουμε λέξεις-κλειδιά, μέσω του λεγόμενου φιλτραρίσματος βάσει περιεχομένου (content-based filtering). Στη συνέχεια, ένας μικρός αριθμός χρηστών επιλέχθηκε, προκειμένου να βαθμολογήσει έναν αριθμό άρθρων. Αφού ολοκληρώθηκαν οι βαθμολογίες, χρησιμοποιήθηκε το συνεργατικό φιλτράρισμα, για να παραχθούν προτάσεις άρθρων με βάσεις τις προτιμήσεις των χρηστών.

Αφού, καταλήξαμε σε δύο διαφορετικές συστάσεις άρθρων, μια με βάση το περιεχόμενο των άρθρων και μια με βάση τις προτιμήσεις των χρηστών, τις συνδέσαμε και καταφέραμε να δημιουργήσουμε έναν ολοκληρωμένο υβριδικό αλγόριθμο, ο οποίος πραγματοποιεί συστάσεις άρθρων στους χρήστες.

Η παρουσίαση των αποτελεσμάτων έγινε στην πλατφόρμα Dash. Η συγκεκριμένη πλατφόρμα είναι ιδανική για την οπτικοποίηση δεδομένων, ενώ δίνει και την δυνατότητα αλληλεπίδρασης με τους χρήστες. Το Dash θεωρήθηκε ιδανικό για την συγκεκριμένη εργασία, καθώς είναι ένα πλαίσιο Python, η γλώσσα η οποία χρησιμοποιήθηκε και για την συγγραφή του κώδικα. Επιπλέον, η εφαρμογή που δημιουργήθηκε είναι ορατή σε διαδικτυακούς περιηγητές (web browsers), τόσο σε υπολογιστές όσο και σε κινητές συσκευές.

Κεφάλαιο 1

Η διαδικτυακή δημοσιογραφία

1.1 Η εξέλιξη της διαδικτυακής δημοσιογραφίας

Αρχικά, προκειμένου να κατανοήσουμε καλύτερα την σημασία της διαδικτυακής δημοσιογραφίας, θα ήταν πρέπον να εξετάσουμε πώς ορίζεται. Σύμφωνα με τον Kevin Kawamoto (2003), διαδικτυακή δημοσιογραφία είναι η αξιοποίηση των υπηρεσιών που προσφέρει το διαδίκτυο, για την εξερεύνηση, παραγωγή, και διαθεσιμότητα νέων και πληροφορίων, από ένα – όλο και μεγαλύτερο κοινό – το οποίο κατέχει γνώσεις χειρισμού ηλεκτρονικών υπολογιστών.

Έχοντας τον παραπάνω ορισμό στον νου μας, και αναλογιζόμενοι την τεχνολογική ‘’έκρηξη’’ που λαμβάνει μέρος τα τελευταία χρόνια παγκοσμίως - η οποία επηρεάζει όλους τους τομείς της ζωής μας – είναι επόμενο να συντελούνται αλλαγές και στην δημοσιογραφία. Πιο συγκεκριμένα, η είσοδος τους διαδικτύου στην καθημερινότητα μας, έχει αλλάξει τον τρόπο με τον οποίο θέλουμε να ενημερωνόμαστε, επηρεάζοντας με την σειρά του τα Μέσα Μαζικής Ενημέρωσης ανά τον κόσμο.

Πλέον, το αναγνωστικό κοινό λαμβάνει την ενημέρωση του κυρίως μέσω διαδικτύου, και όχι μέσω εφημερίδων, όπως συνέβαινε παλιότερα. Αυτό έχει αναγκάσει τα παγκόσμια μέσα να εκσυγχρονιστούν και να δημοσιεύουν τα άρθρα τους στο διαδίκτυο, ούτως ώστε να ανταποκριθούν στις νέες ανάγκες που έχουν δημιουργηθεί.

Πλέον, ο τρόπος με τον οποίο παρουσιάζεται το ενημερωτικό περιεχόμενο είναι κατά κόρον διαδικτυακός. Τηλεοπτικά κανάλια, ραδιοφωνικοί σταθμοί, ακόμα και εφημερίδες, έχουν εστιάσει πλέον στην ψηφιακή μορφή τους, όντας διαδικτυακά μέσα. Η αλλαγή αυτή και η επιλογή του διαδικτύου ως κύριο μέσο αναπαραγωγής ειδήσεων, έχει δώσει παράλληλα την δυνατότητα στις ενημερωτικές ιστοσελίδες να εμπλουτίσουν τον τρόπο λειτουργίας τους, έτσι ώστε να διατηρήσουν και να κερδίσουν παραπάνω αναγνωστικό κοινό.

Πλέον, κάθε ενημερωτική ιστοσελίδα έχει την δυνατότητα να σκιαγραφήσει το κοινό της, γνωρίζοντας καλύτερα τις προτιμήσεις του, έτσι ώστε να προσφέρουν όσο το δυνατόν πιο ελκυστικό περιεχόμενο. Ενδεικτικά, μέσω συγκεκριμένων αλγορίθμων, δίνεται πλέον η δυνατότητα μια ιστοσελίδα να γνωρίζει ακριβώς τις προτιμήσεις των χρηστών της, χρησιμοποιώντας είτε εκπεφρασμένες προτιμήσεις των χρηστών, είτε αναλύοντας τα χαρακτηριστικά των άρθρων που διαβάζουν.

Με αυτόν τον τρόπο, μπορούν να προσαρμόσουν το περιεχόμενό τους στις προτιμήσεις των χρηστών, προσφέροντάς τους ακριβώς τα νέα που επιθυμούν να διαβάσουν. Φυσικά, η συγκεκριμένη διαδικασία είναι δυναμική, καθώς αν αλλάξουν οι προτιμήσεις των χρηστών, η ιστοσελίδα προσαρμόζεται και παρουσιάζει τα νέα άρθρα που αντιστοιχούν στις καινούργιες προτιμήσεις.

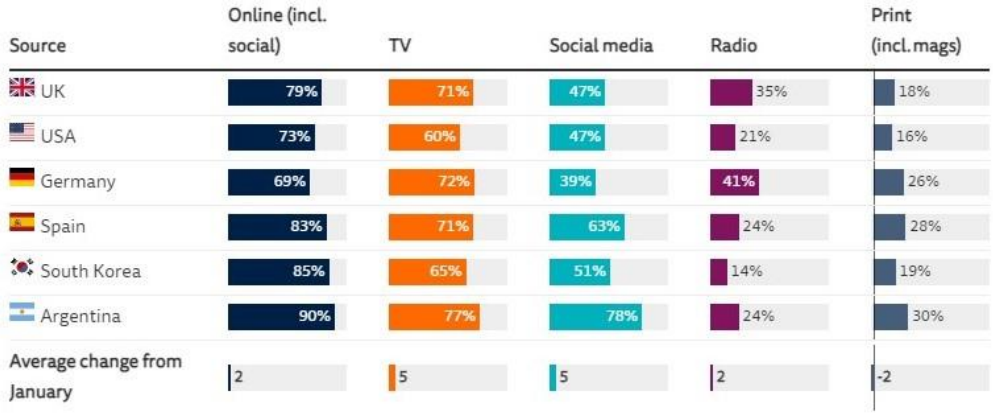
Τα συστήματα παραγωγής συστάσεων, όπως ονομάζονται οι αλγόριθμοι που είναι υπεύθυνοι για παραγωγή συστάσεων άρθρων με βάση τις προτιμήσεις των χρηστών, έχουν πλέον αλλάξει τον χαρακτήρα των ενημερωτικών ιστοσελίδων. Καθώς κάθε χρήστης μπορεί να είναι σίγουρος πως θα δει ακριβώς τα άρθρα που τον ενδιαφέρουν, αυξάνεται η εμπιστοσύνη με την ιστοσελίδα, προσφέροντας της έτσι ένα πλεονέκτημα έναντι άλλων

που δεν προσφέρουν την συγκεκριμένη δυνατότητα. Έτσι, χωρίς να αλλάζει το περιεχόμενο των άρθρων, παρά μόνο το ποια άρθρα συστήνονται σε συγκεκριμένους χρήστες, οι ενημερωτικές ιστοσελίδες μπορούν να προσφέρουν εύκολα, στοχευμένο περιεχόμενο για το κοινό τους. Ωστόσο, τα συστήματα παραγωγής συστάσεων θα αναλυθούν διεξοδικά στην συνέχεια.

Για να βρούμε την απαρχή ωστόσο της διαδικτυακής δημοσιογραφίας, πρέπει να γυρίσουμε πίσω στο 1993. Τότε, το τμήμα Δημοσιογραφίας του Πανεπιστημίου της Φλόριντα, δημοσίευσε – αυτό που θεωρείται – τον πρώτο διαδικτυακό δημοσιογραφικό ιστότοπο (Siapera, Veglis, 2012). Επρόκειτο για έναν ιστότοπο σε πρώιμο στάδιο βεβαίως, και αρκετά στατικό, ο οποίος ενημερωνόταν μόνο τα βράδια, όταν και δεν τον χρησιμοποιούσε κάποιος.

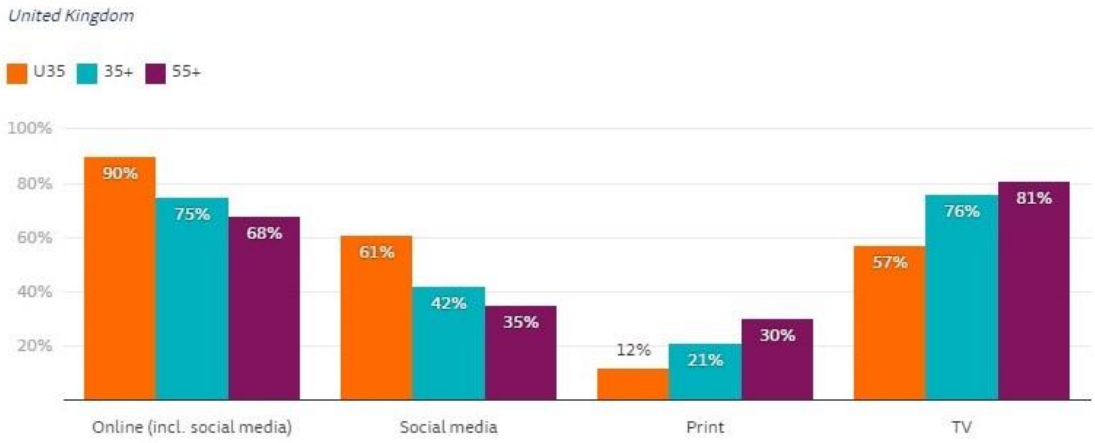
Από τότε, ο αρχικός ενδοιασμός των Μέσων Μαζικής Ενημέρωσης να ακολουθήσουν τον διαδικτυακό τρόπο ενημέρωσης έχει καταρρεύσει πλήρως, καθώς πλέον όλα τα δημοσιογραφικά μέσα έχουν διαδικτυακή παρουσία, στην οποία και δίνουν περισσότερη έμφαση από ότι στις έντυπες μορφές τους.

Την μεγάλη σημασία που έχει αποκτήσει πλέον η διαδικτυακή δημοσιογραφία, αναδεικνύει και μια πρόσφατη έρευνα που πραγματοποίησε το πρακτορείο Reuters. Σύμφωνα με την συγκεκριμένη έρευνα η πανδημία του COVID-19 επιτάχυνε την ψηφιακή αλλαγή στα μέσα μαζικής ενημέρωσης. Όπως φαίνεται και από το παρακάτω διάγραμμα, η πλειοψηφία των ανθρώπων επιλέγει τα διαδικτυακά μέσα ενημέρωσης – είτε social media, είτε ιστοτόπους – για την ενημέρωση τους:



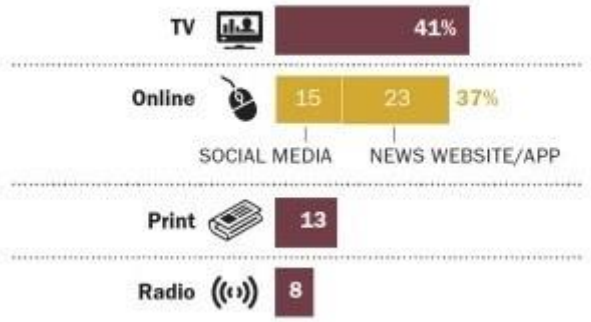
Εικόνα 1: Χρήση μέσων ως πηγή πληροφοριών (Reuters Institute, Digital News Report 2020)

Επιπλέον, η ίδια έρευνα καταλήγει πως άνθρωποι όλων των ηλικιών επιλέγουν πλέον τα διαδικτυακά μέσα για την ενημέρωση τους, δείχνοντας την τεράστια σύνδεση που έχουν πλέον με την καθημερινότητα μας. Ενδεικτικά, σύμφωνα με την έρευνα, στον Ηνωμένο Βασίλειο, τα άτομα κάτω των 35 ετών επιλέγουν τα διαδικτυακά μέσα σε ποσοστό 91%. Ακολουθούν οι άνω των 35 ετών με 75%, ενώ σε ένα αρκετά σημαντικό ποσοστό, τα άτομα άνω των 55 επιλέγουν τα διαδικτυακά μέσα σε ένα ποσοστό της τάξεως του 68%, όπως θα φανεί και από το διάγραμμα που ακολουθεί:



Εικόνα 2:Χρήση μέσων ως πηγή πληροφοριών ανά ηλικιακή ομάδα (Reuters Institute, Digital News Report 2020)

Επιπροσθέτων, και το ερευνητικό κέντρο ''Pew'' έχει υπογραμμίσει την ολοένα και αυξανόμενη άνοδο την διαδικτυακών μέσων ενημέρωσης. Πιο συγκεκριμένα, μετά από έρευνα που πραγματοποιήθηκε το 2018, το ''Pew'' συμπέρανε πως στην Αμερική περίπου 4 στους 10 ενήλικες (37%) προτιμούν να λαμβάνουν την ενημέρωση τους διαδικτυακώς, όπως θα δούμε και στο επόμενο διάγραμμα. Επιπλέον, οι συμμετέχοντας της ίδιας έρευνας απάντησαν σε ποσοστό 77%, πως το διαδίκτυο παίζει σημαντικό ρόλο στον τρόπο πρόσληψης νέων.



Εικόνα 3:Χρήση μέσων ως πηγή πληροφοριών από ενήλικες Αμερικάνους (PEW Research Center, For Local News, Americans Embrace Digital but Still Want Strong Community Connection, 2018)

1.2 Σύγκλιση μέσων

Τα τελευταία χρόνια παρατηρείται πως τα μέσα διατελούν μια διαδικασία συγχώνευσης. Πιο συγκεκριμένα, πλέον υπάρχει η δυνατότητα πολλά διαφορετικά μεταξύ τους μέσα, να προβάλλονται από την ίδια ψηφιακή πλατφόρμα. Χαρακτηριστικά, μέσα όπως το ραδιόφωνο, το διαδίκτυο και ο έντυπος τύπος, κάποτε αποτελούσαν ξεχωριστά μέσα, τα οποία έβρισκαν βήμα μέσα από καλά διαχωρισμένες πλατφόρμες. Παρόλα αυτά, η αυξανόμενη ψηφιοποίηση έχει δώσει την δυνατότητα όλα αυτά τα μέσα να ενωθούν, κάτω από μία ''ομπρέλα''.

Ωστόσο, αυτή η σύγκλιση των μέσων έχει δημιουργήσει πλέον ένα νέο είδος: τα ''έξυπνα δίκτυα'', τα οποία χαρακτηρίζονται από σύγκλιση τεχνολογιών (Dal Zotto, Lugmayr, 2015). Στόχος των συγκεκριμένων δικτύων είναι να βελτιώσουν την απόδοσή τους μέσω της σωστής οργάνωσης και διαμερισμού της πληροφορίας. Τον συγκεκριμένο σκοπό τον επιτυγχάνουν μέσα από τρία στάδια:

1. Παρουσίαση.
2. Λήψη αποφάσεων.
3. Τεχνητή νοημοσύνη (Artificial Intelligence).

Αρχικά, μέσω της παρουσίασης, το δίκτυο επιθυμεί να εξασφαλίσει τον επιτυχή διαμοιρασμό της πληροφορίας – όποιου είδους και αν είναι αυτή - στο κοινό το οποίο επιθυμεί.

Στη συνέχεια, στο δεύτερο στάδιο – της λήψης αποφάσεων – το δίκτυο έχει ως στόχο να δώσει στον χρήση όλες τις απαραίτητες πληροφορίες, έτσι ώστε ο τελευταίος να καταλήξει σε μια συνειδητή απόφαση. Απώτερος σκοπός είναι ο χρήστης να λάβει πληθώρα πληροφοριών, ούτως ώστε να καταλάβει εις βάθος το θέμα που τον απασχολεί.

Το τρίτο στάδιο είναι και αυτό με το οποίο θα ασχοληθεί η συγκεκριμένη εργασία εκτενώς: η τεχνητή νοημοσύνη (AI). Τα τελευταία χρόνια, το AI έχει εκτιναχθεί. Χαρακτηριστικά, συμβουλευτική εταιρεία PricewaterhouseCoopers εκτιμά πως η τεχνητή νοημοσύνη θα συμβάλλει 15,7 τρισεκατομμύρια δολάρια στην παγκόσμια οικονομία μέχρι το 2030 (Marconi, 2020).

Τι ακριβώς όμως κάνει το AI; Μέσω της τεχνητής νοημοσύνης, μηχανές αποκτούν την δυνατότητα να πραγματοποιούν ‘’ανθρώπινες’’ πράξεις, μαθαίνοντας από την εμπειρία. Όσο περισσότερα δεδομένα αποκτούν, τόσο πιο αποτελεσματικές γίνονται. Αυτή η δυνατότητα μπορεί να φανεί εξαιρετικά χρήσιμη για τα ψηφιακά μέσα.

Πιο συγκεκριμένα, χρησιμοποιώντας συγκεκριμένους αλγόριθμους, η τεχνητή νοημοσύνη μπορεί να προσφέρει στους χρήστες εξατομικευμένες συστάσεις, βασισμένες αποκλειστικά και μόνο στις επιλογές του ίδιου του χρήστη. Είναι ένα εργαλείο το οποίο το χρησιμοποιούν πολλά χρόνια εταιρείες-κολοσσοί, όπως είναι η Amazon, η Google, το Netflix και το Spotify.

1.3 Ευχρηστία διαδικτυακών Μέσων Μαζικής Ενημέρωσης

Με την εξέλιξη της τεχνολογίας και της δημιουργίας πλέον πηγών ενημέρωσης που βασίζονται αποκλειστικά στο διαδίκτυο, γεννάται το ερώτημα της βελτίωσης. Πώς μπορεί να βελτιωθεί ένας ιστοχώρος, έτσι ώστε να προσφέρει περιεχόμενο που θα ελκύει τον χρήστη, και επομένως θα ενισχύσει την διατήρηση χρηστών. Πυρήνας της λύσης του συγκεκριμένου προβλήματος, είναι η δημιουργία του ιστοχώρου χρηστοκεντρικά, δηλαδή έχοντας ως προτεραιότητα τις επιθυμίες των χρηστών.

Η ιδέα πίσω από την χρηστοκεντρική προσέγγιση για την δημιουργία μιας ιστοσελίδας είναι η δημιουργία ενός διαλόγου μεταξύ των χρηστών και του ιστοτόπου, έτσι ώστε το τελευταίο να ικανοποιήσει τις επιθυμίες των πρώτων (McKay, 2013). Μια καλά σχεδιασμένη ιστοσελίδα γίνεται κατανοητή από τους χρήστες, καθώς η λειτουργία της γίνεται εύκολα, είναι φιλική προς τον χρήστη, όπως επίσης και αποτελεσματική, οι χρήστες δηλαδή καταφέρνουν και βρίσκουν τελικά αυτό που ψάχνουν.

Ο διάλογος αυτός μεταξύ χρηστών και ιστοσελίδας είναι εξαιρετικά χρήσιμος, καθώς ενισχύει την επικοινωνία μεταξύ των δύο πλευρών, ενισχύοντας έτσι την εμπιστοσύνη των χρηστών πως θα βρουν ακριβώς αυτό που ψάχνουν στην ιστοσελίδα, και μάλιστα χωρίς να καταβάλλουν ιδιαίτερη προσπάθεια. Το τελευταίο στοιχείο εκτελεί και μια ακόμα λειτουργία. Σε έναν κόσμο ο οποίος λειτουργεί με εξαιρετικά γρήγορους ρυθμούς, οι χρήστες επιθυμούν να λαμβάνουν την πληροφορία που επιθυμούν χωρίς να χρονοτριβούν. Για τον παραπάνω λόγο, οι ενημερωτικές ιστοσελίδες οφείλουν να βρίσκουν τρόπους έτσι

ώστε να διατηρούν τον χρήστη και παράλληλα να του παρέχουν όποια πληροφορία θέλει εγκαίρως.

Ένας ακόμα λόγος για τον οποίο οι ενημερωτικές ιστοσελίδες οφείλουν να επενδύσουν στην χρηστοκεντρική προσέγγιση, είναι ο ολοένα και αυξανόμενος αριθμός χρηστών που πλέον μαθαίνουν τις ειδήσεις μέσω του διαδικτύου. Ο συγκεκριμένος τρόπος ενημέρωσης διαφέρει σε αρκετά και σημαντικά σημεία σε σχέση με παραδοσιακούς τρόπους ενημέρωσης, όπως η τηλεόραση, το ραδιόφωνο, και οι εφημερίδες. Το πιο σημαντικό στοιχείο διαφοράς, είναι πως πλέον οι χρήστες έχουν πολλές επιλογές χώρων από τους οποίους μπορούν να ενημερωθούν, και στους οποίους θα βρουν ακριβώς την πληροφορία που ψάχνουν, την στιγμή που επιθυμούν.

Ωστόσο, μια σημαντική ειδοποιός διαφορά είναι πως οι ιστοσελίδες μπορούν πλέον και προσαρμόζονται στις προτιμήσεις του κάθε χρήστη ατομικά. Αυτό είναι ένα στοιχείο το οποίο δεν προσφέρουν τα παραδοσιακά μέσα ενημέρωσης, αλλά και το οποίο μπορεί να κάνει την διαφορά στην επιτυχία μιας ενημερωτικής ιστοσελίδας σε σχέση με μία άλλη. Έτσι, οι χρήστες μπορούν να βλέπουν ειδήσεις που αφορούν αποκλειστικά και μόνο τα δικά τους ενδιαφέρονται, χωρίς να έχουν τον ‘’θόρυβο’’ άλλων νέων τα οποία δεν τους ενδιαφέρουν.

Ένα χαρακτηριστικό παράδειγμα είναι τα Google News. Ψάχνοντας νέα μέσω της Google, η πλατφόρμα αναζήτησης αναλύει αυτόματα τα μοτίβα τα οποία ψάχνουμε, προτείνοντας μας έπειτα παρόμοια νέα. Μια παρόμοια λειτουργία, θα μπορούσε να φανεί εξαιρετικά χρήσιμο στις ενημερωτικές ιστοσελίδες.

Ωστόσο, τα νέα μέσα που προσφέρει η τεχνολογία, καθιστούν την δημιουργία μιας εύχρηστης ενημερωτικής ιστοσελίδας πιο εύκολη, αλλά και αναγκαία. Πιο συγκεκριμένα, όπως αναφέρθηκε και παραπάνω, αναλύοντας τα μοτίβα και τις προτιμήσεις των χρηστών, και χρησιμοποιώντας ειδικό κώδικα, μια ενημερωτική ιστοσελίδα θα ήταν ικανή να μελετήσει τις προτιμήσεις κάθε χρήστη ξεχωριστά. Με τον τρόπο αυτό θα επιτύχει δύο πολύ σημαντικούς στόχους. Αρχικά, θα διαφοροποιηθεί από τον ανταγωνισμό. Ακόμα και αν η θεματολογία των ειδήσεων είναι παρόμοια με άλλες ιστοσελίδες, το γεγονός πως θα παρουσιάζονται στον χρήστη μόνο τα άρθρα που τον ενδιαφέρουν, θα διαφοροποιήσει σημαντικά την συγκεκριμένη ιστοσελίδα. Δεύτερο, ως αποτέλεσμα αυτής της λειτουργίας – δηλαδή της παρουσίασης εξατομικευμένου περιεχομένου, λύνονται δύο ακόμα προβλήματα. Το θέμα του περιεχομένου όταν υπάρχει ένας μεγάλος αριθμός χρηστών, και δεύτερον η διατήρησή τους.

Επιπλέον, η ευχρηστία μιας ενημερωτικής ιστοσελίδας έγκειται και στο πόσο δυναμική είναι. Οι προτιμήσεις των ανθρώπων αλλάζουν από μέρα σε μέρα, και αυτό είναι ένα γεγονός το οποίο πρέπει να ληφθεί σοβαρά υπόψη. Για τον λόγο αυτό, οι αλγόριθμοι που προβάλλουν εξατομικευμένες συστάσεις άρθρων στους χρήστες οφείλουν να είναι δυναμικοί, δηλαδή να αναγνωρίζουν τις αλλαγές των προτιμήσεων ενός χρήστη, και να τους προβάλλουν άρθρα με βάση τις νέες προτιμήσεις του. Αυτή η προσαρμοστικότητα ενισχύει την ευχρηστία της ιστοσελίδας, καθώς οι χρήστες θα διαβάζουν συνεχώς περιεχόμενο που τους ενδιαφέρει, χωρίς ωστόσο να έχουν καταβάλλει προσπάθεια να εξηγήσουν την αλλαγή στις προτιμήσεις τους.

Σχετικά με τον μεγάλο αριθμό χρηστών, αυτό δεν αποτελεί πρόβλημα για μια ιστοσελίδα που παρουσιάζει εξατομικευμένο περιεχόμενο και αυτό γιατί το τι βλέπει κάθε χρήστης

προσαρμόζεται στις δικές τους προτιμήσεις. Επιπλέον, αυτό γίνεται με μια αυτοματοποιημένη διαδικασία, η οποία μαθαίνει καλύτερα τον χρήστη όσο εκείνος περιηγείται στην ιστοσελίδα. Όσο καλύτερα μαθαίνει ο αλγόριθμος τον χρήστη, τόσο πιο προσαρμοστικό θα είναι το περιεχόμενο στις προτιμήσεις του χρήστη. Κατά συνέπεια, όσο παρουσιάζεται στον χρήστη ένα περιεχόμενο το οποίο ταιριάζει στις προτιμήσεις του, τόσο πιο πιθανό είναι αυτός ο χρήστης να παραμείνει στην ιστοσελίδα, ή και να την επισκέπτεται τακτικά.

Αναδεικνύοντας την μεγάλη σημασία της εξατομίκευσης, μια έρευνα που διεξήχθη το 2017 από την εταιρεία Epsilon κατέληξε στο συμπέρασμα πως για το 80% των καταναλωτών είναι περισσότερο πιθανό να προχωρήσουν σε αγορά αν η ιστοσελίδα τους παρέχει μια προσωποποιημένη εμπειρία. Η ίδια έρευνα αναφέρει επίσης, πως το 90% των ερωτηθέντων αξιολογεί το εξατομικευμένο περιεχόμενο θετικά.

Αυτά είναι τα πλεονεκτήματα της εξατομίκευσης του περιεχομένου, η οποία θα αναλυθεί διεξοδικά στο επόμενο κεφάλαιο.

Συμπεράσματα

Στο συγκεκριμένο κεφάλαιο δόθηκε έμφαση στον μεγάλο ρόλο που διαδραματίζουν πλέον τα ψηφιακά μέσα για την ενημέρωση παγκοσμίως. Όπως αναφέρουν μεγάλες έρευνες, ένα πολύ μεγάλο ποσοστό της ανθρωπότητας πλέον βασίζεται αποκλειστικά και μόνο στα ψηφιακά μέσα για την ενημέρωση του, είτε αυτά είναι ιστότοποι, είτε τα μέσα κοινωνικής δικτύωσης.

Ωστόσο, υπάρχει μια τεχνολογία η οποία μπορεί να εκτινάξει την αποτελεσματικότητα των ψηφιακών μέσων. Και αυτή είναι η τεχνητή νοημοσύνη. Στο επόμενο κεφάλαιο θα εμβαθύνουμε στο πώς το AI καταφέρνει και παράγει εξατομικευμένες συστάσεις για τους χρήστες, αλλά επίσης και τους διαφορετικούς τρόπους με τους οποίους το καταφέρνει.

Επιπλέον, σε αυτό το κεφάλαιο έγινε μια πρώτη εισαγωγή για την έννοια της εξατομίκευσης περιεχομένου. Η εξατομίκευση – μια έννοια η οποία διέπει ολόκληρη την εργασία – στοχεύει στην παρουσίαση προσωποποιημένου περιεχομένου για τους χρήστες μιας ενημερωτικής ιστοσελίδας. Πρόκειται για ένα πολύ σημαντικό στοιχείο για την επιτυχία ενός οποιοδήποτε ιστοχώρου, καθώς προσφέροντας περιεχόμενο το οποίο προσαρμόζεται στους χρήστες του, ο ιστότοπος χτίζει μια σχέση εμπιστοσύνης με το κοινό του, και άρα αποκτά προβάδισμα σε σχέση με τους ανταγωνιστές του.

Τέλος, αναλύσαμε την χρηστοκεντρική προσέγγιση για την δημιουργία ιστοσελίδων και την παραγωγή περιεχομένου. Με την συγκεκριμένη προσέγγιση, οι προτιμήσεις του χρήστη μπαίνουν σε προτεραιότητα, έτσι ώστε οι ενημερωτικές ιστοσελίδες να γνωρίζει τις ανάγκες του κοινού τους και να τους προσφέρουν περιεχόμενο προσαρμοσμένο στα γούστα τους. Η εξατομίκευση περιεχομένου αξιολογείται θετικά και από τους χρήστες καθώς η πλειοψηφία αξιολογεί την συγκεκριμένη λειτουργία ως αναγκαία.

Κεφάλαιο 2

Εξατομίκευση περιεχομένου

2.1 Εξατομικευμένο περιεχόμενο

Τα τελευταία χρόνια παρατηρούμε πως συντελείται μια επανάσταση της πληροφορίας, και μάλιστα με πολύ μεγάλη επιτυχία. Πλέον η καθημερινότητα μας έχει αλλάξει δραστικά ως προς τον τρόπο που λαμβάνουμε πληροφορίες. Σε αυτό έχει συμβάλει πολύ η εξέλιξη της τεχνολογίας, καθώς όλοι πλέον μπορούν ανά πάσα στιγμή να μαθαίνουν οτιδήποτε, απλά και μόνο με μια αναζήτηση στη Google από το κινητό τους.

Παρόλη την τεράστια και σημαντική εξέλιξη της πληροφορίας ωστόσο, υπάρχει και ένα αρνητικό. Λόγω της τεράστιας αφθονίας πληροφοριών και πηγών που υπάρχουν πλέον, οι χρήστες δυσκολεύονται ολοένα και περισσότερο να βρουν την σωστή πληροφορία στη σωστή στιγμή. Το γεγονός αυτό παίρνει μεγαλύτερες διαστάσεις αν αναλογιστεί κανείς και την μεγάλη ποικιλία μηχανών από τις οποίες μπορούμε να αντλήσουμε πληροφορίες. Για παράδειγμα, πλέον υπάρχουν τα κινητά τα οποία ανά πάσα στιγμή μπορούν να συνδεθούν στο διαδίκτυο, οι υπολογιστές, ακόμα και έξυπνες τηλεοράσεις οι οποίες προσφέρουν σύνδεση στο ίντερνετ.

Επιπλέον, ο μέσος χρήστης σήμερα δεν έχει την ικανότητα να βρει την πληροφορία που τον ενδιαφέρει. Πιο συγκεκριμένα, οι χρήστες σήμερα δεν είναι ειδικοί στο να ανακαλύπτουν πληροφορίες, και ο τρόπος με τον οποίο αναζητούν πληροφορίες, δεν βοηθά τις μηχανές αναζήτησης να ικανοποιήσουν τις ανάγκες του (Ma, Uchyigit, 2008).

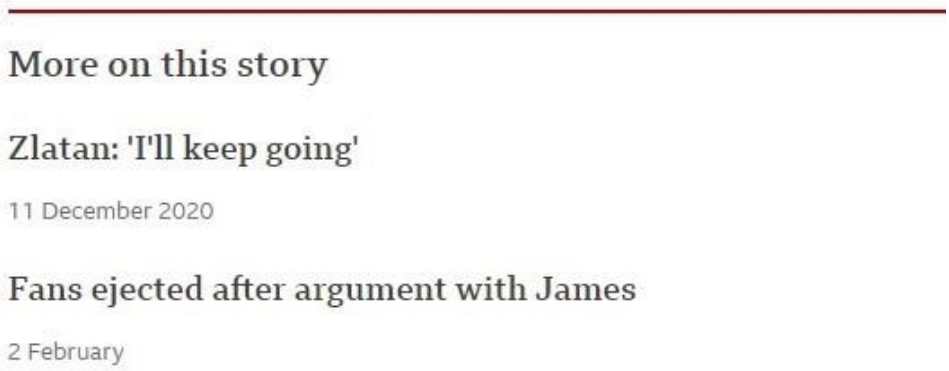
Προκειμένου να λυθεί το συγκεκριμένο πρόβλημα, οι ερευνητές πλέον έχουν στραφεί στην σημασία των προσωποποιημένων υπηρεσιών πληροφορίας (personalized information services). Η προσωποποίηση της πληροφορίας συνδυάζει τεχνικές ανάλυσης του προφίλ του χρήστη, ανακάλυψης πληροφοριών, και τεχνητής νοημοσύνης, προκειμένου να δημιουργήσει υπηρεσίες πληροφορίας οι οποίες θα ανταποκρίνονται πιο αποτελεσματικά στις ανάγκες του κάθε χρήστη ξεχωριστά.

Σε μια προσπάθεια να προσδώσει έναν ορισμό στην εξατομίκευση, ο Kim τονίζει πως η εξατομίκευση είναι η παροχή πληροφορίας σε ένα άτομο ή μια ομάδα ατόμων, αλλά με έναν τρόπο ο οποίος είναι ήδη προσδιορισμένος, όπως επίσης και σε ήδη προσδιορισμένο χρόνο. Σε περίπτωση που οι πηγές της πληροφορίας ανανεωθούν, τότε είναι πολύ σημαντικό να ανανεωθεί και η ίδια η πληροφορία που παρέχεται. Η ανανεωμένη πληροφορία μπορεί να διαμοιραστεί μετά την ανανέωση της πηγής, ή μετά από αίτημα των ίδιων των χρηστών (Kim, 2002).

Ωστόσο, ο ίδιος προσθέτει πως η εξατομίκευση μπορεί να οριστεί και με έναν δεύτερο τρόπο. Ο δεύτερος ορισμός εξετάζει την εξατομίκευση από την πλευρά του one-to-one marketing. Πιο συγκεκριμένα, ο Kim τονίζει πως σκοπός είναι η αύξηση του εισοδήματος και η μείωση της χρηματικής απώλειας μιας εταιρείας, μέσω της καλύτερης κατανόησης – και τελικά ικανοποίησης - των αναγκών, τρόπου ζωής, και προτιμήσεων των πελατών. Η ιδέα είναι πως μέσω της κατανόησης των αναγκών των πελατών, η εταιρεία τελικά θα μπορέσει να προσφέρει καλύτερα και πιο ελκυστικά προϊόντα στο κοινό της.

Επιπλέον, το εξατομικευμένο περιεχόμενο έχει ως απώτερο σκοπό την δυναμική προσαρμοστικότητα στις ανάγκες των χρηστών μιας ιστοσελίδας, έτσι ώστε να είναι σε θέση να προσφέρει το καλύτερο δυνατό περιεχόμενο στο αναγνωστικό κοινό. Με αυτόν τον τρόπο, κάθε χρήστης θα βλέπει αυτόματα το περιεχόμενο που τον ενδιαφέρει, όσο και αν αλλάξουν οι προτιμήσεις του, καθώς η ιστοσελίδα θα μπορεί να προσαρμόζεται συνεχώς.

Ένα χαρακτηριστικό παράδειγμα εξατομίκευσης περιεχομένου είναι η ιστοσελίδα του BBC. Πιο συγκεκριμένα, αν ένας χρήστης διαβάσει κάποια άρθρα από την κατηγορία των αθλητικών, το BBC θα προτείνει άρθρα αθλητικού περιεχομένου, που είναι όμοια με τα άρθρα που διαβάζει ο χρήστης. Για παράδειγμα, διαβάζοντας το άρθρο, ‘’LeBron vs Zlatan: Who won the politics bout?’’, το BBC προτείναι ακόμα δυο άρθρα, παρόμοιου περιεχομένου:



Εικόνα 4:Παραγωγή συστάσεων άρθρων στο BBC.

Στην ίδια λογική κινείται και το CNN, καθώς όπως και το BBC, προτείνει άρθρα στον χρήστη με βάση ήδη διαβασμένα άρθρα. Έτσι, αν διαβάσουμε το άρθρο ‘’ Trump supporters who breached the Capitol: 'It was not Antifa’’, θα έχουμε τις εξής προτάσεις:

Εικόνα 5: Παραγωγή συστάσεων άρθρων στο CNN.

Γενικά, η παραγωγή των συστάσεων γίνεται με τρεις τρόπους. Με βάση το περιεχόμενο των άρθρων, συνεργατικά, με βάση δηλαδή τι έχουν διαβάσει παρόμοιοι χρήστες, και υβριδικά, συνδυάζοντας τους δύο πρώτους τρόπους. Όλοι οι τρόποι παραγωγής συστάσεων, θα αναλυθούν διεξοδικά παρακάτω.

Συνοπτικά, το εξατομικευμένο περιεχόμενο προσφέρει μια μοναδική εμπειρία στον κάθε χρήστη, η οποία δεν μπορεί να αναπαραχθεί σε άλλους χρήστες, καθώς έχουν άλλες προτιμήσεις.

2.1.1 Λειτουργίες εξατομίκευσης

Η τεράστια αφθονία που παρατηρείται στις μέρες μας, έχει δημιουργήσει την ανάγκη για υπηρεσίες πληροφοριών, οι οποίες θα έχουν ως προτεραιότητα την ατομικότητα όσων ψάχνουν για πληροφορίες, προκειμένου να προσωποποιήσουν τα αποτελέσματα από αυτή την αναζήτηση (Fink et al, 2020).

Όπως αναφέρουν και οι Μουρλάς, Γερμανάκος, η εξατομίκευση είναι η προσαρμογή γενικών πληροφοριών πάνω στα χαρακτηριστικά ενός χρήστη (Mourlas, Germanakos, 2009). Αυτά μπορεί να είναι δημογραφικά, ικανότητες, ενδιαφέροντα, και στόχοι. Πιο συγκεκριμένα, οι λειτουργίες της εξατομίκευσης έχουν τους παρακάτω τρεις σκοπούς:

1. Μείωση γνωστικού στρες. Πολλοί χρήστες έχουν δηλώσει πως βιώνουν δυσκολία στο να βρουν την πληροφορία που θέλουν λόγω της μεγάλης αφθονίας.
2. Βελτίωση του τρόπου πρόσληψης της πληροφορίας από τον χρήστη.
3. Εδραίωση αμοιβαίας εμπιστοσύνης μεταξύ του χρήστη και της υπηρεσίας που παρέχει την πληροφορία.

Επιπλέον, πέρα από τους στόχους που έχει η εξατομίκευση, οι λειτουργίες μπορούν να χωριστούν σε τρεις υποκατηγορίες.

1. Προσαρμοστικότητα περιεχομένου: Σε αυτή την κατηγορία το περιεχόμενο πρέπει να διαμορφωθεί σύμφωνα με τις προτιμήσεις του χρήστη και τις δυνατότητες του συστήματος. Επίσης σε αυτήν την κατηγορία, οι εξατομικευμένες υπηρεσίες έχουν ως στόχο να προσφέρουν στους χρήστες το επιθυμητό αποτέλεσμα, χωρίς όμως εκείνοι να το ζητήσουν. Για να πετύχει τον σκοπό της η προσαρμοστικότητα περιεχομένου, πρώτα οφείλει να ξεπεράσει συγκεκριμένες δυσκολίες, οι οποίες αφορούν στα εξής:

- Τι περιεχόμενο να παρουσιαστεί στον χρήστη.
- Πώς να δείξει το περιεχόμενο στον χρήστη.
- Πώς να προστατευτεί η ιδιωτικότητα του χρήστη.
- Πώς να δημιουργήσει ένα παγκόσμιο σύστημα εξατομίκευσης.

Σε πιο εξειδικευμένες περιπτώσεις, για να επιτευχθεί η εξατομίκευση ο χρήστης θα χρειαστεί να εγγραφεί και να δηλώσεις τις προτιμήσεις του, ή συγκεκριμένες υπηρεσίες θα δημιουργήσουν το προφίλ του χρήστη, με βάση την πλοήγηση του στο διαδίκτυο.

2. Προσαρμοστικότητα συστήματος: Η συγκεκριμένη λειτουργία δεν ασχολείται με το περιεχόμενο, όσο με τα ίδια τα συστήματα. Στόχος της είναι τα συστήματα να δουλεύουν απρόσκοπτα, έτσι ώστε να παράγεται με γρήγορες ταχύτητες η εξατομίκευση. Απώτερος σκοπός βέβαια, είναι τόσο η ταχύτητα όσο και η αποτελεσματικότητα της πρόβλεψης να βρίσκονται στα ανώτατα επίπεδα.

3. Προσαρμοστικότητα δικτύων και επικοινωνίας: Τα τελευταία χρόνια εμφανίζονται όλο και περισσότερα νέα δίκτυα που προσφέρουν μεγάλες ποσότητες πληροφορίας. Live-streaming, διαδικτυακές συναντήσεις, και υπηρεσίες on-demand, είναι νέες υπηρεσίες οι οποίες πρέπει να προσαρμοστούν στις ανάγκες των χρηστών τους, αλλά και να έχουν εγγυήσεις για την ποιότητα των υπηρεσιών που προσφέρουν (Quality of Service). Αυτά τα δύο χαρακτηριστικά (προσαρμοστικότητα και Quality of Service) είναι αναγκαία για να τα εγκρίνουν οι χρήστες, καθώς τα συγκρίνουν με ήδη δοκιμασμένα μέσα όπως είναι η τηλεόραση. Επιπλέον, ένα ακόμα ζήτημα που πρέπει να ξεπεραστεί είναι η μεγάλη κινητικότητα των χρηστών. Η αλήθεια είναι πως πλέον πολλοί χρήστες ψάχνουν για

πληροφορίες ενώ βρίσκονται στον δρόμο. Επομένως, τα νέα μέσα πρέπει να λάβουν υπόψιν τους παράγοντες όπως το εύρος ζώνης (bandwith) των κινητών, και την ποιότητα του διαδικτύου των χρηστών, προκειμένου να τους προσφέρουν το καλύτερο δυνατό αποτέλεσμα.

2.1.2 Στόχος εξατομίκευσης περιεχομένου

Όπως αναφέρθηκε και νωρίτερα στο ίδιο κεφάλαιο αλλά και στο προηγούμενο, η εξατομίκευση έχει ως κέντρο τους χρήστες. Στόχος της είναι να τους παρουσιάσει ένα τέτοιο περιβάλλον το οποίο κάθε χρήστης θα βρίσκει ελκυστικό, όντας προσαρμοσμένο στις προτιμήσεις του, αυξάνοντας έτσι την πιθανότητα διατήρησης του εκάστοτε χρήστη.

Προκειμένου ωστόσο να πραγματοποιηθεί με επιτυχία η εξατομίκευση περιεχομένου, πρέπει η έμφαση να δοθεί τόσο στο front-end, όσο και στο back-end. Ιδανικά ένας ιστότοπος που δίνει έμφαση στην προσωποποίηση περιεχομένου, προτείνει στον χρήστη περιεχόμενου που τον ενδιαφέρει, μειώνοντας έτσι την προσπάθειά του να βρει το περιεχόμενο που τον ενδιαφέρει, ενώ παράλληλα ενισχύει την ικανοποίηση του χρήστη.

Τα τελευταία χρόνια η έρευνα σχετικά με την ευχρηστία των ιστοσελίδων έχει οδηγήσει στην δημιουργία τεχνικών προσωποποίησης περιεχομένου. Το κομβικό σημείο για την επιτυχία μιας τέτοιας τεχνικής είναι να αναγνωρίσει επιτυχώς τα ενδιαφέροντα και τα χαρακτηριστικά ενός χρήστη, χωρίς να προϋποθέτει προσπάθεια από τον χρήστη, όπως π.χ μέσω της συμπλήρωσης ερωτηματολογίου. Από την άλλη ωστόσο, η μη προσπάθεια από τον χρήστη κάνει ακόμα πιο σημαντική την σωστή εύρεση των προτιμήσεων του χρήστη, καθώς σε περίπτωση παραγωγής περιεχομένου που δεν ενδιαφέρει τον χρήστη, η σχέση με την ιστοσελίδα θα φθίνει (Blom, Karat, Karat, 2006).

Προκειμένου να επιτευχθεί μια σωστή προσωποποίηση η οποία θα ζητάει όσο το δυνατόν λιγότερη προσπάθεια από τον χρήστη, η αυτοματοποίηση αλλά και η δυναμικότητα της εξατομίκευσης είναι ζωτικής σημασίας. Για να συμβεί αυτό, οι ιστοσελίδες οφείλουν να παίρνουν πληροφορίες για τους χρήστες του από πολλές πηγές, και μέσω ειδικών αλγορίθμων να τις αναλύουν, έτσι ώστε να γνωρίσουν τις προτιμήσεις κάθε χρήστη. Οι αλγόριθμοι αυτοί ωστόσο οφείλουν να είναι δυναμικοί και να προσαρμόζονται σε τυχόν αλλαγές των προτιμήσεων των χρηστών έτσι ώστε να παρέχουν όσο πιο σωστό εξατομικευμένο περιεχόμενο γίνεται.

Ωστόσο υπάρχει τρόπος οι ιστοσελίδες να δημιουργήσουν ένα αποτελεσματικό προσωποποιημένο περιεχόμενο. Για να συμβεί αυτό, πρέπει να εστιάσουν σε τρία σημεία:

1. Το περιεχόμενο.
2. Τον χρήστη.
3. Τους στόχους της ιστοσελίδας.

Πιο συγκεκριμένα, μια ιστοσελίδα πρέπει να έχει την δυνατότητα να γνωρίζει τις προτιμήσεις του χρήστη, να τις συνδέει με τις κατάλληλες προτάσεις που θα τον ενδιαφέρουν, ενώ παράλληλα προσαρμόζεται ανάλογα με τον τρόπο που κινείται και τι ψάχνει ο χρήστης μέσα στην σελίδα. Για να λειτουργήσει αυτός ο τρόπος, πρέπει η ιστοσελίδα αφού αναλύσει τις πληροφορίες που συνέλεξε, να τις τροφοδοτήσει σε ένα σύστημα τεχνητής νοημοσύνης (AI). Αυτό το σύστημα είναι υπεύθυνο έτσι ώστε να χωρίσει αυτές τις πληροφορίες σε προφίλ χρηστών, και έπειτα να τα συνδέσει με τις προτιμήσεις τους, ούτως ώστε να παραχθεί και το κατάλληλο εξατομικευμένο περιεχόμενο.

2.1.3 Η τεχνητή νοημοσύνη στην παραγωγή εξατομικευμένου περιεχομένου

Σύμφωνα με το Πανεπιστήμιο του Harvard, μέχρι το 1950 οι επιστήμονες είχαν ήδη την ιδέα της τεχνητής νοημοσύνης εντυπωμένη στις σκέψεις τους. Πιο συγκεκριμένα, ένας Βρετανός επιστήμονας, ο Alan Turing, εξερεύνησε την πιθανότητα χρήσης της τεχνητής νοημοσύνης. Στόχος της τεχνητής νοημοσύνης είναι η δημιουργία τεχνολογίας και συστημάτων τα οποία επιτρέπουν στους υπολογιστές να λειτουργήσουν με έναν έξυπνο τρόπο.

Λαμβάνοντας το παραπάνω υπόψιν, πολλοί κλάδοι θέλησαν να χρησιμοποιήσουν το AI, έτσι ώστε να αυξήσουν την αποτελεσματικότητα τους. Ανάμεσα σε αυτούς τους κλάδους πλέον συγκαταλέγεται και η δημοσιογραφία. Με την έκρηξη του διαδικτύου, οι περισσότεροι άνθρωποι πλέον λαμβάνουν τα νέα τους μέσα από διαδικτυακούς ιστότοπους. Ωστόσο, η πληθώρα των πηγών, - όπως έχει αναφερθεί – δημιούργησε την ανάγκη οι ενημερωτικές ιστοσελίδες να διαφοροποιηθούν. Ο καλύτερος τρόπος για να συμβεί αυτό είναι μέσω του περιεχομένου. Χρησιμοποιώντας την τεχνητή νοημοσύνη, οι ενημερωτικές ιστοσελίδες ενώ δεν αλλάζουν το περιεχόμενο τους, το προσαρμόζουν πάνω στις ανάγκες του κάθε χρήστη ξεχωριστά, προσφέροντάς του έτσι περισσότερο ενδιαφέροντα μονοπάτια για αυτόν, και όχι νέα που δεν τον αφορούν.

Ωστόσο, ενώ οι επιχειρήσεις που ασχολούνται απευθείας με τους ανθρώπους ως πελάτες (B2C) γνωρίζουν αρκετά καλά το κοινό τους, οι πήγες από τις οποίες λαμβάνουν γνώση για το κοινό τους είναι πολύ λίγες και πολλές φορές δεν παρέχουν χρήσιμες πληροφορίες. Παραδοσιακά, τέτοιες επιχειρήσεις λαμβάνουν πληροφορίες για το κοινό τους με βάση το εισόδημα του και την εργασία του. Ωστόσο, το διαδίκτυο και οι ηλεκτρονικές πλατφόρμες έρχονται να το αλλάξουν.

Ανεπτυγμένες μέθοδοι που χρησιμοποιούν τεχνητή νοημοσύνη δίνουν πλέον την δυνατότητα σε επιχειρήσεις να συγκεντρώσουν, αναγνωρίσουν, και κατηγοριοποιήσουν δεδομένα σχετικά με το κοινό τους. Μέσω άρτια σχεδιασμένων αλγορίθμων που χρησιμοποιούν το AI, όλες οι επιχειρήσεις οι οποίες δίνουν έμφαση στις προτιμήσεις του κοινού τους, όπως μια ενημερωτική ιστοσελίδα, μπορούν να συλλέξουν πληροφορίες-κλειδιά για τους χρήστες.

Επιπλέον, αφού συνελεχθούν όλες οι απαραίτητες πληροφορίες, ο αλγόριθμος θα είναι ικανός να παρατηρήσει μοτίβα μέσα στα δεδομένα, τα οποία τελικά θα αποτελέσουν την βάση έτσι ώστε η ιστοσελίδα να γνωρίζει άριστα τις προτιμήσεις των χρηστών της και να τους παρέχει αυτόματα ακριβώς την πληροφορία που επιθυμούν (Davenport, 2019).

Το συγκεκριμένο περιεχόμενο μπορεί να έχει την μορφή προτάσεων άρθρων που θα ενδιαφέρουν έναν συγκεκριμένο χρήστη, ή ακόμα και ολόκληρων σελίδων που θα αλλάζουν ανάλογα με τις προτιμήσεις του χρήστη. Το περιεχόμενο αυτό θα ταιριάζει με την συμπεριφορά του χρήστη μέσα στην σελίδα, τα δημογραφικά του χαρακτηριστικά, ακόμα και των παρελθοντικών του αναζητήσεων μέσα στον ιστοχώρο.

Για να λειτουργήσει επιτυχώς όμως η τεχνητή νοημοσύνη πρέπει τα άτομα που την χειρίζονται να ‘’διδάξουν’’ τους υπολογιστές τον στόχο που έχουν, έτσι ώστε εκείνοι πλέον να παράγουν αυτόματα το επιθυμητό αποτέλεσμα. Αυτό πραγματοποιείται μέσω μιας διαδικασίας που ονομάζεται μηχανική μάθηση.

2.2 Μηχανική μάθηση

Η επιστήμη της μηχανικής μάθησης (Machine learning) ασχολείται με υπολογιστικούς αλγόριθμους, οι οποίοι βελτιώνονται αυτόματα μέσα από την εμπειρία.

Η σημασία της μηχανικής μάθησης είχε αρχίσει ήδη να εκτιμάται από τις αρχές του 2000, όταν και ξεκίνησε να συνδέει διαφορετικές επιστήμες, όπως η βιολογία, τα μαθηματικά και η στατιστική, έτσι ώστε να βοηθήσει τους υπολογιστές να μάθουν (Marsland, 2009). Ενδεικτικά, πλέον η μηχανική μάθηση μας προσφέρει αμέτρητες τεχνολογικές επιτυχίες, όπως είναι τα αυτόματα αυτοκίνητα, τύπου Tesla, μας έχει βοηθήσει στην εξερεύνηση του διαδικτύου, και έχει συμβάλει τα μέγιστα καλύτερη κατανόηση του ανθρώπινου γονιδιώματος.

Επιβλεπόμενη μάθηση

Η πλειοψηφία των μοντέλων μηχανικής μάθησης χρησιμοποιεί την επιβλεπόμενη μάθηση – όπως θα χρησιμοποιηθεί και στην συγκεκριμένη εργασία. Με την επιβλεπόμενη μάθηση, ο ερευνητής έχει τόσο τις μεταβλητές που εισάγει, όσο και τις μεταβλητές-αποτελέσματα. Έχοντας αυτές τις μεταβλητές, χρησιμοποιεί έναν αλγόριθμο, ώστε να μετατρέψει τις μεταβλητές που εισήχθησαν, σε ένα αποτέλεσμα.

Στόχος είναι, αυτή η διαδικασία της μετατροπής να γίνεται τόσο αποτελεσματικά, ώστε όταν εισάγονται νέα δεδομένα, ο αλγόριθμος να μπορεί να προβλέψει το αποτέλεσμα για την ομάδα δεδομένων που έχει.

Ο λόγος για τον οποίο ονομάζεται επιβλεπόμενη μάθηση, είναι γιατί ο αλγόριθμος πρώτα μαθαίνει για το αποτέλεσμα που πρέπει να βγάλει, μέσα από ένα training-set δεδομένων. Αυτό ομοιάζει με τον τρόπο με τον οποίο οι δάσκαλοι επιβλέπουν τους μαθητές τους κατά την εκμάθηση τους (Brownlee, 2016).

Επιπλέον, η επιβλεπόμενη μάθηση στοχεύσει στην επίλυση δύο συγκεκριμένων προβληματικών. Πιο συγκεκριμένα:

1. **Γραμμική παλινδρόμηση** (**Regression):** Μέσω των συγκεκριμένων μοντέλων, ο αλγόριθμος θέλει να προβλέψει συνεχόμενες τιμές, όπως για παράδειγμα τι θερμοκρασία θα έχει αύριο (Johnston, Mathur, 2019).
2. **Ταξινόμηση (Classification):** Σε αντίθεση με τα μοντέλα γραμμικής παλινδρόμησης, η ταξινόμηση έχει ως στόχο να προβλέψει σωστά την κατηγορία στην οποία ανήκουν τα δεδομένα. Παραδείγματος χάριν, ένας τέτοιος αλγόριθμος μπορεί να προβλέψει εάν ένας άνθρωπος συγκαταλέγεται στα άτομα που έχουν μια συγκεκριμένη ασθένεια ή όχι.

Τέλος, η επιβλεπόμενη μάθηση χρησιμοποιεί κατά κύριο λόγο τους παρακάτω αλγόριθμους:

- Γραμμική παλινδρόμηση.
- Τυχαία δάση για γραμμική παλινδρόμηση και ταξινόμηση.
- Μηχανές διανυσμάτων υποστήριξης για ταξινόμηση.

Μη επιβλεπόμενη μάθηση

Στην μη επιβλεπόμενη μάθηση έχουμε μόνο τα δεδομένα που εισάγουμε, χωρίς τα αντίστοιχα αποτελέσματα. Στόχος είναι ο αλγόριθμος να μάθει την δομή και κατανομή των δεδομένων, έτσι ώστε να συλλέξει περισσότερες πληροφορίες για αυτά.

Στην συγκεκριμένη περίπτωση, σε αντίθεση με την επιβλεπόμενη μάθηση, ο αλγόριθμος δεν έχει από πριν το αποτέλεσμα έτσι ώστε να μάθει από αυτό. Συγκεκριμένα, στόχος αυτών των αλγόριθμων είναι να ανακαλύψουν μόνοι τους τα δεδομένα που έχουν εισαχθεί, και έπειτα να παρουσιάσουν ενδιαφέρουσες πληροφορίες.

Όπως και η επιβλεπόμενη, έτσι και η μη επιβλεπόμενη μάθηση, χωρίζεται σε δύο περαιτέρω προβληματικές:

1. **Ομαδοποίηση:** Όταν χρησιμοποιούμε την ομαδοποίηση στην μη επιβλεπόμενη μάθηση, ο αλγόριθμος θα χωρίσει τα δεδομένα σε διαφορετικές ομάδες, ανάλογα με τα κοινά τους χαρακτηριστικά. Ωστόσο, ο αλγόριθμος δεν γνωρίζει εκ των προτέρων ποιες ομάδες υπάρχουν, και πρέπει να τις ανακαλύψει μόνος του (Amr, 2020).
2. **Συσχέτιση:** Μέσω της συσχέτισης ο αλγόριθμος προσπαθεί να μάθει τους κανόνες που διέπουν ένα μεγάλο ποσοστό από τα δεδομένα μας, έτσι ώστε να βρει κοινές συμπεριφορές.

Όσο αφορά τους συγκεκριμένους αλγόριθμους που χρησιμοποιούνται κατά κόρον στην μη επιβλεπόμενη μάθηση, αυτοί είναι οι εξής:

1. **K-means**: Αυτός ο αλγόριθμος προσπαθεί να βρει την διάμεσο ανάμεσα στα δεδομένα, έτσι ώστε έπειτα να τα χωρίσει σε ομάδες, στις οποίες θα ανήκουν δεδομένα με την πλησιέστερη διάμεσο.
2. **Apriori**: Ο συγκεκριμένος αλγόριθμος προσπαθεί να βρει τα δεδομένα που εμφανίζονται πιο συχνά, και να τα συνδυάζει με ολοένα και μεγαλύτερες ομάδες δεδομένων, έτσι ώστε να ανακαλύψει συσχετίσεις στην βάση δεδομένων (Agrawal and Ramakrishnan Srikant, 1994).

Ημι-επιβλεπόμενη μάθηση

Τέλος, υπάρχει και η ημι-επιβλεπόμενη μάθηση, όπου έχουμε μόνο δεδομένα ως αποτέλεσμα και όχι όλα. Στην πραγματικότητα, πολλοί αλγόριθμοι λειτουργούν πλέον με ημι-επιβλεπόμενη μάθηση, καθώς συνδυάζουν τεχνικές τόσο της επιβλεπόμενης, όσο και της μη επιβλεπόμενης μάθησης.

Πιο συγκεκριμένα, αρκετοί αλγόριθμοι χρησιμοποιούν την μη επιβλεπόμενη μάθηση έτσι ώστε να κάνουν προβλέψεις για τα μη γνωστά δεδομένα. Αφού κάνουν τις προβλέψεις, έπειτα τις εισάγουν στον αλγόριθμο επιβλεπόμενης μάθησης ως training-set, έχοντας πλέον την δυνατότητα να κάνουν προβλέψεις για νέα δεδομένα (Brownlee, 2016).

2.2.1 Εξόρυξη δεδομένων

Ωστόσο, τα τελευταία χρόνια, μια συγκεκριμένη εξέλιξη έχει αρχίσει να επηρεάζει σημαντικά την πορεία της μηχανικής μάθησης, και αυτή είναι η εξόρυξη δεδομένων (data mining). Μέσω της εξόρυξης δεδομένων, στόχος είναι η συλλογή χρήσιμων πληροφοριών από μεγάλες βάσεις δεδομένων, μέσω συγκεκριμένων αλγορίθμων.

Η εξόρυξη δεδομένων έχει βασίσει την λειτουργία της σε τρεις μεθόδους (Gorunescu, 2011):

1. **Στατιστική**: Μέσω τη στατιστικής δίνεται η δυνατότητα να αναγνωριστούν συγκεκριμένες σχέσεις μεταξύ των δεδομένων, εφόσον δεν υπάρχει αρκετή πληροφορία σχετικά με την φύση τους.
2. **Υπολογιστικές μέθοδοι**: Περιγραφική στατιστική, συσχετίσεις, πίνακες, τεχνικές εξερεύνησης, και γραμμικά ή μη-γραμμικά μοντέλα, όλα ανήκουν στις υπολογιστικές μεθόδους που είναι απαραίτητες για την εξόρυξη δεδομένων.
3. **Οπτικοποίηση δεδομένων**: Μέσω της οπτικοποίησης δεδομένων, στόχος είναι η οπτική παρουσίαση της πληροφορίας. Όντας σήμερα στην εποχή της εικόνας, η οπτική απεικόνιση αποτελεί την πιο δυνατή και ελκυστική μέθοδο της εξερεύνησης δεδομένων.

Λειτουργίες της εξόρυξης δεδομένων

Με την αρχή της ενασχόλησης με την εξόρυξη δεδομένων, κανείς μπορεί να πραγματοποιήσεις τις εξής έξι δραστηριότητες:

1. **Περιγραφή**: Είναι πολύ συχνό το φαινόμενο οι ερευνητές να προσπαθούν να ανακαλύψουν τρόπους για να περιγράψουν σχέδια και συσχετίσεις που κρύβονται μέσα στα δεδομένα. Οι περιγραφές των συγκεκριμένων συσχετίσεων είναι ικανές και να δώσουν εξηγήσεις για την φύση τους. Για να επιτευχθεί μια σωστή περιγραφή ωστόσο, είναι αναγκαίο το μοντέλο εξόρυξης δεδομένων που χρησιμοποιείται, να είναι όσο πιο διαφανές γίνεται (Larose, 2015).
2. **Εκτίμηση**: Μέσω της εκτίμησης, το μοντέλο εξόρυξης δεδομένων έχει ως στόχο να βρει ένα αριθμητικό αποτέλεσμα. Τέτοιες εκτιμήσεις μπορεί να είναι για παράδειγμα το ποσό που μπορεί να ξοδέψει μια οικογένεια για διακοπές, ή τι με βαθμό θα αποφοιτήσει ένας μεταπτυχιακός φοιτητής, με βάση τον βαθμό με τον οποίοι αποφοίτησε ως προπτυχιακός.
3. **Πρόβλεψη**: Με μια αρχική ματιά, η πρόβλεψη μοιάζει αρκετά με την ταξινόμηση και την εκτίμηση. Η διαφορά τους ωστόσο βρίσκεται στο ότι η πρόβλεψη αφορά αποκλειστικά και μόνο στο μέλλον. Παράδειγμα εκτίμησης μπορεί να είναι τα χρήματα που θα έχει δαπανήσει η Ελλάδα στον τομέα της υγείας μέχρι το τέλος του 2021.
4. **Ταξινόμηση**: Η ταξινόμηση μοιάζει με την εκτίμηση, με την διαφορά τους να έγκειται στο γεγονός πως για την πρώτη, η μεταβλητή-στόχος είναι κατηγορική και όχι αριθμητική. Ένα μοντέλο εξόρυξης δεδομένων που ασχολείται με την ταξινόμηση, θα έχει μια μεταβλητή η οποία θα μπορεί να χωριστεί σε ξεχωριστές υποκατηγορίες.
5. **Ομαδοποίηση (clustering):** Ελλείψει συγκεκριμένων πληροφοριών για τα δεδομένα, η ομαδοποίηση στοχεύει να δημιουργήσει ένα μοντέλο το οποίο θα ερμηνεύσει συγκεκριμένα δεδομένα ως ένα σύνολο. Το συγκεκριμένο σύνολο θα έχει δεδομένα τα οποία είναι όσο πιο όμοια γίνεται μεταξύ τους (Aggarwal, Reddy, 2007).
6. **Συσχέτιση**: Μέσω της συσχέτισης, μπορούμε να βρούμε την σχέση μεταξύ πολλών δεδομένων. Με αυτή την λειτουργία, το μοντέλο μπορεί να βρει για παράδειγμα, ποια νέα αρέσουν σε έναν χρήστη, και να του προτείνει ανάλογα νέα στο μέλλον.

Συμπεράσματα

Σε αυτό το κεφάλαιο ασχοληθήκαμε με τον ορισμό εξατομίκευσης περιεχομένου, η οποία έχει ως απώτερο στόχο να προσαρμόσει την πληροφορία από μια ιστοσελίδα στα ‘’θέλω’’ του χρήστη, καθώς έτσι θα χτιστεί μια αμοιβαία εμπιστοσύνη ανάμεσα στο κοινό και στον πάροχο της πληροφορίας.

Χρησιμοποιώντας το εξατομικευμένο περιεχόμενο, μια ενημερωτική σελίδα μπορεί να προσφέρει στους χρήστες της ακριβώς το περιεχόμενο που επιθυμούν, χωρίς εκείνοι να καταβάλλουν μεγάλη προσπάθεια από μέρους τους. Επιπλέον, το εξατομικευμένο

περιεχόμενο μπορεί να αποδειχθεί ειδοποιός διάφορα ανάμεσα στις ενημερωτικές ιστοσελίδες.

Οι λύσεις που προσφέρει το εξατομικευμένο περιεχόμενο, επιτυγχάνονται μέσω της μηχανικής μάθησης, όσο και με τις κατηγορίες στις οποίες αυτή χωρίζεται: επιβλεπόμενη, μη-επιβλεπόμενη, και ημι-επιβλεπόμενη μάθηση.

Στην συνέχεια εμβαθύναμε στην εξόρυξη δεδομένων και σε τι αποσκοπούν οι λειτουργίες της, προτού αναφερθούμε στην εξατομίκευση.

Στο επόμενο κεφάλαιο θα δούμε πως δημιουργείται πρακτικά η εξατομίκευση, μέσω τεχνικών παραγωγής συστάσεων (recommender systems).

Κεφάλαιο 3

Τεχνικές παραγωγής συστάσεων

3.1 Συστήματα παραγωγής συστάσεων

Οι τεχνικές παραγωγής συστάσεων (recommender systems) αποτελούν εργαλεία που στοχεύουν στον χρήστη. Οι συστάσεις τις οποίες κάνουν έχουν σκοπό να βοηθήσουν τον χρήστη καθόλη την διάρκεια της περιήγησης του στο διαδίκτυο, παρέχοντας βοήθεια για το ποια προϊόντα να αγοράσει, ποια νέα να διαβάσει, ή ακόμα και τι μουσική να ακούσει (Kembellec, Saleh, Chartron, 2014).

Από την αρχή της δημιουργίας τους τα recommender systems έχουν αποδείξει με μεγάλη επιτυχία το πόσο σημαντικά και αναγκαία είναι. Όπως έχει αναφερθεί ήδη, η εξατομίκευση περιεχομένου βοηθά εξαιρετικά τους χρήστες στο να ανταπεξέλθουν απέναντι στην πληθώρα που υπάρχει αυτή τη στιγμή στο διαδίκτυο, ενώ είναι χρήσιμα εργαλεία και για το marketing (Ricci, Rokach, Shapira, Kantor, 2010). Δεν είναι τυχαίο άλλωστε πως εταιρείες κολοσσοί, όπως είναι η Amazon, η Google, το Netflix και το Spotify, όλες πλέον χρησιμοποιούν ειδικούς αλγόριθμους, προκειμένου να προσφέρουν στους χρήστες περιεχόμενο αποκλειστικά και μόνο βασισμένο στις δικές τους προτιμήσεις.

Όσο αφορά στην δημιουργία των recommender systems, αυτή είναι αποτέλεσμα μιας συνεργασίας πολλών κλάδων, οι οποίοι περιλαμβάνουν μεταξύ άλλων:

- Τεχνητή Νοημοσύνη.
- Αλληλεπίδραση ανθρώπου και υπολογιστή.
- Τεχνολογία πληροφορίας.
- Εξόρυξη δεδομένων.
- Στατιστική.
- Marketing.

Πώς ακριβώς δουλεύουν όμως τα recommender systems; Αρχικά, για να λειτουργήσει ένα τέτοιο σύστημα, πρέπει να πάρει συγκεκριμένες πληροφορίες για τις προτιμήσεις του χρήστη. Τις πληροφορίες αυτές τις λαμβάνει είτε παρατηρώντας τις ενέργειες που πραγματοποιεί ο χρήστης – και μαθαίνει από αυτές -, είτε ζητώντας από τον χρήστη να δηλώσει τι του αρέσει και τι όχι. Αφού έχει συγκεντρώσει όλες τις απαραίτητες πληροφορίες, θα καταλήξει σε συγκεκριμένες προτάσεις, οι οποίες θα απευθύνονται σε κάθε χρήστη ξεχωριστά, ανάλογα με τις πληροφορίες που υπάρχουν για αυτόν (Neumann, 2009).

Ο τρόπος με τον οποίο λειτουργούν οι τεχνικές παραγωγής συστάσεων είναι με την δημιουργία ενός προφίλ για κάθε χρήστη. Μέσω πληροφοριών όπως τι άρθρα έχει διαβάσει, τα tags των συγκεκριμένων άρθρων, τις αξιολογήσεις, και την ομοιότητα με άλλους άλλους χρήστες, τα recommender systems δημιουργούν ένα ολοκληρωμένο προφίλ για τους χρήστες μιας ιστοσελίδας, προκειμένου να προχωρήσουν και στις ανάλογες προτάσεις άρθρων.

Οι λεγόμενες και ως προσαρμοστικές μέθοδοι διαλέγουν αντικείμενα τα οποία συμβαδίζουν με τις απόψεις κάθε νέου χρήστη. Πιο συγκεκριμένα, για αντικείμενα τα οποία ταιριάζουν με τις προτιμήσεις του χρήστη στο παρόν, οι προσαρμοστικές μέθοδοι κάνουν το σύστημα να προσαρμόζεται σε παλαιότερες αξιολογήσεις που είχε δώσει ο καινούργιος χρήστης.

Με αυτόν τον τρόπο, ο χρήστης βαθμολογεί αντικείμενα τα οποία είναι πιο κοντά στις προτιμήσεις του, ενώ η όλη διαδικασία της βαθμολόγησης γίνεται σε ένα ελεγχόμενο περιβάλλον. Το γεγονός αυτό, καθιστά αυτού του είδος την βαθμολόγηση πιο αποτελεσματική από ότι στις μη-προσαρμοστικές μεθόδους.

Επιπλέον, οι προσαρμοστικές μέθοδοι λαμβάνουν υπόψιν τις παλαιότερες αξιολογήσεις του χρήση , όπως επίσης και το συνεχώς εναλλασσόμενο προφίλ του χρήστη. Με αυτόν τον τρόπο, ο χρήστης θα δει περισσότερα αντικείμενα τα οποία θα τον ενδιαφέρουν.

Παραδείγματα προσαρμοστικών μεθόδων αποτελούν τα παρακάτω:

- Παρουσίαση αντικειμένων μέχρι ο χρήστης να δώσει έστω μια βαθμολογία.
- Naive Bayes: Με αυτόν τον τρόπο τα αντικείμενα τα οποία θα επιλεχθούν να παρουσιαστούν στον χρήστη για να τα βαθμολογήσει, θα είναι αυτά τα οποία έχουν την μεγαλύτερη πιθανότητα να βαθμολογηθούν.
- Ομαδοποίηση (Clustering): Αυτός ο τρόπος ψάχνει να βρει τις διαμέσους μεταξύ των αντικειμένων (means), προκειμένου να τα χωρίσει σε ομάδες. Με αυτήν την μέθοδο, ο αλγόριθμος έχει την ικανότητα να επιλέξει ποια αντικείμενα να παρουσιάσει στην συνέχεια στον χρήστη, προκειμένου να τα βαθμολογήσει.

Τέτοιοι αλγόριθμοι μπορούν να επιφέρουν πολλά και σημαντικά πλεονεκτήματα. Αρχικά, ένα πολύ σημαντικό πλεονέκτημα είναι πως η δημιουργία ενός αλγόριθμου παραγωγής συστάσεων θα μειώσει αισθητά το κόστος αναζήτησης προϊόντων, τα οποία θα ταιριάζουν σε συγκεκριμένες ανάγκες ενός πελάτη. Δεύτερον, τα recommender systems μπορούν να χρησιμοποιηθούν ως εργαλεία τα οποία θα αναδείξουν την αλληλεπίδραση που υπάρχει με το κοινό. Με την σειρά του, αυτό θα ενισχύσει την σχέση κοινού και παρόχου της πληροφορίας στο μέλλον.

Όσο αφορά στις μη- προσαρμοστικές μεθόδους, δεν θεωρούνται κομμάτι της εξατομίκευσης περιεχομένου, καθώς δεν προσφέρουν κάποια ταύτιση με τα προφίλ των χρηστών που έχουν δημιουργηθεί. Παρόλα αυτά, παραδείγματα παρόμοιων τεχνικών, είναι μεταξύ άλλων:

1. Η τυχαία διαλογή αντικειμένων.
2. Η δημοτικότητα αντικειμένων, δηλαδή τα αντικείμενα αυτά που έχουν βαθμολογήσει οι πιο πολλοί χρήστες.

Ωστόσο, οι τεχνικές παραγωγής συστάσεων, χωρίζονται σε κατηγορίες, ανάλογα με το τι πληροφορίες ζητούν από τον χρήστη, πώς λαμβάνουν αυτές τις πληροφορίες, και τι σκοπό έχουν (Jannach, Zanker, Felfernig, Friedrich, 2010). Οι τρεις κατηγορίες αυτές είναι:

- Το φιλτράρισμα βάσει περιεχομένου (Content-based filtering).
- Το συνεργατικό φιλτράρισμα (Collaborative filtering).
- Το υβριδικό φιλτράρισμα (Hybrid Filtering).

3.1.1 Φιλτράρισμα βάσει περιεχομένου (Content-based filtering)

Το φιλτράρισμα βάσει περιεχομένου προτείνει αντικείμενα σύμφωνα με τις προτιμήσεις του χρήστη, αφού έχει αναλυθεί η προηγούμενή του δραστηριότητα ή σύμφωνα με όσα έχει δηλώσει ο ίδιος ότι του αρέσουν.

Πιο συγκεκριμένα το σύστημα χωρίζει τα αντικείμενα μιας ιστοσελίδας - εν προκειμένω τα άρθρα, ανάλογα με το αν ενδιαφέρουν ή όχι τον χρήστη. Για να καθορίσει αν ένα άρθρο σχετίζεται θετικά με τον χρήστη, το φιλτράρισμα βάσει περιεχομένου προσπαθεί να βρει κοινά σημεία μεταξύ των άρθρων, όπως οι λέξεις, τα tags, οι τίτλοι, και οι λέξεις κλειδιά (keywords).

Για κάθε άρθρο, το φιλτράρισμα βάσει περιεχομένου υπολογίζει την απόσταση του συγκριτικά με άλλα άρθρα που έχει διαβάσει ο χρήστης, προκειμένου να δημιουργήσει ομάδες με κοντινά άρθρα, δηλαδή με αρκετά κοινά στοιχεία. Στη συνέχεια, ο αλγόριθμος αναλύει άρθρα τα οποία ο χρήστης δεν έχει διαβάσει, με σκοπό να προβλέψει αν τα στοιχεία των άρθρων αυτών θα ενδιαφέρουν τον χρήστη. Τέλος, αφού αναλυθούν όλα τα παραπάνω, ο αλγόριθμος συγκεντρώνει τα άρθρα με τα περισσότερα κοινά στοιχεία, τα οποία και προτείνει στον χρήστη. Όσο πιο πολλές πληροφορίες λαμβάνει το σύστημα για τον χρήστη, τόσο πιο ικανό θα είναι τελικά να δημιουργήσει το προφίλ του, το οποίο θα είναι βασισμένο στα άρθρα εκείνα για τα οποία έχει εκφραστεί θετικά.

Επίσης, προκειμένου το σύστημα να παράξει συγκεκριμένες προτάσεις για τους χρήστες, χρειάζεται δύο πολύ σημαντικές πληροφορίες: το ιστορικό προτιμήσεων του χρήστη, όπως επίσης και έναν τρόπο προκειμένου να συγκρίνει πόσο όμοια είναι δύο αντικείμενα, όπως π.χ να δηλώσει ο χρήστης αν ένα άρθρο του αρέσει ή όχι. Αυτό επιτυγχάνεται με την χρήση της συνάρτησης της ομοιότητας συνημιτόνου (cosine similarity) (Castellano, 2009).

Αφού δημιουργηθεί το προφίλ του χρήστη, θα περιέχει όλα τα αντικείμενα τα οποία του αρέσουν, ενώ θα μπορεί να προσαρμοστεί και σε μια αλλαγή ενδιαφέροντος του χρήστη (Mohanty, Chatterjee, Jain, Elngar, Gupta, 2020).

Για να επιτευχθεί αυτή η διαδικασία, το σύστημα εκτελεί τρία διαφορετικά βήματα:

1. **Ανάλυση περιεχομένου**: Σε περίπτωση που τα δεδομένα δεν είναι δομημένα, θα πρέπει πρώτα να γίνει μια επεξεργασία, έτσι ώστε να μπορέσει το σύστημα να εξάγει κάποιες βασικές πληροφορίες. Σε αυτό το βήμα, το σύστημα εξετάζει όλες τις πληροφορίες που διαθέτει προκειμένου να προσδώσει συγκεκριμένα χαρακτηριστικά στο δεδομένα.
2. **Μάθηση**: Σε αυτό το βήμα ο αλγόριθμος φτιάχνει το προφίλ του χρήστη από τα δεδομένα που έχει συγκροτήσει από το προηγούμενο βήμα. Έπειτα, μέσω της μηχανικής μάθησης, το σύστημα θα είναι ικανό να γνωρίζει τι ακριβώς αρέσει στον χρήστη και τι δεν του αρέσει.
3. **Φιλτράρισμα συστατικών**: Σε αυτό το βήμα το σύστημα χρησιμοποιεί το προφίλ του χρήστη που δημιούργησε από το προηγούμενο στάδιο προκειμένου να καταλήξει σε σχετικά αντικείμενα. Αυτό συμβαίνει αφού ο αλγόριθμος ταιριάξει το προφίλ του χρήστη με τα αντικείμενα τα οποία θα του προταθούν.

Θετικά φιλτραρίσματος βάσει περιεχομένου

Ένα από τα πλεονεκτήματα τα οποία προσφέρει η συγκεκριμένη τεχνική φιλτραρίσματος είναι πως μπορεί να κάνει προτάσεις και για νέα αντικείμενα, όταν δεν υπάρχουν επαρκείς βαθμολογίες για συγκεκριμένα από τα δεδομένα.

Ο αλγόριθμος έχει αυτή την ικανότητα γιατί μπορεί να υπάρχουν άλλα δεδομένα με παρόμοια χαρακτηριστικά τα οποία ο χρήστης έχει ήδη βαθμολογήσει.

Χρησιμοποιώντας τις υπάρχουσες βαθμολογίες, σε συνδυασμό με τα χαρακτηριστικά που έχουν αποδοθεί στα δεδομένα, το σύστημα μπορεί να προσφέρει εξατομικευμένες προτάσεις στον χρήστη (Aggarwal, 2016).

Αρνητικά φιλτραρίσματος βάσει περιεχομένου

Αρχικά οι αλγόριθμοι του content-based filtering βασίζονται σε μεγάλο βαθμό στις λέξεις-κλειδιά (keywords) των δεδομένων. Αυτό έχει ως αποτέλεσμα να μην εμφανίζονται στον χρήστη άρθρα τα οποία δεν περιέχουν τις λέξεις-κλειδιά που έχει χρησιμοποιήσει.

Αυτό συμβαίνει γιατί οι προτάσεις που παράγονται έχουν φτιαχτεί αποκλειστικά και μόνο βασισμένες στις προτιμήσεις του χρήστη, και δεν λαμβάνουν υπόψη τις προτιμήσεις παρόμοιων χρηστών. Το γεγονός αυτό δεν προσδίδει ιδιαίτερη ποικιλία στις προτάσεις του συστήματος (Blattberg, 2008).

Επιπροσθέτως, όπως αναφέρθηκε και παραπάνω, όταν ο αλγόριθμος φιλτράρει βάσει περιεχομένου, μπορεί να κάνει προτάσεις για νέα αντικείμενα, δεν μπορεί όμως να κάνει προτάσεις για νέους χρήστες.

Αυτό συμβαίνει καθώς ο αλγόριθμος χρειάζεται να ξέρει πρώτα τις προτιμήσεις του χρήστη, ούτως ώστε να καταλήξει και σε συγκεκριμένες προτάσεις. Για αυτό το λόγο, αρκετές αξιολογήσεις των χρηστών για τα δεδομένα είναι αναγκαίες.

3.1.2 Συνεργατικό φιλτράρισμα (Collaborative filtering)

Προκειμένου να επιλυθούν τα προβλήματα που παρατηρούνται με το φιλτράρισμα βάσει περιεχομένου, το συνεργατικό φιλτράρισμα χρησιμοποιεί ομοιότητες μεταξύ τόσο των χρηστών όσο και των αντικειμένων, προκειμένου να δημιουργήσει το προφίλ του χρήστη, και να καταλήξει και στις ανάλογες προτάσεις (Herlocker, Konstan, Terveen, Riedl, 2004).

Ο τρόπος με τον οποίο λειτουργεί το φιλτράρισμα είναι πως χρήστες οι οποίοι έχουν δηλώσει πως τους αρέσουν τα ίδια αντικείμενα στο παρελθόν, είναι πολύ πιθανό να έχουν παρόμοιες προτιμήσεις και στο μέλλον. Βασισμένο σε αυτή τη λογική, το σύστημα ψάχνει να βρει χρήστες με παρόμοια ενδιαφέροντα. Παραδείγματος χάριν, ο χρήστης Χ βρίσκεται σε μια σελίδα ειδησεογραφικού ενδιαφέροντος και διαβάζει άρθρα. Για όσα άρθρα δεν έχει διαβάσει ακόμα ο χρήστης Χ, ο αλγόριθμος θα ψάξει και θα εντοπίσει άλλους χρήστες οι οποίοι στο παρελθόν είχαν παρόμοιες προτιμήσεις στα άρθρα που διαβάζουν, με τον χρήστη Χ, και τα έχουν αξιολογήσει. Με τον τρόπο αυτό, ο αλγόριθμος θα μπορέσει να καταλάβει αν ένα συγκεκριμένο άρθρο θα αρέσει τελικά στον χρήστη Χ και να του το προτείνει.

Συνοπτικά, το συνεργατικό φιλτράρισμα προτείνει άρθρα βασιζόμενο στο γεγονός πως χρήστες με κοινές προτιμήσεις, είναι πολύ πιθανό να συνεχίσουν να έχουν τις ίδιες προτιμήσεις και στο μέλλον.

Ωστόσο, ένα πολύ σημαντικό βήμα για την σωστή λειτουργία του συνεργατικού φιλτραρίσματος, είναι οι αξιολογήσεις των χρηστών. Αυτή την στιγμή υπάρχουν δύο τρόποι

με τους οποίους το σύστημα μπορεί να καταλάβει τις προτιμήσεις τον χρηστών. Με τον πρώτο τρόπο – και πιο άμεσο – οι χρήστες βαθμολογούν τα αντικείμενα που τους ενδιαφέρουν. Αυτό μπορεί να γίνει, είτε με τους χρήστες να βαθμολογούν βάση μιας αριθμητικής κλίμακας, είτε μέσω Like/Dislike.

Ο δεύτερος τρόπος είναι πιο έμμεσος. Εδώ οι χρήστες δεν δηλώνουν ρητά πια από τα αντικείμενα τους αρέσουν ή όχι, αλλά το σύστημα το κάνει από μόνο του. Αυτό το καταφέρνει αφού αναλύσει όλες τις κινήσεις που έκανε ο χρήστης μέσα στην ιστοσελίδα. Το σύστημα θα εστιάσει σε κινήσει όπως, ποιες σελίδες επέλεξε να δει ο χρήστης, πόση ώρα κάθισε σε αυτές και αν έχει αγοράσει κάτι. Αυτός ο αυτόματος τρόπος για την συγκέντρωση προτιμήσεων των χρηστών θεωρείται και πιο αποτελεσματικός.

Με τον τρόπο αυτό, το συνεργατικό φιλτράρισμα δίνει την δυνατότητα στο σύστημα να δημιουργήσει ομάδες χρηστών με παρόμοια ενδιαφέροντα. Κατά κάποιο τρόπο λοιπόν οι χρήστες συνεργάζονται μεταξύ τους, προκειμένου ο αλγόριθμος να τους προσφέρει τις καλύτερες δυνατές προτάσεις για τον καθένα ξεχωριστά (Bouza, 2012).

Αυτή τη στιγμή το συνεργατικό φιλτράρισμα είναι η πιο δημοφιλής μέθοδος παραγωγής συστάσεων τους χρήστες. Δεν είναι τυχαίο πως ένα από τα μεγαλύτερα site ηλεκτρονικού εμπορίου, χρησιμοποιεί αυτή τη στιγμή το collaborative filtering προκειμένου να προσφέρει εξατομικευμένες προτάσεις για προϊόντα στο κοινό του.

Το συνεργατικό φιλτράρισμα ωστόσο μπορεί να χωριστεί σε δύο διαφορετικές υποκατηγορίες, τις οποίες θα αναλύσουμε στις αμέσως επόμενες παραγράφους (Ekstrand, 2011):

- Το User Based Collaborative Filtering, και
- Το Item Based Collaborative Filtering.

User Based Collaborative Filtering

Το user based collaborative filtering ήταν πρώτη μέθοδος που χρησιμοποιήθηκε για το συνεργατικό φιλτράρισμα. Χρησιμοποιήθηκε για πρώτη φορά ως μέθοδος παραγωγής συστάσεων για άρθρα του GroupLens Usenet (Ekstrand, Riedl, Konstan, 2011). Το user based collaborative filtering χρησιμοποιεί την βάση του συνεργατικού φιλτραρίσματος. Δηλαδή ο αλγόριθμος ψάχνει να βρει χρήστες των οποίων οι αξιολογήσεις είναι παρόμοιες με αυτές του τωρινού χρήστη, ώστε τελικά να καταλήξεις στις καλύτερες δυνατές εξατομικευμένες προτάσεις.

Προκειμένου ο αλγόριθμος να καταλήξει εάν ένα συγκεκριμένο αντικείμενο αρέσει στον χρήστη ή όχι, το οποίο ωστόσο δεν έχει αξιολογήσει, το σύστημα θα ψάξει να βρει χρήστες που παρουσιάζουν την μεγαλύτερη συμφωνία σε αντικείμενα τα οποία έχουν βαθμολογήσει και οι δύο. Όσο πιο με κοντά είναι βαθμολογίες των χρηστών τόσο πιο πιθανό είναι να τους προταθούν παρόμοια άρθρα.

Για να καθοριστεί ο βαθμός ομοιότητας των χρηστών το user based collaborative filtering χρησιμοποιεί κυρίως την συνάρτηση Pearson (Pearson Correlation). Με αυτή την μέθοδο ο αλγόριθμος αναλύει την στατιστική συσχέτιση μεταξύ των αξιολογήσεων των χρηστών, προκειμένου να βγάλει συμπεράσματα για την ομοιότητα τους.

Παρόλη την αποτελεσματικότητα της συγκεκριμένης μεθόδου, παρουσιάζει και ένα σημαντικό μειονέκτημα. Δυσκολεύεται αρκετά στο να επεκταθεί σε βάσεις δεδομένων με

πολλούς χρήστες. Προκειμένου να επεκταθεί το συνεργατικό φιλτράρισμα σε μεγαλύτερα δεδομένα, θεωρήθηκε αναγκαία η δημιουργία καινούργιων αλγορίθμων, τους οποίους θα τους αναλύσουμε στις αμέσως επόμενες παραγράφους.

Item Based Collaborative Filtering

Την λύση στο πρόβλημα επέκτασης του συνεργατικού φιλτραρίσματος σε μεγάλες βάσεις δεδομένων, φαίνεται να δίνει η μέθοδος του item based collaborative filtering. Σήμερα, η συγκεκριμένη μέθοδος είναι από τις περισσότερο χρησιμοποιημένες στο τομέα του συνεργατικού φιλτραρίσματος (Sarwar, Karypis, Konstan, Riedl, 2001).

Σε αντίθεση με το user based collaborative filtering, το item based collaborative filtering δεν ασχολείται τόσο με τις ομοιότητες μεταξύ των αξιολογήσεων των χρηστών, όσο με τις ομοιότητες μεταξύ των μοτίβων των αξιολογήσεων των αντικειμένων. Για παράδειγμα, αν κάποια δεδομένα έχουν τους ίδιες χρήστες οι έχουν δηλώσει πως τους αρέσουν ή δεν τους αρέσουν, τότε αυτοί οι χρήστες θεωρούνται παρόμοιοι. Αυτό καθιστά πολύ πιθανό το ενδεχόμενο να έχουν και παρόμοιες προτιμήσεις και για μελλοντικά αντικείμενα.

Η εστίαση του item based collaborative filtering στα αντικείμενα προσομοιάζει με αυτόν τον τρόπο στο φιλτράρισμα βάσει περιεχομένου (content-based filtering). Ωστόσο, εμπίπτει στην κατηγορία του συνεργατικού φιλτραρίσματος καθώς καθορίζει την ομοιότητα μεταξύ των αντικειμένων με βάση τα μοτίβα των προτιμήσεων των χρηστών και όχι αναλύοντας μόνο τα δεδομένα για εξάγει πληροφορίες.

Τέλος, για να κατανοήσει ο αλγόριθμος ποια αντικείμενα είναι παρόμοια βάσει των αξιολογήσεων των χρηστών, χρησιμοποιείται η συνάρτηση της απόστασης συνημιτόνου (cosine). Ο αλγόριθμος αναλύει τα άρθρα προκειμένου να καταλήξει στις αποστάσεις μεταξύ τους. Με αυτόν τον τρόπο θα δημιουργήσει ομάδες οι οποίες έχουν λάβει παρόμοια βαθμολογία από τους χρήστες. Χρησιμοποιώντας το cosine similarity, το σύστημα θα υπολογίσει την απόσταση μεταξύ των άρθρων, δημιουργώντας ομάδες με τα πιο κοντινά άρθρα σχετικά με κάθε χρήστη. Τα άρθρα που βρίσκονται στην κοντινότερη απόσταση, θα είναι και αυτά που τελικά θα προταθούν.

Αρνητικά συνεργατικού φιλτραρίσματος

Όπως αναφέρθηκε και παραπάνω, στόχος του συνεργατικού φιλτραρίσματος είναι να αναλύσει τις αξιολογήσεις των χρηστών για δεδομένα, προκειμένου να καταλήξει σε παρόμοια προφίλ και μοτίβα. Για τον λόγο αυτό, είναι εύκολο για ένα τέτοιο σύστημα να υπολογίσει τις ομοιότητες των χρηστών, εφόσον όμως υπάρχουν επαρκείς αξιολογήσεις από τους χρήστες. Για τον λόγο αυτό, το συνεργατικό φιλτράρισμα μπορεί να αντιμετωπίσει προβλήματα στο να παρουσιάσει προτάσεις σε έναν νέο χρήστη, καθώς δεν υπάρχει ακόμα ο επαρκής αριθμός αξιολογήσεων. Τότε παρουσιάζεται το λεγόμενο πρόβλημα της κρύας εκκίνησης (cold start) (Zhao, 2016).

Για να λυθεί το συγκεκριμένο πρόβλημα η πιο αποτελεσματική και άμεση λύση είναι να παρουσιάσουμε στον νέο χρήση κάποια αντικείμενα, πχ άρθρα, και να του ζητήσουμε να τα βαθμολογήσει, παραδείγματος χάριν από το 1 μέχρι το 5. Με αυτόν τον τρόπο θα μπορέσουμε να βοηθήσουμε γρήγορα τον αλγόριθμο στο να δημιουργήσει το προφίλ του νέου χρήστη και να του παρουσιάσει εξατομικευμένες προτάσεις.

Ωστόσο, τα αντικείμενα που θα παρουσιαστούν στον χρήστη για να τα αναλύσει πρέπει να είναι προσεκτικά διαλεγμένα και να μπορούν να προσφέρουν χρήσιμες πληροφορίες στο σύστημα. Πιο συγκεκριμένα, μια σωστά δομημένη στρατηγική, η οποία θα παρουσιάζει στον χρήστη συγκεκριμένα άρθρα, θα βοηθήσει τον αλγόριθμο να κάνει πολύ καλύτερες προτάσεις, από ότι αν οι χρήστες διάλεγαν από μόνοι τους τα άρθρα τα οποία θα βαθμολογούσαν. Αυτό συμβαίνει γιατί η διαδικασία της αξιολόγησης πρέπει είναι όσο πιο εύκολη για τους χρήστες, και να στοχεύει στο να διασφαλίσει την προσπάθεια του χρήστη, αλλά ταυτόχρονα και την αποτελεσματικότητα των προβλέψεων του συστήματος (Nadimi-Shahraki, Bahadorpour, 2014).

3.1.3 Υβριδικό σύστημα παραγωγής συστάσεων (Hybrid recommender system)

Στόχος των συστημάτων παραγωγής προτάσεων είναι να αναλύσουν με τέτοιο βαθμό τις προτιμήσεις των χρηστών, ούτως ώστε να τους παρέχουν εξατομικευμένες προτάσεις. Πλέον έχουν γίνει αναπόσπαστο κομμάτι όχι μόνο του ηλεκτρονικού εμπορίου, αλλά και των μέσων μαζικής ενημέρωσης, τα οποία πλέον επιθυμούν να προσφέρουν κάτι καινούργιο στο κοινό τους, και αυτό είναι η προσωποποίηση του περιεχομένου.

Προκειμένου να επιτευχθεί αυτό έχουν δοκιμαστεί δύο κυρίως μέθοδοι παραγωγής συστάσεων: το φιλτράρισμα βάσει περιεχομένου και το συνεργατικό φιλτράρισμα. Και οι δύο αυτές μέθοδοι επιτυγχάνουν τον σκοπό της εξατομίκευσης, ωστόσο με διαφορετικούς τρόπους, ενώ και οι δύο έχουν τόσο μειονεκτήματα, όσο και πλεονεκτήματα.

Για να χρησιμοποιηθούν τα πλεονεκτήματα και από τους δύο τρόπους, οι δύο αυτές μέθοδοι πολύ συχνά συνδυάζονται, προσφέροντας έναν υβριδικό τρόπο παραγωγής συστάσεων (Burke, 2002). Ο συγκεκριμένος τρόπος θα χρησιμοποιηθεί και σε αυτήν την εργασία.

Για του λόγου το αληθές, πλέον τα περισσότερα συστήματα παραγωγής συστάσεων χρησιμοποιούν αυτή την υβριδική τεχνική, συνδυάζοντας content-based και collaborative filtering. Αυτό συμβαίνει κυρίως για να αντιμετωπιστούν με επιτυχία τα δύο σημαντικότερα προβλήματα των δύο μεθόδων που αναλύθηκαν παραπάνω. Δηλαδή, το πρόβλημα της επέκτασης σε μεγαλύτερες βάσεις δεδομένων για το φιλτράρισμα βάσει περιεχομένου, και το πρόβλημα της κρύας εκκίνησης για το συνεργατικό φιλτράρισμα (Hoekstra. 2010).

Πιο συγκεκριμένα, ας πάρουμε για παράδειγμα μια ειδησεογραφική ιστοσελίδα. Τα άρθρα που υπάρχουν στην σελίδα αυτή περιέχουν συγκεκριμένες λέξεις-κλειδιά, όπως είναι η κατηγορία τους και τα tags. Αυτές οι λέξεις-κλειδιά ονομάζονται μεταδεδομένα (metadata), και με την χρήση τους ο αλγόριθμος κατηγοριοποιεί τα άρθρα. Έπειτα, θα χρησιμοποιήσει τα μεταδεδομένα, έτσι ώστε να παράξει συστάσεις στους χρήστες της σελίδας, κάνοντας χρήση του φιλτραρίσματος βάσει περιεχομένου. Έτσι, το σύστημα καταφέρνει και ξεπερνάει το πρόβλημα της κρύας εκκίνησης που παρουσιάζει το συνεργατικό φιλτράρισμα, για τους χρήστες αυτούς οι οποίοι δεν προβεί ακόμα σε αξιολόγηση των άρθρων.

Αν τα δεδομένα δεν είναι εμφανή στην ιστοσελίδα, την λύση την δίνει η εξόρυξη δεδομένων (data mining). Μέσω ενός συγκεκριμένου αλγορίθμου, και συγκεκριμένα με την χρήση της βιβλιοθήκης Beautiful Soup, έχουμε την δυνατότητα να εξορύξουμε τις πιο συχνές λέξεις από τα άρθρα που μας ενδιαφέρουν, δημιουργώντας έτσι ένα σύννεφο λέξεων (word cloud). Με αυτόν τον τρόπο δημιουργήσαμε μόνοι μας τα μεταδεδομένα που

μας ενδιαφέρουν, και πλέον μπορούμε να τα εισάγουμε στον αλγόριθμο παραγωγής προτάσεων.

Οι υβριδικές τεχνικές μπορούν να λειτουργήσουν με πολλούς τρόπους όπως:

- Φτιάχνοντας ξεχωριστές προβλέψεις με φιλτράρισμα βάσει περιεχομένου και συνεργατικό φιλτράρισμα, και μετά να τις συνδυάζει.
- Με το να προσθέτει στοιχεία ενός από τις δύο τεχνικές στην άλλη προκειμένου να τις ενισχύσει και να δώσει λύση σε πιθανά προβλήματα.
- Με το να ενώσει τις δύο τεχνικές, φτιάχνοντας ένα καινούριο σύστημα (Tuzhilin, 2005).

Επιπροσθέτως, υπάρχουν πλέον αρκετές έρευνες οι οποίες επιβεβαιώνουν την αποτελεσματικότητα των υβριδικών τρόπων παραγωγής συστάσεων. Σύμφωνα με αυτές τις έρευνες, τα hybrid recommender systems λειτουργούν πολύ καλύτερα από ότι το φιλτράρισμα βάσει περιεχομένου ή το συνεργατικό φιλτράρισμα ατομικά, και προσφέρει πιο ακριβείς συστάσεις στους χρήστες.

Συμπεράσματα

Σε αυτό το κεφάλαιο αναλύσαμε τα τρία διαφορετικά είδη συστημάτων παραγωγής συστάσεων, τα οποία χρησιμοποιούνται σήμερα κατά κόρον:

- Φιλτράρισμα βάσει περιεχομένου (content-based filtering)
- Συνεργατικό φιλτράρισμα (collaborative filtering)
- Υβριδικό φιλτράρισμα (Hybrid filtering)

Η διαφορά μεταξύ των δύο πρώτων τρόπων παραγωγής συστάσεων (content-based filtering και collaborative filtering), έγκειται στην διαφορετική φιλοσοφία με την οποία λαμβάνουν πληροφορίες από τους χρήστες, προκειμένου να τους παρουσιάσουν εξατομικευμένες προτάσεις.

Πιο συγκεκριμένα όπως αναλύθηκε παραπάνω, το content-based filtering προσπαθεί να προβλέψει την συμπεριφορά ή τα χαρακτηριστικά των χρηστών βασισμένο στα χαρακτηριστικά ενός αντικειμένου στο οποίο ο χρήστης έχει αντιδράσει είτε θετικά είτε αρνητικά. Παραδείγματος χάριν, αν στον χρήστη αρέσει ένα άρθρο που βρίσκεται στην κατηγορία πολιτική και έχει ως tag ΄΄πολιτική΄΄, είναι πολύ πιθανό το σύστημα να του προτείνει άρθρα που βρίσκονται στην ίδια κατηγορία και έχουν το αντίστοιχο tag.

Από την άλλη το collaborative filtering δεν χρειάζεται τα στοιχεία ενός αντικειμένου προκειμένου να προβεί σε συστάσεις. Πιο συγκεκριμένα, το συνεργατικό φιλτράρισμα χρησιμοποιεί τις αξιολογήσεις που έχουν δώσει οι χρήστες σε ένα αντικείμενο. Έπειτα, ομαδοποιεί τους χρήστες ανάλογα με το πόσο όμοιες ήταν οι βαθμολογίες τους για συγκεκριμένα αντικείμενα. Έτσι, το σύστημα θεωρεί πως χρήστες με παρόμοιες βαθμολογίες θα έχουν και παρόμοιες προτιμήσεις, οπότε προβαίνει στην σύσταση και των ανάλογων άρθρων.

Ωστόσο και οι δύο αυτές τεχνικές παρουσιάζουν προβλήματα και ελλείψεις. Όσον αφορά στο φιλτράρισμα βάσει περιεχομένου, παρουσιάζει πρόβλημα γενίκευσης και υπάρχει δυσκολία όταν πρόκειται για μεγάλες βάσεις δεδομένων. Το πρόβλημα αυτό δεν το αντιμετωπίζει το συνεργατικό φιλτράρισμα, το οποίο όμως έχει το θέμα της ΄΄κρύας αρχής΄΄

(cold start). Αυτό συμβαίνει με νέους χρήστες οι οποίοι ακόμα δεν έχουν αξιολογήσει κάποιο άρθρο.

Προκειμένου να πάρουμε τα θετικά στοιχεία και από τις δύο αυτές μεθόδους, εξαλείφοντας τα αρνητικά τους, πλέον είναι ευρεία η χρήση του υβριδικού φιλτραρίσματος. Με το hybrid filtering μπορούν να συνδυαστούν το φιλτράρισμα βάσει περιεχομένου και το συνεργατικό φιλτράρισμα, οδηγώντας σε ένα αρτιότερο αποτέλεσμα.

Κεφάλαιο 4

Στόχος εργασίας

Λαμβάνοντας υπόψιν την ραγδαία εξάπλωση της μηχανικής μάθησης, και την σημασία που αυτή έχει για την σύγχρονη δημοσιογραφία, σκοπός αυτής της εργασίας είναι να εξερευνήσει την τεχνολογία των συστημάτων παραγωγής προτάσεων (recommender systems), σε ειδησεογραφικές ιστοσελίδες.

Αναλύοντας χιλιάδες άρθρα, και έχοντας βαθμολογίες χρηστών για μια επιλεγμένη ομάδα από τα άρθρα αυτά, στόχος είναι η δημιουργία ενός αλγορίθμου, ο οποίος θα παράγει εξατομικευμένες συστάσεις άρθρων στους χρήστες.

Στην συγκεκριμένη εργασία, προκειμένου ναι έχουμε ένα ολοκληρωμένο προφίλ των χρηστών, χρησιμοποιήθηκε το υβριδικό μοντέλο παραγωγής συστάσεων (Hybrid Filtering).

Επιπλέον, ακόμα ένας στόχος της εργασίας είναι η δημιουργία ενός online περιβάλλοντος παραγωγής συστάσεων. Για την δημιουργία του συγκεκριμένου περιβάλλοντος, χρησιμοποιήθηκε η πλατφόρμα Dash, ενώ επιλέχθηκε το φιλτράρισμα βάσει περιεχομένου για την παραγωγή συστάσεων άρθρων. Η πλατφόρμα Dash θα αναλυθεί εκτενώς και σε επόμενο κεφάλαιο, ωστόσο περιγραφικά πρόκειται για πλαίσιο βασισμένο στην Python, με απώτερο στόχο την δημιουργία εφαρμογών για οπτικοποίηση δεδομένων (data visualization).

Τέλος, η εργασία έχει ως μελλοντικό στόχο την πρόσκληση χρηστών προκειμένου να αξιολογήσουν την ευχρηστία και την αποτελεσματικότητα της εφαρμογής.

Κεφάλαιο 5

Δημιουργία υβριδικού συστήματος παραγωγής συστάσεων

5.1 Η επιλογή του υβριδικού φιλτραρίσματος

Όπως αναφέρθηκε και στον σκοπό της εργασίας, ο αλγόριθμος παράγωγης συστάσεων άρθρων στους χρήστες ακολουθεί το υβριδικό φιλτράρισμα. Ο λόγος επιλογής του συγκεκριμένου αλγορίθμου είναι οι αξιολογήσεις των χρηστών. Πλέον, οι χρήστες του διαδικτύου έχουν την ευκαιρία να αξιολογούν τα άρθρα που διαβάζουν. Για τον λόγο αυτό θεωρήθηκε πως φιλτράροντας μόνο τα χαρακτηριστικά ενός άρθρου δεν θα είναι αρκετό για την καλύτερη εξατομίκευση στην παραγωγή συστάσεων. Επομένως, η χρήση και του συνεργατικού φιλτραρίσματος, και άρα του υβριδικού τρόπου παραγωγής συστάσεων, αξιολογήθηκε ως η ενδεδειγμένη επιλογή.

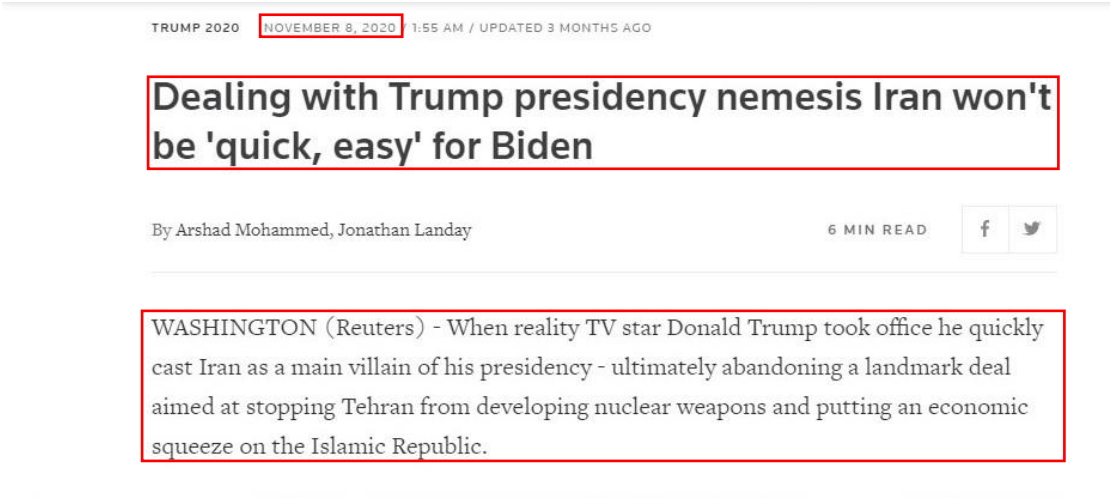
5.1.1 Εξόρυξη δεδομένων

Το πρώτο βήμα για την δημιουργία ενός αλγόριθμου παραγωγής συστάσεων άρθρων είναι η εξόρυξη δεδομένων (data mining). Τα άρθρα που θα εξορυχθούν μέσω αυτής της διαδικασίας θα τα χρησιμοποιήσουμε έπειτα για να εκπαιδεύσουμε τον αλγόριθμο. Ωστόσο, για να μπορέσει να διδαχθεί ικανοποιητικά ο αλγόριθμος πρέπει να τον τροφοδοτήσουμε και με έναν επαρκή αριθμό δεδομένων.

Προκειμένου να εξορυχθούν τα άρθρα χρησιμοποιήθηκε το εργαλείο Web Scrapper, το οποίο είναι διαθέσιμο στον περιηγητή Chrome. Το Web Scrapper έχει την ικανότητα να εξορύσσει μεγάλο αριθμό δεδομένων σε σχετικά σύντομο χρονικό διάστημα. Πιο συγκεκριμένα το Web Scrapper είναι βασισμένο σε μια σειρά από επιλογείς (selectors), οι οποίοι δίνουν εντολή στο εργαλείο αυτό να σαρώσει την ιστοσελίδα που επιθυμούμε και να εξάγει τα δεδομένα που εμείς επιθυμούμε. Αφού έχουμε εξορύξει όλα τα δεδομένα τα οποία θέλουμε, έπειτα το Web Scrapper τα αποθηκεύει ως αρχεία CSV, τα οποία μετά μπορούν να χρησιμοποιηθούν στον αλγόριθμο.

Στην συγκεκριμένη εργασία χρησιμοποιήσαμε το Web Scrapper έτσι ώστε να εξάγουμε πληροφορίες από 1955 άρθρα από την ιστοσελίδα του πρακτορείου Reuters: https://www.cnn.gr/. Από τα συγκεκριμένα άρθρα στόχος ήταν να εξορυχθούν οι παρακάτω πληροφορίες για τα χαρακτηριστικά των άρθρων:

- Τίτλος.
- Ημερομηνία.
- Αρχική παράγραφος.

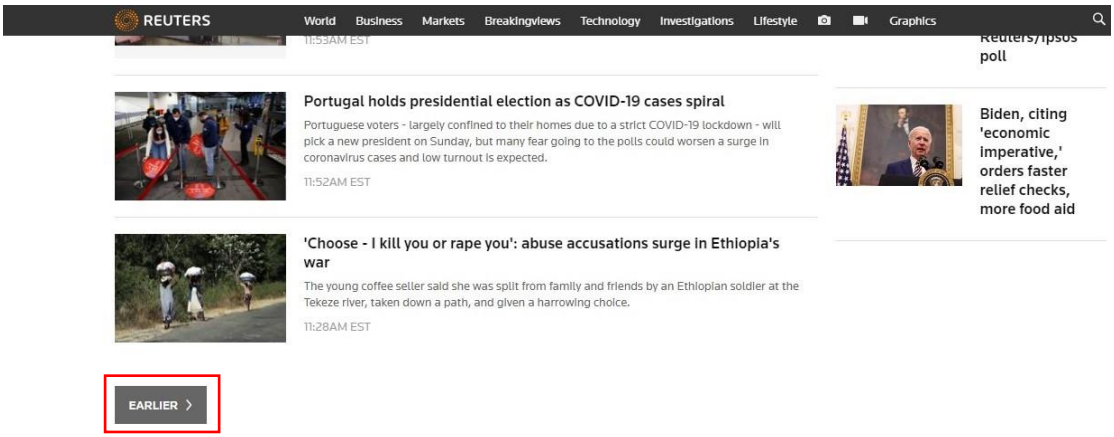


Εικόνα 6: Παράδειγμα άρθρου από το πρακτορείο Reuters, με υπογραμμισμένα τα χαρακτηριστικά του άρθρου που χρησιμοποιήθηκαν.

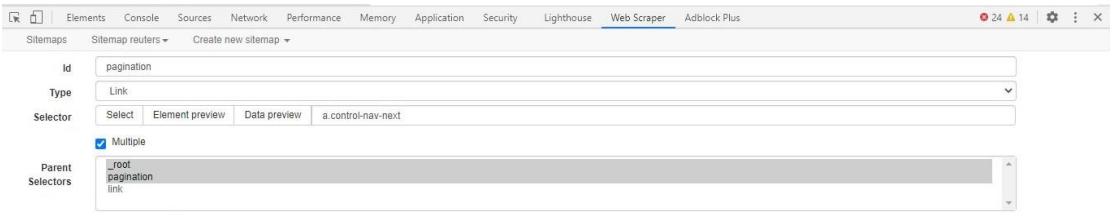
Αναφορικά με τον τρόπο με τον οποίο χρησιμοποιήθηκε το Web Scrapper για να γίνει η εξόρυξη δεδομένων, η διαδικασία που ακολουθήθηκε είναι η εξής: Αρχικά δίνουμε εντολή στο εργαλείο να σαρώσει την ιστοσελίδα του Reuters https://www.reuters.com/news/archive/worldNews?view=page.

Εικόνα 7: Εντολή στον Web Scrapper να σαρώσει την επιθυμητή σελίδα.

Έπειτα, προκειμένου το Web Scrapper να σαρώσει όλες τις υποσελίδες του πρακτορείου Reuters, δημιουργήσαμε την εντολή pagination (σελιδοποίηση), στην οποία θέσαμε ως γονείς επιλογείς (parent selectors) τόσο την αρχική διεύθυνση, όσο και τον ίδιο της το εαυτό. Αυτό έγινε προκειμένου το Web Scrapper να αλλάζει συνεχώς σελίδα όποτε βρίσκει το κουμπί «Earlier» και να μην μείνει μόνο στην πρώτη σελίδα.

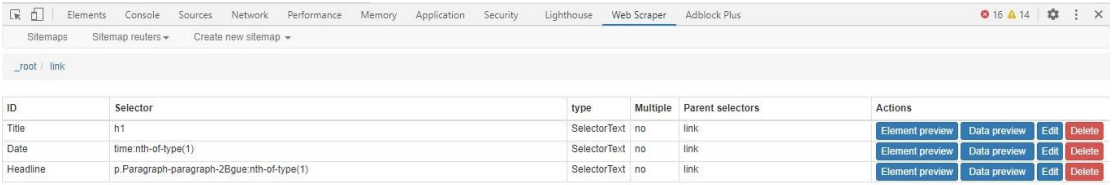


Εικόνα 8: Το κουμπί «Earlier» το οποίο οδηγεί σε παλαιότερα άρθρα.



Εικόνα 9: Εντολή pagination έτσι ώστε το Web Scrapper να σαρώσει άρθρα από περισσότερες από μια υποσελίδες του Reuters.

Το επόμενο βήμα είναι η δημιουργία της εντολής link. Αυτή η εντολή περιέχει τρία βήματα, από τα οποία και θα πάρουμε τα στοιχεία των άρθρων που μας ενδιαφέρουν. Σε αυτή την εντολή ουσιαστικά καθοδηγούμε το Web Scrapper στο να εξορύξει την ημερομηνία του άρθρου, τον τίτλο του και την αρχική παράγραφο. Σε κάθε ένα από αυτά τα τρία βήματα θέτουμε ως γονέα την διεύθυνση κάθε άρθρου. Με τον τρόπο αυτό, το εργαλείο γνωρίζει πως όταν ανοίγει ένα νέο άρθρο, θα πρέπει να μας δώσει τις τρεις πληροφορίες που του ζητήσαμε.



Εικόνα 10: Εντολή ώστε το Web Scrapper να εξορύξει τα ζητούμενα χαρακτηριστικά του άρθρου.

Έπειτα και από την τελευταία αυτή εντολή το Web Scrapper πλέον ψάχνει όλες τις σελίδες μέσα στο Reuters και παράγει τα δεδομένα που επιθυμούμε.

5.1.2 Φιλτράρισμα βάσει περιεχομένου

Αφού εξορύξαμε τα δεδομένα που επιθυμούσαμε, πλέον μπορούμε να τα εισαγάγουμε στον κώδικά μας. Το πρώτο βήμα στην δημιουργία του συστήματος παραγωγής συστάσεων είναι το φιλτράρισμα βάσει περιεχομένου. Όπως έχει αναλυθεί και παραπάνω, σε αυτό το βήμα θα δοθεί έμφαση στα χαρακτηριστικά των άρθρων. Πιο συγκεκριμένα, θα αναλύσουμε τα στοιχεία των άρθρων που μας.

Αρχικά εισαγάγουμε τις απαραίτητες βιβλιοθήκες, έτσι ώστε να μπορούμε να επεξεργαστούμε τα δεδομένα που αποκτήσαμε. Πιο συγκεκριμένα, χρησιμοποιήθηκαν οι εξής βιβλιοθήκες:

- **Pandas**: Η συγκεκριμένη βιβλιοθήκη μας βοηθά να επεξεργαστούμε τα δεδομένα που έχουμε και να πραγματοποιήσουμε αναλύσεις πάνω σε αυτά.
- **Από την βιβλιοθήκη Rake_nltk πήραμε το Rake**: Αυτή η βιβλιοθήκη είναι συντομογραφία για: Rapid Automatic Keyword Extraction. Μέσω αυτής μπορούμε να καθορίσουμε λέξεις κλειδιά μέσα στα άρθρα.
- **Numpy**: Με αυτή τη βιβλιοθήκη μας δίνεται η δυνατότητα να πραγματοποιήσουμε πολύπλοκες μαθηματικές λειτουργίες με τα δεδομένα μας.
- **Από την βιβλιοθήκη Sklearn.metrics.pairwise πήραμε το cosine_similarity**: Με την συγκεκριμένη βιβλιοθήκη θα μπορέσουμε να μετρήσουμε την ομοιότητα μεταξύ των δεδομένων μας.
- **Από την βιβλιοθήκη Sklearn.feature_extraction.text πήραμε το CountVectorizer:** Αυτή η βιβλιοθήκη θα μας βοηθήσει στο να μετατρέψουμε το κείμενα σε

μετρήσιμο διάνυσμα, έτσι ώστε να μετρήσουμε με τι συχνότητα εμφανίζονται συγκεκριμένες λέξεις.

```
[ ]  import pandas as pd
     !pip install rake-nltk
     from rake_nltk import Rake
     import numpy as np
     from sklearn.metrics.pairwise import cosine_similarity
     from sklearn.feature_extraction.text import CountVectorizer
```

Εικόνα 11: Εισαγωγή κατάλληλων βιβλιοθηκών.

Αφού εισάγουμε με επιτυχία τις βιβλιοθήκες που επιθυμούμε, έπειτα εισάγουμε το αρχείο CSV με τα ζητούμενα στοιχεία από τα άρθρα, το οποίο δημιούργησε το Web Scrapper, μετά την εξόρυξη δεδομένων.

```
[ ]  df_reuters.iloc[0]
```

```
web-scraper-order                                    1606153665-2067
web-scraper-start-url    https://www.reuters.com/news/archive/worldNews...
pagination                                                    Earlier
pagination-href          https://www.reuters.com/news/archive/worldNews...
link                     Armenia, Azerbaijan agree to defuse Nagorno-Ka...
link-href                https://www.reuters.com/article/us-armenia-aze...
Title                    Armenia, Azerbaijan agree to defuse Nagorno-Ka...
Date                                                       October 30, 2020
Headline                 GENEVA (Reuters) - The foreign ministers of Ar...
Name: 0, dtype: object
```

Εικόνα 12: Εισαγωγή αρχείου CSV με τα εξορυγμένα δεδομένα.

Ωστόσο, επειδή η συγκεκριμένη μορφή δεν μας βοηθάει να κάνουμε αναλύσεις, θα χωρίσουμε τα άρθρα σε ένα πίνακα, ανάλογα με τα στοιχεία που επιθυμούμε. Στην συγκεκριμένη περίπτωση μας ενδιαφέρει ο τίτλος του άρθρου και η αρχική του παράγραφος:

```
df_reuters = df_reuters[['Title','Headline']]
df_reuters.head(5)
```

	Title	Headline
0	Armenia, Azerbaijan agree to defuse Nagorno-Ka...	GENEVA (Reuters) - The foreign ministers of Ar...
1	Dealing with Trump presidency nemesis Iran won...	WASHINGTON (Reuters) - When reality TV star Do...
2	Exclusive: Russian hackers targeted California...	WASHINGTON (Reuters) - The group of Russian ha...
3	Spain to send more police to Senegal to curb i...	DAKAR (Reuters) - Spain will increase its poli...
4	Canada expects six million COVID-19 vaccine do...	(Reuters) - Canada expects to receive six mill...

Εικόνα 13: Εισαγωγή άρθρων σε πίνακα.

Αφού δημιουργήσουμε τον πίνακα με τα άρθρα, θα δημιουργήσουμε μια νέα στήλη, η οποία θα ονομάζεται Key_Words. Όπως υποδηλώνει και το όνομά της, αυτή η στήλη θα περιλαμβάνει λέξεις κλειδιά, από όποια στήλη δηλώσουμε. Εν προκειμένω την στήλη του Τίτλου. Για να πραγματοποιήσουμε την συγκεκριμένη ενέργεια, θα χρησιμοποιήσουμε την βιβλιοθήκη Rake.

```
[ ] df_reuters['Key_words'] = ""

    for index, row in df_reuters.iterrows():
        headline = row['Headline']

        r = Rake()

        r.extract_keywords_from_text(headline)

        key_words_dict_scores = r.get_word_degrees()

        row['Key_words'] = list(key_words_dict_scores.keys())

    df_reuters.drop(columns = ['Headline'], inplace = True)
```

```
[ ] df_reuters.set_index('Title', inplace = True)
    df_reuters.head(5)
```

Title	Key_words
Armenia, Azerbaijan agree to defuse Nagorno-Karabakh conflict	[reuters, friday, major, powers, said, urgent,...
Dealing with Trump presidency nemesis Iran won't be 'quick, easy' for Biden	[main, villain, reuters, presidency, stopping,...
Exclusive: Russian hackers targeted California, Indiana Democratic parties	[group, reuters, 2016, u, california, new, yor...
Spain to send more police to Senegal to curb illegal migration, says foreign minister	[illegal, migration, reuters, dakar, foreign, ...

Εικόνα 14: Εύρεση λέξεων κλειδιών στους τίτλους των άρθρων.

Στη συνέχει και αφού έχουμε καταλήξει στις λέξεις κλειδιά των τίτλων, θα χρησιμοποιήσουμε την βιβλιοθήκη sklearn.feature_extraction.text, και από αυτή την ενέργεια του CountVectorizer. Με αυτόν τον τρόπο θα δημιουργήσουμε μια σειρά από τους τίτλους, έτσι ώστε να βρει ο αλγόριθμος ποιες λέξεις εμφανίζονται συχνότερα στους τίτλους.

```
[ ]
    count = CountVectorizer()
    count_matrix = count.fit_transform(df_reuters['key_words_string'])


    indices = pd.Series(df_reuters.index)
    indices[:5]
```

```
0    Armenia, Azerbaijan agree to defuse Nagorno-Ka...
1    Dealing with Trump presidency nemesis Iran won...
2    Exclusive: Russian hackers targeted California...
3    Spain to send more police to Senegal to curb i...
4    Canada expects six million COVID-19 vaccine do...
Name: Title, dtype: object
```

Εικόνα 15: Χρήση του CountVectorizer για την εύρεση των πιο συχνών λέξεων στα άρθρα

Το προηγούμενο βήμα ήταν πολύ σημαντικό, καθώς αφού πλέον έχουμε βρει την συχνότητα συγκεκριμένων λέξεων στους τίτλους των άρθρων, μπορούμε τώρα να βρούμε ποια άρθρα είναι όμοια μεταξύ τους. Αυτό είναι το τελευταίο και κρισιμότερο βήμα, προκειμένου ο αλγόριθμος να φιλτράρει ανά περιεχόμενο τα δεδομένα που του έχουμε δώσει. Αφού πραγματοποιήσει το συγκεκριμένο βήμα, θα είναι έτοιμος να προβεί και στις ανάλογες προτάσεις άρθρων, ανάλογα με την ομοιότητα των τίτλων τους.

```
[ ]  cosine_sim = cosine_similarity(count_matrix, count_matrix)
     cosine_sim

     array([[1.        , 0.03806935, 0.03872015, ..., 0.13043478, 0.04089304,
             0.09567297],
            [0.03806935, 1.        , 0.06780635, ..., 0.03806935, 0.03580574,
             0.04188539],
            [0.03872015, 0.06780635, 1.        , ..., 0.03872015, 0.03641785,
             0.04260143],
            ...,
            [0.13043478, 0.03806935, 0.03872015, ..., 1.        , 0.08178608,
             0.09567297],
            [0.04089304, 0.03580574, 0.03641785, ..., 0.08178608, 1.        ,
             0.04499213],
            [0.09567297, 0.04188539, 0.04260143, ..., 0.09567297, 0.04499213,
             1.        ]])
```

Εικόνα 16: Χρήση του Cosine Similarity για την εύρεση ομοιοτήτων μεταξύ των τίτλων.

Προκειμένου ο αλγόριθμος να προχωρήσει σε συστάσεις άρθρων ανάλογα με την ομοιότητα τους θα του δώσουμε εντολή να σαρώσει όλους τους τίτλους των άρθρων που έχουμε εισάγει και να δημιουργήσει ένα σκορ ομοιότητας. Έπειτα, θα θέσουμε ως βάση έναν συγκεκριμένο τίτλο άρθρου, εν προκειμένω αυτός ο τίτλος είναι ο: ‘’Armenia, Azerbaijan agree to defuse Nagorno-Karabakh conflict’’, και δίνουμε εντολή στον αλγόριθμο να μας παρουσιάσει τα άρθρα εκείνα που έχουν το πλησιέστερο σκορ ομοιότητας με τον συγκεκριμένο τίτλο:

0	Russia says discussing U.N. presence in Nagorn...	0.383065
0	Azerbaijan's president says deal is signed to ...	0.354005
0	Leader of Nagorno-Karabakh says ceasefire with...	0.344031
0	Putin calls for Turkish involvement in Nagorno...	0.327144
0	Turkey says in talks on how to monitor Karabak...	0.321029
0	Russia reinforces border guards in Armenia aft...	0.318511

Εικόνα 17: Παραγωγή συστάσεων άρθρων χρησιμοποιώντας το φιλτράρισμα ανά περιεχόμενο.

Έτσι βλέπουμε πως αν κάποιος διαβάσει το άρθρο με τίτλο ‘’Armenia, Azerbaijan agree to defuse Nagorno-Karabakh conflict’, ο αλγόριθμος θα αναλύσει τα στοιχεία του άρθρου, έτσι ώστε να βρει ομοιότητες με τα υπόλοιπα άρθρα, και στο τέλος θα του παρουσιάσει τα άρθρα που φαίνονται στο γράφημα 17.

5.1.3 Συνεργατικό φιλτράρισμα

Ωστόσο, το φιλτράρισμα βάσει περιεχομένου δεν καλύπτει όλες τις πτυχές που επιθυμούμε, έτσι ώστε να παραχθούν καλύτερες εξατομικευμένες προτάσεις για κάθε χρήστη. Για τον λόγο αυτό, εξελίσσουμε τον αλγόριθμο ώστε να συμπεριλαμβάνει και το συνεργατικό φιλτράρισμα.

Όπως έχει αναφερθεί και στα προηγούμενα κεφάλαια, το συνεργατικό φιλτράρισμα χρειάζεται κάποιου είδους αξιολόγηση των άρθρων από τον χρήστη (π.χ σε μια κλίμακα από 1-5). Αυτή τη μέθοδο εφαρμόσαμε και στην συγκεκριμένη εργασία. Πιο συγκεκριμένα, πραγματοποιήθηκε μια διαδικτυακή συνέντευξη, μέσω της οποίας ζητήθηκε από 10 άτομα, να βαθμολογήσουν 50 άρθρα, από αυτά που είχαμε εξορύξει από την ιστοσελίδα του πρακτορείου Reuters.

	A	B	C
1	UserID	ArticleID	Ratings
2	1	Armenia, Azerbaijan agree to defuse Nagorno-Karabakh conflict	2
3	1	Dealing with Trump presidency nemesis Iran won't be 'quick, easy' for Biden	1
4	1	Exclusive: Russian hackers targeted California, Indiana Democratic parties	3
5	1	Spain to send more police to Senegal to curb illegal migration, says foreign minister	4

Εικόνα 18: Βαθμολογία άρθρων από τους χρήστες.

Αφότου αποκτήσουμε τις αξιολογήσεις μας για τα άρθρα, μπορούμε πλέον να προχωρήσουμε στην δημιουργία του κώδικα με βάση το συνεργατικό φιλτράρισμα. Πρώτο βήμα είναι η εισαγωγή μιας απαραίτητης βιβλιοθήκης: της Pandas. Η βιβλιοθήκη αυτή θα μας δώσει την δυνατότητα να πραγματοποιήσουμε αναγκαίες αναλύσεις για τα δεδομένα που έχουμε, έτσι ώστε να πετύχουμε μια σωστή παραγωγή προτάσεων άρθρων στους χρήστες. Έπειτα προσθέτουμε και το αρχείο το οποίο περιέχει τις βαθμολογίες.

```
[ ]  df_ratings.head(5)
```

	UserID	ArticleID	Ratings
0	1	Armenia, Azerbaijan agree to defuse Nagorno-Ka...	2
1	1	Dealing with Trump presidency nemesis Iran won...	1
2	1	Exclusive: Russian hackers targeted California...	3
3	1	Spain to send more police to Senegal to curb i...	4
4	1	Canada expects six million COVID-19 vaccine do...	5

Εικόνα 19: Εισαγωγή του αρχείου με τις βαθμολογίες άρθρων από τους χρήστες.

Στη συνέχεια, το επόμενο βήμα είναι να καθοριστεί ο συντελεστής προσδιορισμού r2. Ο συγκεκριμένος συντελεστής (R-squared) είναι ένα στατιστικό μέτρο το οποίο αναδεικνύει το ποσοστό της διακύμανσης για μια εξαρτημένη μεταβλητή η οποία εξηγείται από μια ανεξάρτητη μεταβλητή μέσα σε ένα μοντέλο παλινδρόμησης. Επιπλέον, πολύ σημαντική για την εύρεση του R2 είναι η χρήση της βιβλιοθήκης Numpy, έτσι ώστε να πραγματοποιήσουμε πολύπλοκες μαθηματικές λειτουργίες για τα δεδομένα μας.

```
[ ]  import numpy as np
     selfjoined['r1r2'] = selfjoined['Ratings_x']*selfjoined['Ratings_y']
     selfjoined['r1square'] = np.square(selfjoined['Ratings_x'])
     selfjoined['r2square'] = np.square(selfjoined['Ratings_y'])
```

```
[ ]  selfjoined.head(5)
```

	UserID	ArticleID_x	count_x	Ratings_x	ArticleID_y	count_y	Ratings_y	r1r2	r1square	r2square
1	3	Polish cardinal accused of sexual abuse dies a...	2	3	Syrian satirist questioned over 'Macron' flogg...	1	2	6	9	4
3	3	Polish cardinal accused of sexual abuse dies a...	2	3	Pompeo says Europe, U.S. need to work together...	2	1	3	9	1
4	3	Polish cardinal accused of sexual abuse dies a...	2	3	Portugal's president ponders COVID-19 state of...	1	5	15	9	25
10	3	No more bullying': fresh start to U.S.-Mexico ...	2	2	Polish cardinal accused of sexual abuse dies a...	2	3	6	4	9
11	3	No more bullying': fresh start to U.S.-Mexico ...	2	2	Syrian satirist questioned over 'Macron' flogg...	1	2	4	4	4

Εικόνα 20: Καθορισμός συντελεστής προσδιορισμού R2.

Έπειτα, θα θέσουμε ως βάση έναν συγκεκριμένο τίτλο, έτσι ώστε να ελέγξουμε την συσχέτιση του με τους υπόλοιπους τίτλους άρθρων.

```
[ ] aggdata=selfjoined.groupby(['ArticleID_x','ArticleID_y'])['Ratings_x','Ratings_y','r1r2','r1square','r2square','count_x','count_y'].sum()
    aggdata.head(5)
```

```
/usr/local/lib/python3.6/dist-packages/ipykernel_launcher.py:1: FutureWarning: Indexing with multiple keys (implicitly converted to a tuple of keys) will be depre
  """Entry point for launching an IPython kernel.
```

ArticleID_x	ArticleID_y	Ratings_x	Ratings_y	r1r2	r1square	r2square	count_x	count_y
Australia alleges military carried out unlawful killings in Afghanistan	COVID-19 cases soar in Brazil's largest indigenous reservation	5	5	8	17	17	6	6
	Factbox: As mediation calls mount, who has leverage in Ethiopia?	4	2	8	16	4	3	2
	France reports 271 new COVID-19 deaths	4	4	16	16	16	3	2
	Italy has 48 hours to approve new COVID-19 restrictions: Health Minister	4	4	16	16	16	3	3

Εικόνα 21: Καθορισμός συντελεστής προσδιορισμού R2 συγκεκριμένου τίτλου.

Αφού βρούμε τον συντελεστή προσδιορισμού, ακολουθεί ένα πολύ σημαντικό βήμα, το οποίο είναι η συσχέτιση των άρθρων μεταξύ τους. Ο συντελεστής συσχέτισης (correlation coefficient) ανακαλύφθηκε από τον Καρλ Πίρσον το 1896. Ο συντελεστής συσχέτισης χρησιμοποιείται ευρέως στην στατιστική μέχρι και σήμερα (Ratner, 2009). Σύμφωνα με τον συντελεστή αυτό, όσο πιο κοντά βρίσκονται δύο δεδομένα στο 1, τόσο πιο όμοια θεωρούνται.

```
[ ] aggdata['corelation'] = (n*aggdata['r1r2'] - aggdata['Ratings_x']*aggdata['Ratings_y']) / \
    np.sqrt((n * aggdata['r1square']  - np.square(aggdata['Ratings_x']))*(n * aggdata['r2square']  - np.square(aggdata['Ratings_y'])))
```

```
[ ] aggdata['corelation'] = aggdata['corelation']*n/(n+50)
```

```
[ ] aggdata.reset_index(inplace=True)
    aggdata.head(5)
```

	ArticleID_x	ArticleID_y	Ratings_x	Ratings_y	r1r2	r1square	r2square	count_x	count_y	corelation
0	Australia alleges military carried out unlawfu...	COVID-19 cases soar in Brazil's largest indige...	5	5	8	17	17	6	6	0.309532
1	Australia alleges military carried out unlawfu...	Factbox: As mediation calls mount, who has lev...	4	2	8	16	4	3	2	0.668874
2	Australia alleges military carried out unlawfu...	France reports 271 new COVID-19 deaths	4	4	16	16	16	3	2	0.668874
3	Australia alleges military carried out unlawfu...	Italy has 48 hours to approve new COVID-19 res...	4	4	16	16	16	3	3	0.668874
4	Australia alleges military carried out unlawfu...	Pakistan minister deletes tweet containing Mac...	4	3	12	16	9	3	1	0.668874

Εικόνα 22: Καθορισμός συντελεστή συσχέτισης μεταξύ τίτλων των άρθρων.

Η εύρεση του συγκεκριμένου συντελεστή βοηθά στο επόμενο βήμα τον αλγόριθμο να καταλήξει σε συστάσεις άρθρων, με βάση το συνεργατικό φιλτράρισμα. Για να το επιτύχουμε αυτό θέτουμε ως μέτρο σύγκρισης έναν συγκεκριμένο τίτλο άρθρου. Τώρα θα χρησιμοποιήσουμε ως βάση τον τίτλο ‘’England will need five days of lockdown for each day relaxed at Christmas: adviser’’.

```
[ ] def recommendation(Title):
        recommended_movies = []
        data =aggdata2[aggdata2['Title_x']==Title]
        data = data.sort_values(by='corelation',ascending=False)
        return data
```

```
collabaritive=recommendation('England will need five days of lockdown for each day relaxed at Christmas: adviser')[['Title_y','corelation']]
collabaritive['Title'] = collabaritive['Title_y'].str.split(' \(').str.get(0)
collabaritive.drop('Title_y',axis=1,inplace=True)
collabaritive.head(5)
```

	corelation	Title
40	0.668874	Taiwan hopes for close U.S. cooperation in cal...
49	0.668874	Thank you Tegel: Berliners bid emotional farew...
71	0.668874	Exclusive: Russian hackers targeted California...

Εικόνα 23: Παραγωγή συστάσεων άρθρων με βάση το συνεργατικό φιλτράρισμα.

Όπως βλέπουμε, αφού δώσουμε την εντολή στον αλγόριθμο να βρει παρόμοια άρθρα με το ‘’England will need five days of lockdown for each day relaxed at Christmas: adviser’’, θα συνέστηνε τα εξής:

- Taiwan hopes for close U.S. cooperation in call with Biden adviser
- Thank you Tegel: Berliners bid emotional farewell to Cold War airport
- Exclusive: Russian hackers targeted California, Indiana Democratic parties
- Japan opens airport coronavirus test lab for departing travellers

5.1.4 Υβριδικό φιλτράρισμα

Εφόσον ο αλγόριθμος κατάφερε με επιτυχία να παράξει συστάσεις άρθρων τόσο με βάση το φιλτράρισμα βάσει περιεχομένου όσο και με το συνεργατικό φιλτράρισμα, πλέον μπορούμε να συνδυάσουμε αυτά τα δύο αποτελέσματα και να δημιουργήσουμε ένα υβριδικό σύστημα. Το υβριδικό σύστημα παραγωγής συστάσεων θα είναι ικανό να αποκτήσει μια γενίκευση, το οποίο αποτελεί το κύριο πρόβλημα του content-based filtering, αλλά υπάρχει και μια άμυνα απέναντι στην έλλειψη αξιολογήσεων από τους χρήστες – το κύριο πρόβλημα του collaborative filtering – αφού ο αλγόριθμος μπορεί να προχωρήσει σε συστάσεις στηριζόμενος σε στοιχεία των άρθρων, ελλείψει αξιολογήσεων.

Το πρώτο βήμα για την δημιουργία ενός αλγορίθμου που θα πραγματοποιεί προτάσεις με βάση το υβριδικό φιλτράρισμα, είναι να εισάγουμε τις προτάσεις που είχαν παραχθεί μέσω του φιλτραρίσματος βάσει περιεχομένου. Οι συγκεκριμένες προτάσεις είχαν αποθηκευτεί ως αρχείο CSV, και μπορούν πολύ εύκολα να ενσωματωθούν στον κώδικα.

Αφού εισάγουμε τις προτάσεις του content-based filtering θα τις ενώσουμε με τις προτάσεις του συνεργατικού φιλτραρίσματος, χρησιμοποιώντας την εντολή ‘’merge’’.

```
[ ]  hybrid=contentrating.merge(collabaritive,left_on='0',right_on='Title')
     hybrid=hybrid[['Title','1','corelation']]
     hybrid['wcorelation'] = (hybrid['1'] + hybrid['corelation'])/2
```

```
[ ]  hybrid.head(5)
```

	Title	1	corelation	wcorelation
0	Japan opens airport coronavirus test lab for d...	0.093250	0.668874	0.381062
1	Thank you Tegel: Berliners bid emotional farew...	0.045502	0.668874	0.357188
2	Taiwan hopes for close U.S. cooperation in cal...	0.040129	0.668874	0.354501
3	Exclusive: Russian hackers targeted California...	0.038720	0.668874	0.353797

Εικόνα 24: Ένωση συστάσεων άρθρων από το φιλτράρισμα βάσει περιεχομένου και από το συνεργατικό φιλτράρισμα, με την εντολή merge.

Συγκρίνοντας τις προτάσεις του συνεργατικού φιλτραρίσματος για το άρθρο με τίτλο ‘’England will need five days of lockdown for each day relaxed at Christmas: adviser’’, με τις προτάσεις του υβριδικού φιλτραρίσματος, παρατηρούμε ομοιότητες και διαφορές.

Όπως φαίνεται από τα γραφήματα 23 και 24, τα άρθρα που προτείνουν τα δύο διαφορετικά είδη φιλτραρίσματος είναι τα ίδια. Πιο συγκεκριμένα είναι τα:

- Taiwan hopes for close U.S. cooperation in call with Biden adviser
- Thank you Tegel: Berliners bid emotional farewell to Cold War airport
- Exclusive: Russian hackers targeted California, Indiana Democratic parties
- Japan opens airport coronavirus test lab for departing travellers

Ωστόσο, ενώ το συνεργατικό φιλτράρισμα κατέληξε στην προαναφερθείσα σειρά των συστημένων άρθρων, κατά σειρά μεγαλύτερης συσχέτισης, στο υβριδικό φιλτράρισμα παρατηρείται μια διαφορά. Ενώ τα προτεινόμενα άρθρα είναι τα ίδια, το υβριδικό φιλτράρισμα, έχοντας λάβει υπόψιν του και τις συστάσεις του φιλτραρίσματος βάσει περιεχομένου, προτείνει τα ίδια άρθρα, αλλά με διαφορετική σειρά. Πιο συγκεκριμένα, το υβριδικό φιλτράρισμα θα πρότεινε την παρακάτω σειρά των άρθρων, για τον τίτλο ‘’England will need five days of lockdown for each day relaxed at Christmas: adviser’’:

- Japan opens airport coronavirus test lab for departing travellers
- Thank you Tegel: Berliners bid emotional farewell to Cold War airport
- Taiwan hopes for close U.S. cooperation in call with Biden adviser
- Exclusive: Russian hackers targeted California, Indiana Democratic parties

Κεφάλαιο 6

Περιβάλλον εξατομίκευσης

6.1. Τι είναι το Dash

Όπως αναφέρθηκε και παραπάνω, σκοπός της εργασίας είναι να μελετήσει κατά πόσο είναι εφικτό μια ιστοσελίδα ενημερωτικού ενδιαφέροντος να βελτιώσει την εμπειρία των χρηστών τους μέσω της παραγωγής εξατομικευμένου περιεχομένου. Για να συμβεί αυτό, ειδικοί αλγόριθμοι μηχανικής μάθησης, αναλαμβάνουν να μελετήσουν στοιχεία των άρθρων, τις προτιμήσεις των χρηστών ή και τα δύο, προκειμένου να παράξουν εξατομικευμένες προτάσεις άρθρων. Επιπλέον, η συγκεκριμένη εργασία είχε επίσης ως σκοπό την οπτική παρουσίαση ενός τέτοιου ιστοχώρου. Πιο συγκεκριμένα, στόχος ήταν να δημιουργηθεί ένα περιβάλλον, μέσα στο οποίο θα εντάσσονταν τα χιλιάδες άρθρα που είχαν εξορυχθεί. Έπειτα, οι χρήστες θα επέλεγαν ένα οποιοδήποτε άρθρο, έτσι ώστε να παραχθούν και οι αντίστοιχες προτάσεις.

Για να επιτευχθεί ο συγκεκριμένος σκοπός, επιλέξαμε να χρησιμοποιήσουμε το Dash, ως το περιβάλλον στο οποίο θα παρουσιαστούν τα αποτελέσματα του συστήματος παραγωγής συστάσεων. Το Dash είναι μια βιβλιοθήκη η οποία στοχεύει στην αλληλεπίδραση των χρηστών. Όσοι χρησιμοποιούν την γλώσσα Python για να αναλύσουν δεδομένα, να τα οπτικοποιήσουν και να φτιάξουν ειδικά μοντέλα αλγορίθμων, θα βρουν το Dash εξαιρετικά χρήσιμο.

Οι εφαρμογές που δημιουργούνται στο Dash, τρέχουν με το Flask, χρησιμοποιώντας αιτήσεις HTTP. Το Flask είναι ένα διαδικτυακό περιβάλλον γραμμένο σε Python. Το Flask δεν χρειάζεται συγκεκριμένα εργαλεία ή βιβλιοθήκες, όπως επίσης δεν έχει βιβλιοθήκες τρίτων οι οποίες να παρέχουν κοινά στοιχεία.

Επιπλέον, οι εφαρμογές που γράφονται με κώδικα Dash μπορούν να είναι διαδραστικές, ενώ καθιστούν εύκολη την δημιουργία πολύπλοκων εφαρμογών, οι οποίες διαθέτουν πολλά διαδραστικά στοιχεία. Επίσης, όλα τα αισθητικά στοιχεία των εφαρμογών Dash, όπως το μέγεθος, οι τοποθετήσεις τους, και τα χρώματα, μπορούν να μεταβληθούν.

Επίσης, οι εφαρμογές που γράφονται στο Dash δημοσιεύονται στο διαδίκτυο. Παρόλα αυτά, δεν χρειάζονται κώδικα Javascript ή HTML. Αυτό συμβαίνει επειδή, όπως προείμαμε, το Dash παρέχει ένα περιβάλλον Python, το οποίο είναι πλούσιο σε διαδραστικά, διαδικτυακά στοιχεία. Για παράδειγμα, όταν αλλάζει ένα στοιχεία στην Dash εφαρμογή – στην συγκεκριμένη εργασία αυτό αφορά την επιλογή άρθρων από την λίστα που έχει δημιουργηθεί – ο κώδικας που έχει γραφτεί στο Dash θα μας δώσει το αποτέλεσμα αυτής της αλλαγής – πιο συγκεκριμένα τις προτάσεις για το άρθρο που έχει επιλεχθεί. -

6.2 Δημιουργία της εφαρμογής στο Dash

Λαμβάνοντας υπόψιν όλα τα παραπάνω, θεωρήθηκε πως το περιβάλλον του Dash θα ήταν ιδανικό για την δημιουργία της εφαρμογής παραγωγής συστάσεων άρθρων. Αρχικά, πρέπει να εισαχθούν όλες οι απαραίτητες βιβλιοθήκες, έτσι ώστε να τρέξει ο κώδικας. Όπως φαίνεται, εγκαθιστούμε το Dash μέσω της βιβλιοθήκης Dash. Επιπροσθέτως, αντί να να γράψουμε κώδικα σε HTML, χρησιμοποιήθηκε η βιβλιοθήκη dash-html για να δημιουργήσει αντίστοιχο περιβάλλον, χρησιμοποιώντας όμως την γλώσσα Python.

```
import dash
import dash_html_components as html
import dash_core_components as dcc
import dash_bootstrap_components as dbc
import pandas as pd
import recomms
from dash.dependencies import Input, Output
import ast
from scipy import stats
from ast import literal_eval
from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer
from sklearn.metrics.pairwise import linear_kernel, cosine_similarity
```

Εικόνα 25: Εισαγωγή απαραίτητων βιβλιοθηκών για την λειτουργία της εφαρμογής.

Στην συνέχεια αφού εισάγουμε το αρχείο με τα άρθρα που εξορύχθησαν, θα δημιουργήσουμε την εφαρμογή, όσο και τον server στον οποίο θα τρέξει.

```
df = pd.read_csv('reuters.csv', index_col=[0])

app = dash.Dash(external_stylesheets=[dbc.themes.BOOTSTRAP])

server = app.server
```

Εικόνα 26: Δημιουργία εφαρμογής και server.

Σε αυτό το σημείο, θα πρέπει να αναφερθεί πως όλες οι εφαρμογές Dash αποτελούνται από δύο μέρη: το layout, το οποίο αφορά στην μορφή της εφαρμογής, και τα call-backs, τα οποία αφορούν την λειτουργία της, και πιο συγκεκριμένα την αλληλεπίδραση με τους χρήστες για την παραγωγή αποτελεσμάτων.

Αρχικά, θα ασχοληθούμε με την μορφής της εφαρμογής. Πρώτον, στόχος είναι η δημιουργία ενός πίνακα, ο οποίος θα περιέχει όλα τα άρθρα που εξορύξαμε. Θα δημιουργήσουμε τον συγκεκριμένο πίνακα, δίνοντας εντολή στο Dash να διατρέξει το αρχείο με άρθρα, και να δημιουργήσει τον πίνακα με βάση τον τίτλο των άρθρων.

```
def Table(df):
    #Creating table of article titles
    rows = []
    for i in range(len(df)):
        row = []
        for col in df.columns:
            value = df.iloc[i][col]

            if col == 'Title':
                cell = html.Td(html.A(href=df.iloc[i]['Title'], children=value))
            elif col == 'Title':
                continue
            else:
                cell = html.Td(children=value)
            row.append(cell)
        rows.append(html.Tr(row))
    return html.Table(
        # Header
        [html.Tr([html.Th(col) for col in ['V', 'V']])] + rows
        )

separation_string = '''



'''

products_dictionary = dicts(df, 'Title')
#Taking elements from the excel
```

Εικόνα 27: Εντολή για δημιουργία πίνακα με βάση τον τίτλο των άρθρων.

Έτσι, ο πίνακας που δημιουργήθηκε θα έχει την εξής μορφή:

Εικόνα 28: Πίνακας για επιλογή άρθρου από τους χρήστες.

Έπειτα, δημιουργούμε τον τίτλο που θα έχει η σελίδα, όπως και τον τίτλο των δύο επιμέρους πινάκων: των άρθρων και των συστάσεων. Στην συγκεκριμένη εργασία, η σελίδα θα ονομάζεται Article recommendations.

```
app.layout = html.Div(style=colors, children=[
                html.H2(children='Article recommendations',
                style={
                'textAlign': 'center',
                'color': colors['text'],
                'backgroundColor': colors["background-image"],
                'font-family': 'Bangers'
            }
    ),
```

Εικόνα 29: Δημιουργία μορφής της σελίδας παραγωγής συστάσεων.

Αφού ολοκληρώθηκε η μορφή της εφαρμογής, το επόμενο στάδιο είναι η δημιουργία της αλληλεπίδρασης με τους χρήστες (callback). Στην συγκεκριμένη εφαρμογή η αλληλεπίδραση επιτυγχάνεται με τον χρήστη να επιλέγει ένα άρθρο από τον πρώτο πίνακα, ο οποίος περιέχει όλα τα άρθρα που εξορύξαμε, και στην συνέχεια το σύστημα παραγωγής συστάσεων παράγει τις σχετικές συστάσεις. Για την δημιουργία των συστάσεων χρησιμοποιήθηκε το φιλτράρισμα βάσει περιεχομένου (content based filtering).

Για να δημιουργηθεί ο κώδικας μέσω του content based filtering, συγκρίναμε ομοιότητες στοιχείων μεταξύ των άρθρων, όπως είναι ο τίτλος τους. Πιο συγκεκριμένα, στόχος ήταν να βρεθούν οι πιο κοινές λέξεις μεταξύ των τίτλων των άρθρων, χρησιμοποιώντας την εντολή του TfidfVectorizer. Έπειτα, στοχεύσαμε να βρούμε τον βαθμό ομοιότητας μεταξύ των τίτλων χρησιμοποιώντας το cosine similarity.

```
def contentrecommender(article):
    #Creating recommendations based on keywords,showing top results
    df_cont = df[['Title',]]
    df_cont.drop_duplicates(inplace=True)
    df_cont = df_cont.reset_index(drop=True)
    tf = TfidfVectorizer(analyzer='word', ngram_range=(1, 2), min_df=0, stop_words='english')
    tfidf_matrix = tf.fit_transform(df_cont['Title'])
    cosine_sim = linear_kernel(tfidf_matrix, tfidf_matrix)
    titles = df_cont[['Title']]
    indices = pd.Series(df_cont.index, index=df_cont['Title'])
    idx = indices[article]
    sim_scores = list(enumerate(cosine_sim[idx]))
    sim_scores = sorted(sim_scores, key=lambda x: x[1], reverse=True)
    sim_scores = sim_scores[1:11]
    article_indices = [i[0] for i in sim_scores]
    return Table(titles.iloc[article_indices])
```

Εικόνα 30: Δημιουργία συστήματος παραγωγής συστάσεων βασισμένο στο φιλτράρισμα βάσει περιεχομένου.

Με την περάτωση όλων των παραπάνω βημάτων, πλέον η εφαρμογή είναι έτοιμη, και τρέχει χρησιμοποιώντας τον παρακάτω κώδικα, στην διεύθυνση http://127.0.0.1:8050/.

```
if __name__ == '__main__':
    app.run_server(debug=True)
```

Εικόνα 31: Εκτέλεση εφαρμογής.

Εικόνα 32: Η εφαρμογή ολοκληρωμένη.

Συμπεράσματα

Σκοπός της παρούσης εργασίας ήταν να εξερευνήσει κατά πόσο είναι δυνατή η δημιουργία ενός περιβάλλοντος παραγωγής συστάσεων για μια δημοσιογραφική ιστοσελίδα. Απώτερος στόχος είναι η εξατομίκευση του περιεχομένου για τους χρήστες της, έτσι ώστε να παρουσιάζεται το είδος του περιεχομένου που θα είναι της αρέσκειας κάθε διαφορετικού χρήστη.

Όπως αναλύθηκε και παραπάνω, αυτήν την στιγμή υπάρχουν τρία είδη συστημάτων παραγωγής συστάσεων:

- Το φιλτράρισμα βάσει περιεχομένου (content based filtering).
- Το συνεργατικό φιλτράρισμα (collaborative filtering).
- Το υβριδικό φιλτράρισμα (υβριδικό φιλτράρισμα).

Τα τελευταία χρόνια, το υβριδικό φιλτράρισμα κερδίζει έδαφος στην παραγωγή συστάσεων καθώς χρησιμοποιεί και τα υπόλοιπα δύο είδη φιλτραρίσματος, εξουδετερώνοντας έτσι τα μειονεκτήματά τους και δημιουργώντας περισσότερο ολοκληρωμένες συστάσεις. Για τον λόγο αυτό, η συγκεκριμένη εργασία χρησιμοποίησε το υβριδικό φιλτράρισμα, επιβεβαιώνοντας την χρησιμότητα του.

Συγκεκριμένα, μέσω του υβριδικού φιλτραρίσματος κατέστη δυνατό να ενωθούν τα αποτελέσματα του content based και του collaborative filtering, δημιουργώντας πιο σωστές συστάσεις. Αυτό φαίνεται από το γεγονός ότι πως ενώ οι συστάσεις από το συνεργατικό φιλτράρισμα και το υβριδικό είναι ακριβώς οι ίδιες, το υβριδικό φιλτράρισμα τις κατατάσσει σε διαφορετική σειρά από ότι το συνεργατικό, καθώς έχει περισσότερα στοιχεία από τα οποία μπορεί να μάθει. Τα συγκεκριμένα άρθρα είναι τα:

- Taiwan hopes for close U.S. cooperation in call with Biden adviser
- Thank you Tegel: Berliners bid emotional farewell to Cold War airport
- Exclusive: Russian hackers targeted California, Indiana Democratic parties
- Japan opens airport coronavirus test lab for departing travellers

Η κατάταξή τους με βάση το υβριδικό φιλτράρισμα διαμορφώνεται ως εξής:

- Japan opens airport coronavirus test lab for departing travellers
- Thank you Tegel: Berliners bid emotional farewell to Cold War airport
- Taiwan hopes for close U.S. cooperation in call with Biden adviser
- Exclusive: Russian hackers targeted California, Indiana Democratic parties

Στην συνέχεια, ένα πολύ σημαντικό κομμάτι της εργασίας είναι κατά πόσο μπορεί να φτιαχτεί μια εφαρμογή η οποία θα παράγει εξατομικευμένες προτάσεις άρθρων στους χρήστες μια ενημερωτικής ιστοσελίδας. Για να συμβεί αυτό επιλέξαμε να χρησιμοποιήσουμε την πλατφόρμα Dash. Η συγκεκριμένη πλατφόρμα προσφέρει ένα περιβάλλον Python, στο οποίο μπορεί να συνταχθεί ο κώδικας προκειμένου να δημιουργηθεί μια εφαρμογή για το διαδίκτυο, χωρίς ωστόσο να χρειάζεται HTML κώδικα.

Έπειτα από έρευνα της λειτουργίας του Dash, κατέστη εφικτή η δημιουργία μιας εφαρμογής παραγωγής συστάσεων άρθρων. Η συγκεκριμένη εφαρμογή παρουσιάζει δύο πίνακες. Ο ένας πίνακας περιέχει περισσότερους από χίλιους τίτλους άρθρων, από τα άρθρα που εξορύχθησαν από το πρακτορείο Reuters. Οι χρήστες μπορούν να επιλέξουν οποιοδήποτε άρθρο επιθυμούν.

Ο δεύτερος πίνακας που περιέχει η εφαρμογή συμπεριλαμβάνει τις συστάσεις με βάση το άρθρο που επέλεξαν οι χρήστες. Οι συστάσεις παράγονται με βάση το φιλτράρισμα βάσει περιεχομένου, συγκρίνοντας τις ομοιότητες μεταξύ του άρθρου που επέλεξαν οι χρήστες, και των υπολοίπων, καταλήγοντας στα άρθρα με τις περισσότερες ομοιότητες.

Με βάση τα παραπάνω, η εργασία κρίνει πως είναι εφικτή η δημιουργία μιας εφαρμογής παραγωγής συστάσεων άρθρων για οποιοδήποτε ενημερωτική ιστοσελίδα, με απώτερο στόχο την εξατομίκευση του περιεχομένου για όλους τους χρήστες της.

Ως μελλοντικό έργο, η συγκεκριμένη εργασία σκοπεύει να ερευνήσει την αποτελεσματικότητα της εφαρμογής παραγωγής συστάσεων, λαμβάνοντας υπόψιν τις αξιολογήσεις χρηστών που θα χρησιμοποιήσουν την εφαρμογή.

Βιβλιογραφία

1. A. (n.d.). New Epsilon research indicates 80% of consumers are more likely to make a purchase when brands offer personalized experiences. EPSILON. Retrieved February 17, 2021, from https://us.epsilon.com/pressroom/new-epsilon-research-indicates-80-of-consumers-are-more-likely-to-make-a-purchase-when-brands-offer-personalized-experiences
2. Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering, 17(6), 734–749. https://doi.org/10.1109/tkde.2005.99
3. Aggarwal, C. C. (2016). Recommender Systems: The Textbook (1st ed. 2016 ed.). Springer.
4. Aggarwal, C. C., & Reddy, C. K. (2013). Data Clustering: Algorithms and Applications (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series) (1st ed.). Chapman and Hall/CRC.
5. Agrawal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules. VLDB. http://www.vldb.org/conf/1994/P487.PDF
6. Alonso, B. (2020, July 28). How to Use Content Personalization to Win Customers. Kibo Commerce. https://kibocommerce.com/blog/content-personalization/
7. Amr, T. (2020). Hands-On Machine Learning with scikit-learn and Scientific Python Toolkits: A practical guide to implementing supervised and unsupervised machine learning algorithms in Python. Packt Publishing.
8. Bouza, A. (2012). Hypothesis-Based Collaborative Filtering. Lulu Com.
9. Brownlee, J. (2019). Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch. https://books.google.gr/books?id=n--oDwAAQBAJ&printsec=frontcover&dq=supervised+machine+learning&hl=en&sa=X&ved=2ahUKEwjOs_u8w5HuAhXoiIsKHWx3CMIQ6AEwA3oECAIQAg#v=onepage&q=supervised%20machine%20learning&f=false
10. Burke, R. (2002). Hybrid Recommender Systems: Survey and Experiments. User Modeling and User-Adapted Interaction, 12(4), 1–11. https://doi.org/10.1023/a:1021240730564
11. Collaborative Filtering | Recommendation Systems |. (n.d.). Google Developers. Retrieved January 18, 2021, from https://developers.google.com/machine-learning/recommendation/collaborative/basics
12. Content-based Filtering | Recommendation Systems |. (n.d.). Google Developers. Retrieved January 17, 2021, from https://developers.google.com/machine-learning/recommendation/content-based/basics
13. Content-based Filtering Advantages & Disadvantages. (n.d.). Google Developers. Retrieved January 18, 2021, from https://developers.google.com/machine-learning/recommendation/content-based/summary
14. Data Mining Tasks. (n.d.). Wide Skills. Retrieved January 4, 2021, from https://www.wideskills.com/data-mining-tutorial/05-data-mining-tasks#:~:text=There%20are%20a%20number%20of,association%2C%20clustering%2C%20summarization%20etc.&text=Descriptive%20data%20mining%20tasks%20usually,from%20the%20available%20data%20set.
15. Davenport, T. H. (2019). The AI Advantage: How to Put the Artificial Intelligence Revolution to Work (Management on the Cutting Edge) (Reprint ed.). The MIT Press.

16. Ekstrand, M. D., Riedl, J. T., & Konstan, J. A. (2011). Collaborative Filtering Recommender Systems (Foundations and Trends(r) in Human-Computer Interaction). Now Publishers Inc.
17. For Local News, Americans Embrace Digital but Still Want Strong Community Connection. (2019). Pew Research Center. https://www.journalism.org/wp-content/uploads/sites/8/2019/03/PJ_2019.03.26_Local-News_FINAL.pdf
18. Gorunescu, F. (2011). Data Mining: Concepts, Models and Techniques (Vol. 12). Springer-Verlag Berlin and Heidelberg GmbH & Co. K.
19. Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems.
20. Hoekstra, R. (2010). The knowledge reengineering bottleneck. Semantic Web, 1(1,2), 2–4. https://doi.org/10.3233/sw-2010-0004
21. Introduction | Dash for Python Documentation | Plotly. (n.d.). Dash. Retrieved January 24, 2021, from https://dash.plotly.com/introduction
22. Jannach, D., Zanker, M., Felfernig, A., & Friedrich, G. (2010). Recommender Systems: An Introduction (Illustrated ed.). Cambridge University Press.
23. Johnston, B., & Mathur, I. (2019). Applied Supervised Learning with Python: Use scikit-learn to build predictive models from real-world datasets and prepare yourself for the future of machine learning. Packt Publishing.
24. Karat, J., Blom, J. O., & Karat, J. (2006). Designing Personalized User Experiences in eCommerce. Springer Publishing.
25. Kawamoto, K. (2003). Digital Journalism: Emerging Media and the Changing Horizons of Journalism (1st ed.). Rowman & Littlefield Publishers.
26. Kembellec, G., Chartron, G., & Saleh, I. (2014). Recommender Systems (Iste) (1st ed.). Wiley-ISTE.
27. K-means clustering. (n.d.). Wikipedia. Retrieved January 5, 2021, from https://en.wikipedia.org/wiki/K-means_clustering
28. Larose, D. T. (2015). Data Mining and Predictive Analytics (Wiley Series on Methods and Applications in Data Mining) (2nd ed.). Wiley.
29. Lugmayr, A., & Zotto, D. C. (2018). Media Convergence Handbook - Vol. 2: Firms and User Perspectives (Media Business and Innovation) (Softcover reprint of the original 1st ed. 2016 ed.). Springer.
30. Luo, S. (2019, February 6). Introduction to Recommender System - Towards Data Science. Medium. https://towardsdatascience.com/intro-to-recommender-system-collaborative-filtering-64a238194a26
31. Marconi, F. (2020). Newsmakers: Artificial Intelligence and the Future of Journalism. Columbia University Press.
32. Marsland, S. (2009). Machine Learning: An Algorithmic Perspective (Chapman & Hall/Crc Machine Learning & Pattern Recognition) (1st ed.). Chapman and Hall/CRC.
33. McKay, E. N. (2013). UI is Communication: How to Design Intuitive, User Centered Interfaces by Focusing on Effective Communication (1st ed.). Morgan Kaufmann.
34. Mohanty, S. N., Chatterjee, J. M., Jain, S., Elngar, A. A., & Gupta, P. (2020). Recommender System with Machine Learning and Artificial Intelligence: Practical Tools and Applications in Medical, Agricultural and Other Industries (1st ed.). Wiley-Scrivener.
35. Mourlas, C., & Germanakos, P. (2008). Intelligent User Interfaces: Adaptation and Personalization Systems and Technologies. Information Science Reference.

36. Nadimi-Shahraki, M.-H., & Bahadorpour, M. (2014). Cold-start Problem in Collaborative Recommender Systems: Efficient Methods Based on Ask-to-rate Technique. Journal of Computing and Information Technology, 22(2), 105–111. https://doi.org/10.2498/cit.1002223
37. Neumann, A. W. (2010). Recommender Systems for Information Providers: Designing Customer Centric Paths to Information (Contributions to Management Science) (Softcover reprint of hardcover 1st ed. 2009 ed.). Physica.
38. NumPy. (n.d.). NumPy. Retrieved January 23, 2021, from https://numpy.org/
39. Ratner, B. (2009). The correlation coefficient: Its values range between +1/−1, or do they? Journal of Targeting, Measurement and Analysis for Marketing, 17(2), 139–142. https://doi.org/10.1057/jt.2009.5
40. Reuters Institute Digital News Report 2020. (2020). Reuters Institute, University of Oxford. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-06/DNR_2020_FINAL.pdf
41. Ricci, F., Rokach, L., Shapira, B., & Kantor, P. B. (2010). Recommender Systems Handbook (2011th ed.). Springer.
42. Roberge, J., & Castelle, M. (2020). The Cultural Life of Machine Learning: An Incursion into Critical AI Studies (1st ed. 2021 ed.). Palgrave Macmillan.
43. Rocca, B. (2019, June 12). Introduction to recommender systems - Towards Data Science. Medium. https://towardsdatascience.com/introduction-to-recommender-systems-6c66cf15ada
44. Roy, A. (n.d.). Introduction To Recommender Systems- 1: Content-Based Filtering And Collaborative Filtering. Medium. Retrieved January 17, 2021, from https://towardsdatascience.com/introduction-to-recommender-systems-1-971bd274f421
45. R-Squared. (n.d.). Investopedia. Retrieved January 23, 2021, from https://www.investopedia.com/terms/r/r-squared.asp#:%7E:text=R%2Dsquared%20(R2),variables%20in%20a%20regression%20model.
46. Sarwar, B., Karypis, G., Konstan, J., & Reidl, J. (2001). Item-based collaborative filtering recommendation algorithms. Proceedings of the Tenth International Conference on World Wide Web - WWW '01, 2–5. https://doi.org/10.1145/371920.372071
47. Siapera, E., & Veglis, A. (2012). The Handbook of Global Online Journalism (1st ed.). Wiley-Blackwell.
48. sklearn.feature_extraction.text.CountVectorizer — scikit-learn 0.24.1 documentation. (n.d.). Sklearn. Retrieved January 22, 2021, from https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html
49. sklearn.metrics.pairwise.cosine_similarity — scikit-learn 0.24.1 documentation. (n.d.). Sklearn. Retrieved January 23, 2021, from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html
50. Sotirakou, (2015), Σχεδίαση εφαρμογής για κινητές συσκευές για ειδησεογραφικό περιεχόμενο, με δυνατότητες εξατομίκευσης και χρήσης τεχνικών παιχνιδοποίησης, Published, Master's Thesis, National and Kapodistrian University of Greece, https://masters.ntlab.gr/wp-content/uploads/2018/07/ergasia_sotirakou2015.pdf
51. Tutorials — pandas 0.15.2 documentation. (n.d.). Pandas. Retrieved January 22, 2021, from https://pandas.pydata.org/pandas-docs/version/0.15/tutorials.html

52. Uchyigit, G., & Ma, M. Y. (2008). Personalization Techniques And Recommender Systems (Series in Machine Perception and Artifical Intelligence) (Volume 70). Wspc.
53. Wieringa, J. (n.d.). Intro to Beautiful Soup. Programming Historian. Retrieved January 21, 2021, from https://programminghistorian.org/en/lessons/intro-to-beautiful-soup
54. Wikipedia contributors. (n.d.). Apriori algorithm. Wikipedia. Retrieved January 18, 2021, from https://en.wikipedia.org/wiki/Apriori_algorithm#cite_note-apriori-1
55. Wikipedia contributors. (n.d.-b). Recommender system. Wikipedia. Retrieved January 10, 2021, from https://en.wikipedia.org/wiki/Recommender_system#cite_note-54
56. X. Zhao, (2016), Cold-Start Collaborative Filtering, Published, Master's Thesis, University College London, https://www.easybib.com/cite/results?source=encyclopedia&query=https%3A%2F%2Fdiscovery.ucl.ac.uk%2Fid%2Feprint%2F1474118%2F1%2FThesis_final_revision.pdf

Παραρτήματα

Κώδικας για την δημιουργία συστήματος παραγωγής συστάσεων με υβριδικό φιλτράρισμα

Σύνδεσμος για τον πλήρη κώδικα δημιουργίας του υβριδικού φιλτραρίσματος:
https://github.com/iltoum/DataAnalysisPython/blob/master/DIPLWMATIKI_FINAL.ipynb

CONTENT BASED

```
In [ ]: import pandas as pd
        !pip install rake-nltk
        from rake_nltk import Rake
        import numpy as np
        from sklearn.metrics.pairwise import cosine_similarity
        from sklearn.feature_extraction.text import CountVectorizer
```

```
Requirement already satisfied: rake-nltk in /usr/local/lib/python3.6/dist-packages (1.0.4)
Requirement already satisfied: nltk in /usr/local/lib/python3.6/dist-packages (from rake-nltk) (3.2.5)
Requirement already satisfied: six in /usr/local/lib/python3.6/dist-packages (from nltk->rake-nltk) (1.15.0)
```

```
In [ ]: from google.colab import files
        uploaded = files.upload()
```

```
Saving reuters.csv to reuters.csv
```

```
In [ ]: import io
        df_reuters = pd.read_csv(io.BytesIO(uploaded['reuters.csv']))
```

```
In [ ]: df_reuters.iloc[0]
```

```
Out[ ]: web-scraper-order                                    1606153665-2067
        web-scraper-start-url    https://www.reuters.com/news/archive/worldNews...
        pagination                                                        Earlier
        pagination-href          https://www.reuters.com/news/archive/worldNews...
        link                     Armenia, Azerbaijan agree to defuse Nagorno-Ka...
        link-href                https://www.reuters.com/article/us-armenia-aze...
        Title                    Armenia, Azerbaijan agree to defuse Nagorno-Ka...
        Date                                                         October 30, 2020
        Headline                 GENEVA (Reuters) - The foreign ministers of Ar...
        Name: 0, dtype: object
```

Εικόνα 33: Εισαγωγή βιβλιοθηκών και αρχείου για φιλτράρισμα βάσει περιεχομένου.

```
In [ ]: df_reuters = df_reuters[['Title','Headline']]
        df_reuters.head(5)
```

Out[ ]:

	Title	Headline
0	Armenia, Azerbaijan agree to defuse Nagorno-Ka...	GENEVA (Reuters) - The foreign ministers of Ar...
1	Dealing with Trump presidency nemesis Iran won...	WASHINGTON (Reuters) - When reality TV star Do...
2	Exclusive: Russian hackers targeted California...	WASHINGTON (Reuters) - The group of Russian ha...
3	Spain to send more police to Senegal to curb i...	DAKAR (Reuters) - Spain will increase its poli...
4	Canada expects six million COVID-19 vaccine do...	(Reuters) - Canada expects to receive six mill...

```
In [ ]: df_reuters['Headline']=df_reuters['Headline'].apply(str)
```

```
In [ ]: df_reuters['Key_words'] = ""

        for index, row in df_reuters.iterrows():
            headline = row['Headline']

            r = Rake()

            r.extract_keywords_from_text(headline)

            key_words_dict_scores = r.get_word_degrees()

            row['Key_words'] = list(key_words_dict_scores.keys())

        df_reuters.drop(columns = ['Headline'], inplace = True)
```

Εικόνα 34: Εξαγωγή λέξεων κλειδιών.

```
In [ ]: df_reuters.set_index('Title', inplace = True)
        df_reuters.head(5)
```

Out[ ]:

	Key_words
Title	
Armenia, Azerbaijan agree to defuse Nagorno-Karabakh conflict	[foreign, ministers, killed, nagorno, fighting...
Dealing with Trump presidency nemesis Iran won't be 'quick, easy' for Biden	[landmark, deal, aimed, stopping, tehran, isla...
Exclusive: Russian hackers targeted California, Indiana Democratic parties	[indiana, email, accounts, knowledge, accordin...
Spain to send more police to Senegal to curb illegal migration, says foreign minister	[increase, spain, territory, west, african, co...
Canada expects six million COVID-19 vaccine doses early in 2021	[2021, regulatory, approval, parliamentary, co...

```
In [ ]: df_reuters['key_words_string'] = [','.join(map(str, l)) for l in df_reuters['Key_words']]
        df_reuters
```

Out[ ]:

	Key_words	key_words_string
Title		
Armenia, Azerbaijan agree to defuse Nagorno-Karabakh conflict	[foreign, ministers, killed, nagorno, fighting...	foreign,ministers,killed,nagorno,fighting,conf...
Dealing with Trump presidency nemesis Iran won't be 'quick, easy' for Biden	[landmark, deal, aimed, stopping, tehran, isla...	landmark,deal,aimed,stopping,tehran,islamic,re...
Exclusive: Russian hackers targeted California, Indiana Democratic parties	[indiana, email, accounts, knowledge, accordin...	indiana,email,accounts,knowledge,according,cal...
Spain to send more police to Senegal to curb illegal migration, says foreign minister	[increase, spain, territory, west, african, co...	increase,spain,territory,west,african,coast,da...
Canada expects six million COVID-19 vaccine doses early in 2021	[2021, regulatory, approval, parliamentary, co...	2021,regulatory,approval,parliamentary,committ...

Εικόνα 35: Αποτελέσματα εξαγωγής λέξεων κλειδιών.

```
In [ ]: df_reuters2 = df_reuters.drop('Key_words', 1)
        df_reuters2.head(5)
```

Out[ ]:

	key_words_string
Title	
Armenia, Azerbaijan agree to defuse Nagorno-Karabakh conflict	foreign,ministers,killed,nagorno,fighting,conf...
Dealing with Trump presidency nemesis Iran won't be 'quick, easy' for Biden	landmark,deal,aimed,stopping,tehran,islamic,re...
Exclusive: Russian hackers targeted California, Indiana Democratic parties	indiana,email,accounts,knowledge,according,cal...
Spain to send more police to Senegal to curb illegal migration, says foreign minister	increase,spain,territory,west,african,coast,da...
Canada expects six million COVID-19 vaccine doses early in 2021	2021,regulatory,approval,parliamentary,committ...

```
In [ ]: count = CountVectorizer()
        count_matrix = count.fit_transform(df_reuters['key_words_string'])


        indices = pd.Series(df_reuters.index)
        indices[:5]
```

```
Out[ ]: 0    Armenia, Azerbaijan agree to defuse Nagorno-Ka...
        1    Dealing with Trump presidency nemesis Iran won...
        2    Exclusive: Russian hackers targeted California...
        3    Spain to send more police to Senegal to curb i...
        4    Canada expects six million COVID-19 vaccine do...
        Name: Title, dtype: object
```

Εικόνα 36: Χρήση CountVectorizer για εύρεση συχνότερων λέξεων.

0	Turkey says Azerbaijan achieved 'sacred succes...	0.430528
0	Armenia, Azerbaijan, Russia say sign deal to e...	0.404577
0	Russia says discussing U.N. presence in Nagorn...	0.383065
0	Azerbaijan's president says deal is signed to ...	0.354005
0	Leader of Nagorno-Karabakh says ceasefire with...	0.344031
0	Putin calls for Turkish involvement in Nagorno...	0.327144
0	Turkey says in talks on how to monitor Karabak...	0.321029
0	Russia reinforces border guards in Armenia aft...	0.318511

Εικόνα 37: Τελικές προτάσεις με φιλτράρισμα βάσει περιεχομένου

COLLABORATIVE

```
In [ ]: import pandas as pd
```

```
In [ ]: from google.colab import files
        uploaded = files.upload()
```

```
Saving ratings2.csv to ratings2.csv
```

```
In [ ]: import io
        df_ratings = pd.read_csv(io.BytesIO(uploaded['ratings2.csv']))
```

```
In [ ]: df_ratings.head(5)
```

Out[ ]:

	UserID	ArticleID	Ratings
0	1	Armenia, Azerbaijan agree to defuse Nagorno-Ka...	2
1	1	Dealing with Trump presidency nemesis Iran won...	1
2	1	Exclusive: Russian hackers targeted California...	3
3	1	Spain to send more police to Senegal to curb i...	4
4	1	Canada expects six million COVID-19 vaccine do...	5

Εικόνα 38: Εισαγωγή βιβλιοθηκών και αρχείου για συνεργατικό φιλτράρισμα.

```
In [ ]: raters=df_ratings.groupby(['ArticleID'])['Ratings'].count()
        print(raters)
```

```
ArticleID
Australia alleges military carried out unlawful killings in Afghanistan              3
Burkina opposition candidate alleges 'massive fraud' ahead of Sunday vote            1
COVID-19 cases soar in Brazil's largest indigenous reservation                       3
Czechs mark anniversary of Velvet Revolution amid pandemic                           1
Dealing with Trump presidency nemesis Iran won't be 'quick, easy' for Biden          1
England will need five days of lockdown for each day relaxed at Christmas: adviser   1
Ethiopia's Tigray forces' leader denies regional capital circled                     1
Exclusive: Russian hackers targeted California, Indiana Democratic parties           3
Factbox: As mediation calls mount, who has leverage in Ethiopia?                     2
France reports 271 new COVID-19 deaths                                               2
Gunmen kill at least 11 in attack on Iraqi army post in Baghdad, sources say         1
Italy has 48 hours to approve new COVID-19 restrictions: Health Minister             3
Japan opens airport coronavirus test lab for departing travellers                    2
No more bullying': fresh start to U.S.-Mexico relations eyed under Biden             2
Pakistan minister deletes tweet containing Macron Nazi jibe                          1
Palestinian Authority resuming cooperation with Israel, Palestinian official says    3
Poland looks into coronavirus risks at mink farms                                    1
Polish cardinal accused of sexual abuse dies aged 97                                 2
Pompeo says Europe, U.S. need to work together to address Turkey                     2
Portugal's president ponders COVID-19 state of emergency                             1
Saudi minister Al-Jubeir says Vienna attack is a crime contrary to all religions     1
South Korea's Moon congratulates Biden, to ensure no gap in U.S. alliance            2
Syrian satirist questioned over 'Macron' flogging stunt in Berlin                    1
Taiwan hopes for close U.S. cooperation in call with Biden adviser                   3
Thank you Tegel: Berliners bid emotional farewell to Cold War airport                3
Turkey says it will congratulate U.S. election winner once result finalised          1
U.N. chief congratulates Biden, says U.S. 'essential' to global cooperation          1
Veteran Syrian diplomat Mekdad named foreign minister, state media says              2
Name: Ratings, dtype: int64
```

Εικόνα 39: Ομαδοποίηση δεδομένων με βάση το ID του χρήστη και την βαθμολογία

```
In [ ]: selfjoined = data.merge(data,on='UserID')
```

```
In [ ]: selfjoined.head(5)
```

Out[ ]:

	UserID	ArticleID_x	count_x	Ratings_x	ArticleID_y	count_y	Ratings_y
0	3	Polish cardinal accused of sexual abuse dies a...	2	3	Polish cardinal accused of sexual abuse dies a...	2	3
1	3	Polish cardinal accused of sexual abuse dies a...	2	3	Syrian satirist questioned over 'Macron' flogg...	1	2
2	3	Polish cardinal accused of sexual abuse dies a...	2	3	No more bullying': fresh start to U.S.-Mexico ...	2	2
3	3	Polish cardinal accused of sexual abuse dies a...	2	3	Pompeo says Europe, U.S. need to work together...	2	1
4	3	Polish cardinal accused of sexual abuse dies a...	2	3	Portugal's president ponders COVID-19 state of...	1	5

```
In [ ]: selfjoined = selfjoined[selfjoined['ArticleID_x']<selfjoined['ArticleID_y']]
```

```
In [ ]: import numpy as np
        selfjoined['r1r2'] = selfjoined['Ratings_x']*selfjoined['Ratings_y']
        selfjoined['r1square'] = np.square(selfjoined['Ratings_x'])
        selfjoined['r2square'] = np.square(selfjoined['Ratings_y'])
```

```
In [ ]: selfjoined.head(5)
```

Out[ ]:

	UserID	ArticleID_x	count_x	Ratings_x	ArticleID_y	count_y	Ratings_y	r1r2	r1square	r2square
1	3	Polish cardinal accused of sexual abuse dies a...	2	3	Syrian satirist questioned over 'Macron' flogg...	1	2	6	9	4

Εικόνα 40: Μαθηματική συνάρτηση για τον υπολογισμό της τετραγωνικής τιμής κάθε στοιχείου στον πίνακα.

```
In [ ]: aggdata=selfjoined.groupby(['ArticleID_x','ArticleID_y'])['Ratings_x','Ratings_y','r1r2','r1square','r2square','count_x','count_y'].sum()
aggdata.head(5)
```

```
/usr/local/lib/python3.6/dist-packages/ipykernel_launcher.py:1: FutureWarning: Indexing with multiple keys (implicitly converted to a tup
le of keys) will be deprecated, use a list instead.
  """Entry point for launching an IPython kernel.
```

Out[ ]:

		Ratings_x	Ratings_y	r1r2	r1square	r2square	count_x	count_y
ArticleID_x	**ArticleID_y**							
Australia alleges military carried out unlawful killings in Afghanistan	**COVID-19 cases soar in Brazil's largest indigenous reservation**	5	5	8	17	17	6	6
	Factbox: As mediation calls mount, who has leverage in Ethiopia?	4	2	8	16	4	3	2
	France reports 271 new COVID-19 deaths	4	4	16	16	16	3	2
	Italy has 48 hours to approve new COVID-19 restrictions: Health Minister	4	4	16	16	16	3	3
	Pakistan minister deletes tweet containing Macron Nazi jibe	4	3	12	16	9	3	1

Εικόνα 41: Καθορισμός συντελεστής προσδιορισμού R2 συγκεκριμένου τίτλου.

```
In [ ]: n=aggdata.shape[0]
```

```
In [ ]: aggdata['corelation'] = (n*aggdata['r1r2'] - aggdata['Ratings_x']*aggdata['Ratings_y']) / \
np.sqrt((n * aggdata['r1square']  - np.square(aggdata['Ratings_x']))*(n * aggdata['r2square']  - np.square(aggdata['Ratings_y'])))
```

```
In [ ]: aggdata['corelation'] = aggdata['corelation']*n/(n+50)
```

```
In [ ]: aggdata.reset_index(inplace=True)
aggdata.head(5)
```

Out[ ]:

	ArticleID_x	ArticleID_y	Ratings_x	Ratings_y	r1r2	r1square	r2square	count_x	count_y	corelation
0	Australia alleges military carried out unlawfu...	COVID-19 cases soar in Brazil's largest indige...	5	5	8	17	17	6	6	0.309532
1	Australia alleges military carried out unlawfu...	Factbox: As mediation calls mount, who has lev...	4	2	8	16	4	3	2	0.668874
2	Australia alleges military carried out unlawfu...	France reports 271 new COVID-19 deaths	4	4	16	16	16	3	2	0.668874
3	Australia alleges military carried out unlawfu...	Italy has 48 hours to approve new COVID-19 res...	4	4	16	16	16	3	3	0.668874

Εικόνα 42: Καθορισμός συσχέτισης άρθρων μεταξύ τους.

```
In [ ]: def recommendation(Title):
    recommended_movies = []
    data =aggdata2[aggdata2['Title_x']==Title]
    data = data.sort_values(by='corelation',ascending=False)
    return data
```

```
In [ ]: collabaritive=recommendation('England will need five days of lockdown for each day relaxed at Christmas: adviser')[['Title_y','corelatio
n']]
collabaritive['Title'] = collabaritive['Title_y'].str.split(' \(').str.get(0)
collabaritive.drop('Title_y',axis=1,inplace=True)
collabaritive.head(5)
```

Out[ ]:

	corelation	Title
40	0.668874	Taiwan hopes for close U.S. cooperation in cal...
49	0.668874	Thank you Tegel: Berliners bid emotional farew...
71	0.668874	Exclusive: Russian hackers targeted California...
80	0.668874	Japan opens airport coronavirus test lab for d...

Εικόνα 43: Συστάσεις άρθρων με συνεργατικό φιλτράρισμα.

HYBRID

```
In [ ]: contentrating = pd.read_csv('contentrating.csv')
        contentrating.head(10)
```

Out[ ]:

	Unnamed: 0	0	1
0	0	Armenia, Azerbaijan agree to defuse Nagorno-Ka...	1.000000
1	0	Armenia, Azerbaijan agree to defuse Nagorno-Ka...	0.621874
2	0	Turkey says Azerbaijan achieved 'sacred succes...	0.430528
3	0	Armenia, Azerbaijan, Russia say sign deal to e...	0.404577
4	0	Russia says discussing U.N. presence in Nagorn...	0.383065
5	0	Azerbaijan's president says deal is signed to ...	0.354005
6	0	Leader of Nagorno-Karabakh says ceasefire with...	0.344031
7	0	Putin calls for Turkish involvement in Nagorno...	0.327144
8	0	Turkey says in talks on how to monitor Karabak...	0.321029
9	0	Russia reinforces border guards in Armenia aft...	0.318511

```
In [ ]: hybrid=contentrating.merge(collabaritive,left_on='0',right_on='Title')
        hybrid=hybrid[['Title','1','corelation']]
        hybrid['wcorelation'] = (hybrid['1'] + hybrid['corelation'])/2
```

Εικόνα 44: Συνδυασμός προτάσεων με φιλτράρισμα βάσει περιεχομένου και συνεργατικό φιλτράρισμα.

```
In [ ]: hybrid.head(5)
```

Out[ ]:

	Title	1	corelation	wcorelation
0	Japan opens airport coronavirus test lab for d...	0.093250	0.668874	0.381062
1	Thank you Tegel: Berliners bid emotional farew...	0.045502	0.668874	0.357188
2	Taiwan hopes for close U.S. cooperation in cal...	0.040129	0.668874	0.354501
3	Exclusive: Russian hackers targeted California...	0.038720	0.668874	0.353797

Εικόνα 45: Τελικές προτάσεις άρθρων με χρήση υβριδικού φιλτραρίσματος.

Κώδικας για την δημιουργία εφαρμογής στο Dash με σκοπό την παραγωγή συστάσεων άρθρων.

```
import dash
import dash_html_components as html
import dash_core_components as dcc
import dash_bootstrap_components as dbc
import pandas as pd
import recomms
from dash.dependencies import Input, Output
import ast
from scipy import stats
from ast import literal_eval
from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer
from sklearn.metrics.pairwise import linear_kernel, cosine_similarity

df = pd.read_csv('reuters.csv', index_col=[0])

app = dash.Dash(external_stylesheets=[dbc.themes.BOOTSTRAP])

server = app.server
```

Εικόνα 46: Εισαγωγή βιβλιοθηκών και δημιουργία εφαρμογής και server.

```
def dicts(df, colname):
    vals = list(set(df[colname]))
    l = []
    for i in vals:
        dic = {}
        dic['label'] = i
        dic['value'] = i
        l.append(dic)
    return l

def Table(df):
    #Creating table of article titles
    rows = []
    for i in range(len(df)):
        row = []
        for col in df.columns:
            value = df.iloc[i][col]

            if col == 'Title':
                cell = html.Td(html.A(href=df.iloc[i]['Title'], children=value))
            elif col == 'Title':
                continue
            else:
                cell = html.Td(children=value)
            row.append(cell)
        rows.append(html.Tr(row))
    return html.Table(
        # Header
        [html.Tr([html.Th(col) for col in ['V', 'V']])] + rows
        )
```

Εικόνα 47: Δημιουργία πίνακα με άρθρα για να επιλέξουν οι χρήστες.

```
markdown_text_2 = '''
__Recommended articles:__
'''

colors = {
    #'background': '#1DB954',
    "text": "#111111",
    "background-image" : "url('/assets/wallpaperskin_retouched.png')",
    "background-size": "cover",
}
```

Εικόνα 48: Δημιουργία πίνακα που θα περιέχει τις συστάσεις των άρθρων.

```
app.layout = html.Div(style=colors, children=[
                    html.H2(children='Article recommendations',
                    style={
                    'textAlign': 'center',
                    'color': colors['text'],
                    'backgroundColor': colors["background-image"],
                    'font-family': 'Bangers'
                }
    ),
    html.Label('Choose an article'),
    dcc.Dropdown(
        id='article-selector',
        options=products_dictionary,
        placeholder='Select'
    ),

    dcc.Markdown(children=markdown_text_2),

    html.Div(id='output_2'),

    dcc.Markdown(children=separation_string),
    ])
```

Εικόνα 49: Δημιουργία drop-down για τον πίνακα που θα περιέχει τα άρθρα.

```
@app.callback(
    Output('output_2', 'children'),
    [Input('article-selector', 'value')]
    )

def contentrecommender(article):
    #Creating recommendations based on keywords,showing top results
    df_cont = df[['Title',]]
    df_cont.drop_duplicates(inplace=True)
    df_cont = df_cont.reset_index(drop=True)
    tf = TfidfVectorizer(analyzer='word', ngram_range=(1, 2), min_df=0, stop_words='english')
    tfidf_matrix = tf.fit_transform(df_cont['Title'])
    cosine_sim = linear_kernel(tfidf_matrix, tfidf_matrix)
    titles = df_cont[['Title']]
    indices = pd.Series(df_cont.index, index=df_cont['Title'])
    idx = indices[article]
    sim_scores = list(enumerate(cosine_sim[idx]))
    sim_scores = sorted(sim_scores, key=lambda x: x[1], reverse=True)
    sim_scores = sim_scores[1:11]
    article_indices = [i[0] for i in sim_scores]
    return Table(titles.iloc[article_indices])
```

Εικόνα 50: Δημιουργία αλληλεπίδρασης με χρήστες και παρουσίαση προτάσεων άρθρων.

```
if __name__ == '__main__':
    app.run_server(debug=True)
```

Εικόνα 51: Εκτέλεση εφαρμογής.