



ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΘΕΜΑ

Εφαρμογή AI και Big Data στον Τουρισμό

Εισηγητής: ΦΑΣΟΥΛΑΣ ΦΩΤΗΣ

Επιβλέπων καθηγητής : ΚΩΣΤΑΝΤΙΝΟΣ ΒΟΓΚΛΗΣ

ΠΕΡΙΕΧΟΜΕΝΑ:

<i>Εισαγωγή.....</i>	<i>3</i>
<i>1 Τεχνητή Νοημοσύνη</i>	<i>4</i>
<i>2 ΑΙ, μηχανική μάθηση και Επιστήμη Δεδομένων.....</i>	<i>5</i>
<i>3 ΑΙ Για Μεγάλα Δεδομένα.....</i>	<i>7</i>
3.1 ΑΙ για διαμόρφωση λύσεων προβλημάτων.....	8
3.2 ΑΙ για τη συλλογή δεδομένων	9
3.3 ΑΙ για καθαρισμό και προετοιμασία δεδομένων.....	9
3.4 ΑΙ για επεξεργασία δεδομένων	10
3.5 ΑΙ για εκμετάλλευση δεδομένων.....	11
<i>4 ΑΙ και μεγάλα δεδομένα στον τουρισμό.....</i>	<i>11</i>
<i>5 Εξόρυξη δεδομένων.....</i>	<i>14</i>
5.1 Ανοιχτά Δεδομένα	16
5.2 API	17
5.3 Απόξεση Δεδομένων.....	17
5.4 Νομικές προοπτικές εξόρυξης κειμένου και δεδομένων	18
5.5 Τυπικές περιπτώσεις χρήσης απόξεσης ιστού στον τουρισμό.....	19
5.6 BeautifulSoup	20
5.7 selenium	20
5.8 Scrapy	21
<i>Παράρτημα – Οδηγίες Κώδικα</i>	<i>21</i>

Εισαγωγή

Η τεχνητή νοημοσύνη (AI) και η ανάλυση μεγάλων δεδομένων (Big Data) αποτελούν δύο από τις πιο ραγδαία εξελισσόμενες τεχνολογίες που αναδιαμορφώνουν τον τομέα του τουρισμού παγκοσμίως. Με την ικανότητα να επεξεργάζονται και να αναλύουν τεράστιους όγκους δεδομένων σε πραγματικό χρόνο, αυτές οι τεχνολογίες επιτρέπουν στις τουριστικές επιχειρήσεις να βελτιώνουν την εμπειρία των πελατών, να αυξάνουν την αποδοτικότητα των λειτουργιών τους και να προβλέπουν με μεγαλύτερη ακρίβεια τις ανάγκες και τις τάσεις της αγοράς. Μέσω της εφαρμογής τεχνητής νοημοσύνης, μπορούν να αναπτυχθούν εξατομικευμένες υπηρεσίες και προϊόντα, όπως προτάσεις προορισμών, βελτιστοποίηση τιμολόγησης και ανάλυση συναισθημάτων από κριτικές, συμβάλλοντας στην αναβάθμιση της ανταγωνιστικότητας και της βιωσιμότητας των τουριστικών προορισμών.

Η ανάλυση μεγάλων δεδομένων επιτρέπει στους προορισμούς να κατανοούν καλύτερα τη συμπεριφορά των ταξιδιωτών, λαμβάνοντας υπόψη δεδομένα από πολλές πηγές όπως πλατφόρμες κοινωνικής δικτύωσης, ηλεκτρονικές κρατήσεις, κριτικές και άλλες μορφές ψηφιακού αποτυπώματος. Με τη σωστή χρήση αυτών των δεδομένων, οι τουριστικοί οργανισμοί μπορούν να βελτιστοποιήσουν τη στρατηγική μάρκετινγκ, να ενισχύσουν τη διαχείριση των πόρων τους και να ενσωματώσουν πρακτικές βιώσιμης ανάπτυξης. Αυτή η πτυχιακή εργασία στοχεύει στην ανάλυση των τρόπων με τους οποίους η AI και τα Big Data εφαρμόζονται στον τουρισμό, εξετάζοντας παραδείγματα από την πραγματικότητα και διερευνώντας τις μελλοντικές προοπτικές αυτών των τεχνολογιών στη βιομηχανία του τουρισμού.

Στην παρούσα εργασία επιχειρούμε μια εκτενή και συστηματική βιβλιογραφική επισκόπηση σχετικά με τη χρήση Big Data και τεχνητής νοημοσύνης στον τουρισμό. Ο στόχος μας είναι να αναλύσουμε και να συνοψίσουμε τις υπάρχουσες θεωρητικές και εμπειρικές μελέτες που αφορούν την εφαρμογή της τεχνητής νοημοσύνης και της ανάλυσης μεγάλων δεδομένων στον τομέα του τουρισμού. Μέσω της επισκόπησης αυτής, εξετάζονται οι κύριες τάσεις, προκλήσεις και καινοτομίες που έχουν αναδυθεί στην ερευνητική βιβλιογραφία, προσφέροντας μια ολοκληρωμένη κατανόηση των σχετικών τεχνολογιών και των επιπτώσεών τους στις τουριστικές επιχειρήσεις και τους προορισμούς.

1 Τεχνητή Νοημοσύνη

Ο κλασικός ορισμός της Τεχνητής Νοημοσύνης (AI) χρονολογείται από τη δεκαετία του 1950 (McCarthy 2007) και περιλαμβάνει την ιδέα ότι η τεχνητή νοημοσύνη θα επιτρέψει στους υπολογιστές να εκτελούν εργασίες και δραστηριότητες που θεωρούνται ευφυείς, δηλαδή απαιτούν νοημοσύνη ανθρώπινου επιπέδου. Δεδομένης της δυσκολίας του ορισμού της ανθρώπινης νοημοσύνης, ένας πιο πρακτικός ορισμός μπορεί να αναφέρεται στις ικανότητες και τις δυνατότητες που η τεχνητή νοημοσύνη στοχεύει να αυτοματοποιήσει:

- Η επικοινωνία, συμπεριλαμβανομένων όλων των μορφών και όλων των τύπων μέσων (Κείμενο, Εικόνα, Ήχος, Βίντεο).
- Η αντίληψη, έχει λάβει μεγάλη προσοχή τα τελευταία χρόνια λόγω της ανάπτυξης νέων συσκευών εισόδου/εξόδου (π.χ. αισθητήρες, Internet of Things (IoT) και κυβερνο-φυσικά συστήματα).
- Η γνώση, καθίσταται διαθέσιμη για την αποθήκευση, την ανάκτηση και την επεξεργασία για ποικίλες εφαρμογές.
- Ο σχεδιασμός, λειτουργεί ως (δεύτερης μορφής) στήριξης / εφεδρική υποστήριξη (για τη λήψη αποφάσεων και την ανταπόκριση (π.χ. στη ρομποτική και αυτόνομη οδήγηση).
- Η συλλογιστική, είναι η προσομοίωση ανθρώπινων διαδικασιών σκέψης και μάθησης.

Καθώς όλες αυτές οι δυνατότητες? είναι αλληλένδετες, το ίδιο ισχύει και για τα αντίστοιχα υπόπεδιά της τεχνητής νοημοσύνης, που χαρακτηρίζεται από διάφορες προσεγγίσεις και εφαρμογές. Ένας κατάλογος των κλασικών βασικών τομέων της έρευνας για την τεχνητή νοημοσύνη συμπεριλαμβάνουν την επίλυση προβλημάτων, τους ευφυείς παράγοντες, τους φυσικούς παράγοντες, την επεξεργασία γλώσσας (NLP), την αναγνώριση ομιλίας, την όραση υπολογιστή, την ρομποτική, την γνώση αναπαράσταση άκρων και την μηχανική μάθηση.

Παρά τα σκαμπανεβάσματα, συμπεριλαμβανομένων των λεγόμενων "χειμώνων" τεχνητής νοημοσύνης (AI) (Lim, 2018), η πρόοδος της τεχνητής νοημοσύνης είναι ξεκάθαρη σε όλους τους προαναφερθέντες κλάδους. Ωστόσο, το ερώτημα εάν οι υπολογιστές είναι πράγματι ευφυείς παραμένει ανοιχτό με μια θετική κλίση. Για παράδειγμα, ακόμη και στην επεξεργασία γλώσσας (NLP) οποίος είναι ένας κλάδος AI με εντυπωσιακά επιτεύγματα, οι υπολογιστές δεν "καταλαβαίνουν" πραγματικά το περιεχόμενο ενός κειμένου. Στην πραγματικότητα, δεν μπορούν να υποστηρίξουν στο εκατό τις εκατό τις αλληλεπιδράσεις στη φυσική γλώσσα μέχρι στιγμής, τα υπάρχοντα συνομιλητικά chatbots χρησιμοποιούν μεγάλα σύνολα δεδομένων από

ανταλλασσόμενες προτάσεις αλλά δεν "γνωρίζουν" το περιεχόμενο των προτάσεων αυτών, ούτε θα μπορούσαν να εξηγήσουν τις απαντήσεις τους. Το ίδιο συμβαίνει και με τα συστήματα αυτόματης μετάφρασης.

Επιπλέον, ο ρόλος της βιομηχανίας στην ανάπτυξη της τεχνητής νοημοσύνης επιβεβαιώνεται από την αύξηση των πτυχιούχων Διδακτορικών Σπουδών της (τεχνητής νοημοσύνης) που πηγαίνουν στον κλάδο της βιομηχανίας: στη Βόρεια Αμερική, μια αύξηση 65% το 2019, από 44,4% το 2010, θα μπορούσε να παρατηρηθεί και ως φαινόμενο. Επιπλέον, οι κυβερνήσεις καταστρώνουν επίσης τις στρατηγικές τους πάνω στην τεχνητή νοημοσύνη. Για παράδειγμα, Ο Καναδάς δημοσίευσε τη στρατηγική του το 2017, ακολουθούμενη από περισσότερες από άλλες 30 χώρες και περιφέρειες τον Δεκέμβριο του 2020 (Zhang et al., 2021). Οι επενδύσεις στο AI και τα μεγάλα δεδομένα (Big Data) είναι κάτι από το οποίο μπορεί να ωφεληθεί σημαντικά ο τουριστικός τομέας.

2 AI, μηχανική μάθηση και Επιστήμη Δεδομένων

Η πρόσφατη πρόοδος της Τεχνητής Νοημοσύνης οφείλεται στη Μηχανική Μάθηση (Machine Learning). Η Μηχανική Μάθηση είναι ένα υπόπεδίο της τεχνητής νοημοσύνης, και ο στόχος της είναι να πραγματοποιήσει υπολογιστικά μοντέλα τα οποία κάνουν τους υπολογιστές να μαθαίνουν "τι να κάνουν" αντί να τους προγραμματίσεις "πώς να το κάνουν". Ένα βασικό/ουσιώδες μέρος της ιστορίας των υπολογιστών συνδέεται με την έννοια των προγραμμάτων αποθήκευσης, στο οποίο οι οδηγίες σχετικά με τον τρόπο επίλυσης ενός προβλήματος καταχωρούνται σε μια μνήμη και εκτελούνται με βάση τα δεδομένα εισόδου προκειμένου να παραχθούν τα δεδομένα εξόδου. Έτσι, τα προγράμματα ήταν το αποτέλεσμα της μετάφρασης αλγορίθμων χρησιμοποιώντας μια (τεχνητή) γλώσσα προγραμματισμού. Ως εκ τούτου, τα προγράμματα εφαρμόζουν αλγόριθμους για την αυτοματοποίηση καλά καθορισμένων δραστηριοτήτων ή διαδικασιών. Η Μηχανική Μάθηση, ωστόσο, βασίζεται σε μια διαφορετική προσέγγιση, με τη βασική ιδέα να είναι ότι κάνουν τους υπολογιστές να μαθαίνουν από παραδείγματα. Κατά συνέπεια, αντί να εξηγεί βήμα - βήμα πώς να λύσει ένα πρόβλημα, τα μοντέλα Μηχανικής Μάθησης μαθαίνουν ποια είναι η σωστή έξοδος βασισμένα σε συγκεκριμένα δεδομένα εισόδου ή την διαδικασία βελτιστοποίησης.

Η υλοποίηση/ενσάρκωση και αξιοποίησή της ως ένα νέο υπολογιστικό μοντέλο κατέστη δυνατή μόνο χάρη στην πρόοδο του τεχνολογικού υλικού των υπολογιστών. Ο λόγος για αυτό είναι ότι τα συστήματα μηχανικής μάθησης απαιτούν μεγάλα σύνολα δεδομένων και μεγάλες

μνήμες καθώς και πόρους επεξεργασίας (Mich, 2020a). Για παράδειγμα, η υιοθέτηση ((τσιπ επιταχυντή)) σχεδιασμένα για μηχανική μάθηση μείωσαν την χρόνο εκπαίδευσης συστημάτων ταξινόμησης εικόνων με βάση το ImageNet από 6,2 λεπτά το 2018 σε 47 δευτερόλεπτα 2020. Εκτός από την εκμάθηση με παράδειγμα, υπάρχουν πολλά άλλα μοντέλα ή αλγόριθμοι μηχανικής μάθησης και μπορεί να ταξινομηθούν σε τρεις κύριες κατηγορίες: υπό επίβλεψη, χωρίς επίβλεψη και βασισμένη στην ενίσχυση. Μερικά συστήματα μηχανικής μάθησης συνδυάζουν τα πρότυπα των διαφορετικών ειδών ή συμπεριλαμβάνει νέες ιδέες, όπως αυτή της προσήλωσης (Bollinger, 2021). Καθορίζοντας ποιο είναι το πιο κατάλληλο για μια δεδομένη εργασία είναι συνήθως ένα από τα σημαντικότερα προβλήματα. Η λύση του απαιτεί εμπειρογνώμοσύνη και γνώση τόσο στην μηχανική μάθηση όσο και στο αντικείμενο.

Ένα σημαντικό κίνητρο για την έρευνα της μηχανικής μάθησης και των εφαρμογών της αφορούσε τα μεγάλα δεδομένα(Big Data). Η ποικιλία των πηγών εισόδου, των συσκευών αποθήκευσης και των συστημάτων του διαδικτύου συνέβαλαν στην αδιάκοπη δημιουργία μεγάλων συνόλων ή δεξαμενών δεδομένων. Διαφορετικά σε μορφή, τύπο, ποιότητα και δομή, αυτά τα δεδομένα, ή μεγάλα δεδομένα ή και μεγάλα δεδομένα στον τουρισμό έχει τροφοδοτήσει την ανάγκη για την εξόρυξη και την ερμηνευση των δεδομένων αυτών. Ωστόσο, οι παραδοσιακές τεχνικές επεξεργασίας δεδομένων, τα μοντέλα βάσεων δεδομένων και οι λειτουργίες του συστήματος πληροφοριών αντιμετωπίζουν πολλές προκλήσεις όταν αντιμετωπίζουν τα σύνολα δεδομένων που είναι πολλές τάξεις μεγέθους μεγαλύτερες από το συνηθισμένο.

Το AI και η μηχανική μάθηση όχι μόνο εκμεταλλεύονται τα μεγάλα δεδομένα, αλλά προσφέρουν και νέες λύσεις μέσω μιας αμφίδρομης σχέσης. Μέσω αυτής, η πρόοδος στην τεχνητή νοημοσύνη και την μηχανική μάθηση συνέβαλε σε μια νέας τομής έρευνα και εφαρμογής γνωστή ως Data Science (DS) (Murtagh & Devlin, 2018). Το Data Science στοχεύει στην ερμηνεία των διαθέσιμων δεδομένων, αποκαλύπτοντας κρυφές σχέσεις και μοτίβα. Ως διεπιστημονικό πεδίο, τα έργα DS απαιτούν μια εξειδικευμένη προσέγγιση σε στατιστική, Μαθηματικά, Πληροφορική, Μηχανική Μάθηση, οπτικοποίηση δεδομένων, δεξιότητες επικοινωνίας και παρουσίασης. Επιπλέον, όλες αυτές οι δυνατότητες πρέπει να χρησιμοποιούνται σε διαφορετικούς τομείς, καθιστώντας την τεχνογνωσία κλάδο μιας υποχρεωτικής πτυχής για κάθε ορθή και επιτυχή δραστηριότητα/ενέργεια αξιοποίησης και διαχείρισης δεδομένων (Chauhan & Σουντ, 2021). Για να χρησιμοποιηθεί ως είσοδος για αλγορίθμους μηχανικής μάθησης, τα δεδομένα πρέπει πρώτα να προεπεξεργαστούν (π. χ. να ταξινομηθούν ή να επισημανθούν), συνήθως με τη βοήθεια πληροφοριών του συγκεκριμένου κλάδου. Στη συνέχεια, η έξοδος των αλγορίθμων μηχανικής μάθησης μπορούν να χρησιμοποιηθούν για την υποστήριξη διαδικασιών βασισμένων στη γνώση, όπως σε συστήματα επιχειρηματικής ευφυΐας, συστήματα διαχείρισης πελατειακών σχέσεων και σε πολλές άλλες εφαρμογές. (Sharker, 2021). Οι κλασικές αρχές των συστημάτων πληροφοριών προτείνουν ότι,

οι λύσεις μηχανικής μάθησης και Data Science επιτρέψαν στις εταιρείες και τους οργανισμούς να βελτιώσουν τις επιχειρηματικές τους διαδικασίες, να μεταμορφώσουν τα επιχειρηματικά τους μοντέλα και να υποστηρίξουν νέα επιχειρηματικά μοντέλα. Οι μεγάλες εταιρείες τεχνολογίας οι λεγόμενες FAANGs (Facebook, Apple, Amazon, Netflix και Google), είναι τα σαφέστερα παραδείγματα για το πώς τα δεδομένα μπορούν να χρησιμοποιηθούν για τη δημιουργία επιχειρηματικής αξίας. Πολλές από τις πρόσφατες επιτυχίες τους οφείλονται σε υπηρεσίες που χρησιμοποιούν λύσεις AI και big data.

3 AI Για Μεγάλα Δεδομένα

Σε μια προσπάθεια να χαρακτηρίσουμε τα μεγάλα δεδομένα, και να απαντήσουμε στην ερώτηση "πόσο μεγάλα είναι τα μεγάλα δεδομένα;" ένας γενικότερη έννοια εισάγει τρεις συντελεστές: όγκος, ταχύτητα και ποικιλία. Οι τρεις αυτοί συντελεστές συσχετίζονται με την κατανόηση του γιατί τα μοντέλα AI και τα εργαλεία είναι απαραίτητα για τη μόχλευση μεγάλων δεδομένων.

- Ένας μεγάλος όγκος δεδομένων μπορεί να συλλεχθεί από διάφορες πηγές και εργαλεία, συμπεριλαμβανομένων επιχειρηματικών συναλλαγών, έξυπνων συσκευών (IoT), βιομηχανικού εξοπλισμού, κινητές συσκευές, αναγνώστες προσδιορισμού ραδιοσυχνότητας (RFID), ασύρματος αισθητήρας δικτύων, κοινωνικά δίκτυα κλπ. Το μέγεθος και ο αριθμός των διαθέσιμων συνόλων δεδομένων έχουν αυξηθεί πέρα από τις προσδοκίες, αλλά αποθηκεύοντάς το με νέα και φθηνότερη αποθήκευση οι επιλογές και οι πλατφόρμες έχουν γίνει επίσης πολύ πιο εύκολες τώρα απ' ό,τι ήταν στο παρελθόν.
- Τα δεδομένα ταχύτητας παράγονται και χρησιμοποιούνται με μεγάλη ταχύτητα. Κυβερνοσωματικά συστήματα, ενσωμάτωση στοιχείων υλικού και λογισμικού σε σύνθετα πλαίσια (π.χ. Σε αυτοκίνητα χωρίς οδηγό, ρομποτική) και πολλές άλλες ευρέως χρησιμοποιούμενες εφαρμογές (π.χ. Chatbots ή προσωπικοί βοηθοί) παράγουν γρήγορες ροές δεδομένων και συχνά απαιτούν επεξεργασία σχεδόν σε πραγματικό χρόνο.
- Ποικιλία δεδομένων έρχεται σε διάφορες μορφές. Μόνο ένα μικρό κλάσμα από αυτά αποτελούνται από δομημένα δεδομένα και μπορούν να διαχειριστούν με κλασικές βάσεις δεδομένων. Τα δεδομένα που είναι αδόμητα με τη μορφή κειμένων, εικόνων, βίντεο, αισθητήρων και δεδομένα ticker κλπ. είναι τεράστια σε αριθμό, π.χ. μια ιστοσελίδα περιλαμβάνει συνήθως έναν συνδυασμό από τις παραπάνω μορφές δεδομένων.

Άλλα μοντέλα για τα μεγάλα δεδομένα περιλαμβάνουν μια ευρύτερη ποικιλία συντελεστών. Μεταξύ των πιο συχνών προστιθέμενων παραμέτρων είναι η εγκυρότητα ακολουθούμενη από την μεταβλητότητα:

- Τα δεδομένα εγκυρότητα πρέπει να ελέγχονται ποιοτικά. Από τα πρώτα προγράμματα υπολογιστών, η συνάφεια της ποιότητας των δεδομένων έχει επισημανθεί μέσω του όρου, "Garbage In Garbage Out" (GIGO). Στην πραγματικότητα, τα προβλήματα στα δεδομένα εισόδου επηρεάζουν την έξοδο, και όλες οι παράμετροι που χαρακτηρίζουν τα μεγάλα δεδομένα καθιστούν την ποιότητα των δεδομένων ένα δύσκολο και πολύπλευρο πρόβλημα. Το γεγονός ότι υπάρχουν πολλά σημεία δεδομένων, πολλές πηγές, πολλές αλλαγές, πολλές μορφές κ.λπ. αποτελεί πρόκληση για κάθε εφαρμογή που χρησιμοποιεί μεγάλα δεδομένα. Κυρίως για τα υπολογιστικά μοντέλα μηχανικής μάθησης που χρησιμοποιούν τέτοια δεδομένα για αυτοεκπαίδευση.
- Οι ροές δεδομένων μεταβλητότητας αλλάζουν συχνά και ποικίλλουν πολύ. Αυτό το χαρακτηριστικό είναι συχνό σε δεδομένα που σχετίζονται με τον τουρισμό. για παράδειγμα, οι τουριστικές ροές αλλάζουν τόσο δραματικά μερικές φορές , ανάλογα με τα γεγονότα, τις καιρικές συνθήκες, την ημερήσια ή εποχιακή προσφορά και τη ζήτηση γενικότερα. Τέτοιες αλλαγές πρέπει να αντιμετωπίζονται άμεσα για την αντιμετώπιση των φορτίων αιχμής των δεδομένων, καθώς και για την εξαγωγή πολύτιμων γνώσεων που θα καθοδηγήσουν τις επακόλουθες αποφάσεις των εταιρειών και όχι μόνο.

Τα μεγάλα δεδομένα περιγράφονται επίσης χρησιμοποιώντας τον όρο "αξία" ως βασική διάσταση για να υπογραμμίζει το γεγονός ότι τα δεδομένα πρέπει να ερμηνεύονται λαμβάνοντας υπόψη τη φύση τους δηλαδή, στατιστική, υποθετική κ.λπ., αλλά το πιο σημαντικό είναι ότι αξία σημαίνει ότι "το να έχουμε πρόσβαση σε μεγάλα δεδομένα δεν είναι καλό αν δεν μπορούμε να τα μετατρέψουμε σε αξία" (Marr, 2015). Τα μεγάλα δεδομένα μπορούν επίσης να θεωρηθούν ως ένας κινούμενος στόχος, ανάλογα με τους διαθέσιμους υπολογιστικούς πόρους και δυνατότητες. Μια τάση που μοιράζεται κάθε τεχνολογική καινοτομία: με άλλα λόγια, οι νέες τεχνολογίες, και ειδικότερα η πρόοδος του AI, επεκτείνουν συνεχώς το πεδίο των εφαρμογών μεγάλων δεδομένων.

Η συνεργασία της τεχνητής νοημοσύνης και των μεγάλων δεδομένων μπορεί να περιγραφεί με βάση τα πέντε κύρια βήματα ενός γενικού μοντέλου διαδικασίας Digital Science:

- Διαμόρφωση προβλήματος
- Συλλογή δεδομένων
- Καθαρισμός δεδομένων
- Επεξεργασία δεδομένων
- Εκμετάλλευση δεδομένων

3.1 AI για διαμόρφωση λύσεων προβλημάτων

Αδιαμφισβήτητο αυτό είναι το πιο κρίσιμο βήμα , καθώς τα αποτελέσματα εξαρτώνται σε

μεγάλο βαθμό από τα αρχικά ερωτήματα και τους δηλωμένους στόχους. Τα μεγάλα δεδομένα από μόνα τους δεν έχουν αξία εκτός αν μπορείτε να εξαγάγετε σημαντικά αποτελέσματα "αξία" από αυτά. Για το σκοπό αυτό, είναι απαραίτητο να θέσετε πρώτα ένα βασικό ερώτημα, τη διαμόρφωση ενός επιχειρηματικού προβλήματος ή τον καθορισμό της στρατηγικής μιας επιχείρησης. Από την άλλη πλευρά, η πειθαρχία του κλάδου των πληροφοριακών συστημάτων διδάσκει ότι, οι λύσεις τεχνητής νοημοσύνης μπορούν να προτείνουν νέους τρόπους χρήσης των μεγάλων δεδομένων για την προσθήκη αξίας στα προϊόντα ή τις υπηρεσίες μιας επιχείρησης. Τέλος, οι τεχνικές τεχνητής νοημοσύνης μπορούν επίσης να υποστηρίξουν την δημιουργικότητα προκειμένου να διερευνηθούν νέες καινοτόμες ιδέες και να αντιμετωπιστούν τα αναφερθέντα πρόβλημα με τη χρήση των μεγάλων δεδομένων (Boden, 1998).

3.2 AI για τη συλλογή δεδομένων

Η τεχνητή νοημοσύνη συμβάλλει στην "καταγραφή δεδομένων" του πραγματικού κόσμου, δηλαδή στη διαδικασία των αντικείμενων, των ανθρώπων και των διαδικασιών που "μετασχηματίζονται" σε ψηφιακά δεδομένα (Southerton, 2020). Για παράδειγμα, οι διαδικτυακές διαδράσεις επιτρέπουν τη συλλογή δεδομένων που είναι χρήσιμα στη μελέτη της συμπεριφοράς και της κοινωνικής ζωής. Η διαδικασία εξαγωγής περιεχομένου από έναν ιστότοπο ή από μια πλατφόρμα κοινωνικής δικτύωσης ονομάζεται επίσης απόξεση δεδομένων, και ορισμένα από τα διαθέσιμα εργαλεία ισχύουν ακόμη και στην επεξεργασία γλώσσας (NLP) (Lane et al., 2019) ή υπολογιστικής όρασης (Brownlee, 2019). Σε ορισμένες περιπτώσεις, τα μεγάλα δεδομένα ανήκουν σε διαφορετικούς οργανισμούς ή εταιρείες όπως συμβαίνει με τα έργα "έξυπνων πόλεων". Νομική ιδιοκτησία, θέματα προστασίας της ιδιωτικής ζωής αλλά και ασφάλειας, υπάρχουν τρία θέματα που πρέπει να αντιμετωπιστούν σωστά προκειμένου να αποκτηθούν και να διαχειριστούν τα απαιτούμενα δεδομένα. Αντιστρόφως, για τις κυβερνήσεις ή τις μεγάλες επιχειρήσεις, το ερώτημα που πρέπει να τεθεί είναι "Ποιος θα πρέπει να αναλάβει τις πρωτοβουλίες για τα μεγάλα δεδομένα που επηρεάζουν τους διάφορους οργανισμούς;". Τέλος, αν και όχι τόσο δημοφιλής είναι οι τεχνικές τεχνητής νοημοσύνης που μπορούν να βοηθήσουν στον εντοπισμό νομικών απαιτήσεων ή απαιτήσεων προστασίας της ιδιωτικής ζωής (Zeni et al., 2015) ή να αναζητήσουν σχετικά δεδομένα.

3.3 AI για καθαρισμό και προετοιμασία δεδομένων

Σε κάθε αυτόματο σύστημα αποτελεί τεράστιο ζήτημα η ποιότητα των δεδομένων ειδικά όταν πρόκειται για μεγάλα δεδομένα. Ο καθαρισμός δεδομένων περιλαμβάνει σφάλματα και διπλότυπες εξαλείψεις, μια διεργασία που απαιτεί διαφορετικές μεθόδους για τις διάφορες μορφές εισόδου και αντιμετωπίζει διαφορετικά πρότυπα και κανόνες πρόσβασης. Σε

συγκεκριμένες περιπτώσεις, είναι απαραίτητο τα διαθέσιμα δεδομένα που προέρχονται από διαφορετικές πηγές να φιλτράρονται, να μετασχηματίζονται ή να ενσωματώνονται και τέλος να αποθηκεύονται σε διάφορα συστήματα. Το σύνολο της συλλογής των δεδομένων μιας εταιρείας - δομημένων και αδόμητων - ονομάζεται επίσης "λίμνη δεδομένων". Συνήθως αποθηκεύονται σε μια πλατφόρμα cloud, οι λίμνες δεδομένων είναι τύποι "αποθηκών" που συγκεντρώνουν αντίγραφα των δεδομένων ενός οργανισμού και, όπως και με τις αποθήκες, οι λίμνες δεδομένων πρέπει να διαχειρίζονται αρτίως ώστε να είναι σε θέση να ανακτούν τα σύνολα δεδομένων που είναι χρήσιμα για τα επόμενα βήματα. Συνήθως, τα καθαρισμένα δεδομένα πρέπει να προετοιμαστούν προκειμένου να χρησιμοποιηθούν ως δεδομένα εισόδου για το στάδιο της ανάλυσης ως εκ τούτου, το στάδιο αυτό είναι αυστηρά συνυφασμένο με το επόμενο στάδιο και περιγράφεται επίσης ως μέρος μιας διαδικασίας επεξεργασίας δεδομένων (Lakshmanan et al., 2020). Η τεχνητή νοημοσύνη, και ειδικότερα η μηχανική μάθηση, μπορεί να βοηθήσει σε πολλές από τις εργασίες που σχετίζονται με τον καθαρισμό και την προετοιμασία δεδομένων (Ilyas, 2020).

3.4 AI για επεξεργασία δεδομένων

Τα δεδομένα που προεπεξεργάστηκαν στο προηγούμενο βήμα μπορούν να διερευνηθούν και να αναλυθούν με τη χρήση αλγορίθμων μηχανικής μάθησης. Επιπλέον, ανάλογα με τον τύπο των συνόλων δεδομένων που εμπλέκονται στην ανάλυση και τους σχετικούς τομείς τεχνητής νοημοσύνης, υπάρχουν πολλές διαφορετικές προσεγγίσεις που μπορούν να υιοθετηθούν. Πρώτα απ' όλα, μπορούν να χρησιμοποιηθούν τεχνικές οπτικοποίησης δεδομένων για τη διερεύνηση των δεδομένων και την αναζήτηση ακραίων τιμών ή κανονικοτήτων, οι οποίες υποστηρίζονται επίσης από παραδοσιακά στατιστικά μοντέλα. Ο απώτερος στόχος είναι να καθοριστεί μια στρατηγική όπου μπορούν να εξαχθούν πολύτιμες γνώσεις από τα δεδομένα. Για παράδειγμα, εάν η είσοδος περιλαμβάνει μεγάλα σύνολα αξιολογήσεων προϊόντων προς αυτόματη βαθμολόγηση, ένα υποσύνολο αυτών μπορεί να υποστεί προεπεξεργασία με τη χρήση σημασιολογικών δικτύων, εννοιολογικών μοντέλων (Harmelen et al., 2010) ή άλλων γλωσσολογικών μεθόδων. Οι αλγόριθμοι μηχανικής μάθησης μπορούν στη συνέχεια να εφαρμοστούν για τον προσδιορισμό κανόνων βαθμολόγησης που αντιστοιχούν σε εντοπισμένα γλωσσικά πρότυπα ή εννοιολογικές σχέσεις. Εάν το σύνολο δεδομένων περιλαμβάνει εικόνες, από την άλλη πλευρά, τότε άλλοι τομείς της τεχνητής νοημοσύνης μπορούν να συνεισφέρουν. Το ίδιο ισχύει και για τις ροές δεδομένων IoT (Internet of Things) κ.λπ. Το πρόβλημα ωστόσο, της τεχνητής νοημοσύνης και των μεγάλων δεδομένων στον τουρισμό δεν έγκειται μόνο στην ίδια την επιλογή του τρόπου εφαρμογής της μεθόδου ή του συστήματος της τεχνητής νοημοσύνης, αλλά και στην πολυπλοκότητά της. Για το λόγο αυτό απαιτούνται ειδικοί στην ανάλυση δεδομένων στην τεχνητή νοημοσύνη και στον τομέα της εφαρμογής, γεγονός που αναδεικνύει την ανάγκη για μια διεπιστημονική και συνεργατική ομάδα για έργα μεγάλων δεδομένων. Το

θετικό είναι ότι πολλές βιβλιοθήκες και πλατφόρμες υποστηρίζουν τον χειρισμό μεγάλων δεδομένων και μεθόδων επιστημονικών δεδομένων , συμπεριλαμβανομένης της μηχανικής μάθησης (Krensky et al., 2021).

3.5 AI για εκμετάλλευση δεδομένων

Τα αποτελέσματα που προκύπτουν από την αναλύση μεγάλων δεδομένων μπορούν να χρησιμοποιηθούν για τη λήψη μέτρων ή αποφάσεων. (Sharda et al., 2021) Φαίνεται ότι δεν υπάρχουν όρια στις εφαρμογές των λύσεων που βασίζονται στην τεχνητή νοημοσύνη και πως κάθε τομέας και κάθε εργασία θα μπορούσε με τον ένα ή τον άλλο τρόπο να υποστηριχθεί από ευφυή συστήματα που χρησιμοποιούν μεγάλα δεδομένα. Μεταξύ των πιο πρόσφατων και διαδεδομένων εργαλείων που βασίζονται στην τεχνητή νοημοσύνη, έχουν αφήσει το στίγμα τους τα συστήματα συστάσεων (π.χ. στο Amazon), οι εικονικοί βοηθοί (π.χ. η Siri, που εφαρμόζει συστήματα επεξεργασίας γλώσσας (NLP) και αναγνώρισης ομιλίας), τα συστήματα ανίχνευσης απάτης (π.χ. αυτά που χρησιμοποιούνται για τις πιστωτικές κάρτες) και η απεικόνιση της κυκλοφορίας (π.χ. στους χάρτες της Google, χρησιμοποιώντας διάφορες πηγές δεδομένων από smartphones ή GPS). Η επαυξημένη και η εικονική πραγματικότητα προσφέρουν επίσης νέους τρόπους επικοινωνίας. Επιπρόσθετα παραδείγματα περιλαμβάνουν την όραση υπολογιστών που υποστηρίζει την αναγνώριση προσώπου, τη ρομποτική που εφαρμόζεται σε συστήματα αυτοκινήτων χωρίς οδηγό κ.ο.κ. Εν αντιθέσει προβλήματα και κίνδυνοι είναι εκείνα που σχετίζονται με την επεξηγηματική δύναμη και τα ζητήματα λογοδοσίας, καθώς πολλοί αλγόριθμοι μηχανικής μάθησης αποτυγχάνουν να δικαιολογήσουν την έξοδό τους και με τη σειρά τους να καθορίσουν την ευθύνη των εμπλεκόμενων υποκειμένων (Garigliano & Mich, 2019).

4 AI και μεγάλα δεδομένα στον τουρισμό

Ο τουρισμός είναι ένας πολυδιάστατος και σύνθετος τομέας και όλα τα χαρακτηριστικά του συμβάλλουν στο να τον καταστήσουν το ιδανικό πεδίο για την εφαρμογή συστημάτων τεχνητής νοημοσύνης που αξιοποιούν μεγάλα σύνολα δεδομένων. Πρώτον, λόγω του μεγάλου αριθμού εμπλεκόμενων ανθρώπων στις διαδικασίες και δραστηριότητες, τα δεδομένα του τουρισμού είναι συχνά μεγάλα δεδομένα. Επιπλέον, δεν φαίνεται να υπάρχουν όρια στους τύπους των πηγών δεδομένων και των τεχνολογιών τεχνητής νοημοσύνης που θα μπορούσαν να είναι χρήσιμες για τη διερεύνηση και τη βελτίωση τόσο της τουριστικής προσφοράς όσο και της ζήτησης, συμπεριλαμβανομένων των στρατηγικών μάρκετινγκ και της διαχείρισης του υπερτουρισμού. Οι τεχνολογίες αυτές θα μπορούσαν ακόμη και να διευκολύνουν την ανάκαμψη από περιόδους κρίσης και πολλά άλλα. Συνεπώς, παρά το γεγονός ότι δεν είναι πρακτικά εφικτό να επανεξεταστούν όλες οι σημαντικές εμπειρίες και τα έργα στον τομέα αυτό, μπορούν να

δοθούν ορισμένες κατευθυντήριες γραμμές:

- Τα projects τεχνητής νοημοσύνης και μεγάλων δεδομένων μπορούν να επωφεληθούν από "εξωτερικές" πηγές δεδομένων, δηλαδή όχι μόνο από εκείνες που παραδοσιακά λαμβάνονται υπόψη στις τουριστικές εφαρμογές (Li et al., 2018). Για παράδειγμα, εκτός από το περιεχόμενο που δημιουργείται από χρήστες (UGC), οι μετεωρολογικές προβλέψεις και τα δεδομένα για την κλιματική αλλαγή θα μπορούσαν να χρησιμοποιηθούν για την αντιμετώπιση του προβλήματος της εποχικής τουριστικής προσφοράς. Επιπλέον, και οι δύο πτυχές της εμπειρογνωμοσύνης σε θέματα διαπραγμάτευσης και της γνώσης εναρμόνισης δεδομένων είναι εξίσου κρίσιμες για επιτυχημένες πρωτοβουλίες όπως και για τα projects έξυπνων πόλεων.
- Καινοτόμες στρατηγικές για τον τουρισμό μπορούν να σχεδιαστούν μόνο με νέα οράματα και ιδέες. Οι δημιουργικές τεχνικές μπορούν στη συνέχεια να εφαρμοστούν λαμβάνοντας υπόψη τις γνώσεις που αποκτώνται από την ανάλυση μεγάλων συνόλων δεδομένων. Εναλλακτικά, θα μπορούσαν να χρησιμοποιηθούν νέοι τρόποι οπτικοποίησής τους μεταξύ των πιο πρόσφατων αρχών της κυκλικής οικονομίας, της βιώσιμης ανάπτυξης και της περιβαλλοντικής οικονομίας, μπορούν να αποτελέσουν έμπνευση για τις κυβερνήσεις και την διαχείριση προορισμών, τις τοπικές τουριστικές επιτροπές και τους ενδιαφερόμενους φορείς του τουρισμού γενικότερα.
- Όσον αφορά τις πολλές τεχνολογίες συλλογής δεδομένων, αυτές που προσδιορίζουν την ανθρώπινη θέση ή εξίσου κάθε είδους φαινόμενο είναι πολύ σημαντικές για την τουριστική γεωγραφία. Οι νέες τεχνολογίες και οι τεχνικές παρακολούθησης (δράσης), όπως το GPS, τα κινητά τηλέφωνα, οι αισθητήρες, καθώς και τα δεδομένα από drones, μπορούν να χρησιμοποιηθούν για τη διαχείριση ζητημάτων υπερτουρισμού, π.χ. για τη βελτιστοποίηση διαδρομών, τη διαχείριση της κυκλοφορίας, την εξατομικευμένη ξενάγηση κ.λπ. (Mich, 2020b).
- Η τεχνητή νοημοσύνη και τα μεγάλα δεδομένα θα μπορούσαν επίσης να συμβάλουν στη μείωση του κατακερματισμού των τουριστικών (κινητών) εφαρμογών όπως, συχνά, πολλές εφαρμογές είναι πολύ εξειδικευμένες και δεν είναι αρκετά εξατομικευμένες. Μια από τους λόγους για αυτό είναι ότι το καθένα βασίζεται σε ένα ενιαίο σύνολο δεδομένων και δεν είναι σχεδιασμένο για να ικανοποιεί τις πραγματικές απαιτήσεις των ενδιαφερομένων.
- Η τεχνητή νοημοσύνη και τα μεγάλα δεδομένα θα μπορούσαν επίσης να συμβάλουν στη μείωση του κατακερματισμού των εφαρμογών (κινητών) τουριστικών υπηρεσιών, καθώς συχνά πολλές εφαρμογές είναι πολύ συγκεκριμένες και όχι αρκετά εξατομικευμένες. Ένας από τους λόγους γι' αυτό είναι ότι κάθε μία βασίζεται σε ένα μόνο σύνολο δεδομένων και δεν έχει σχεδιαστεί για να ανταποκρίνονται στις πραγματικές απαιτήσεις των ενδιαφερομένων.

- Οι εξελίξεις στην τεχνητή νοημοσύνη πρέπει να παρακολουθούνται συνεχώς για την περαιτέρω βελτίωση οποιασδήποτε από τις κλασικές εφαρμογές των μεγάλων δεδομένων. Για παράδειγμα, τα συστήματα που αναλύουν UGC για συστάσεις πελατών ή για παρακολούθηση της φήμης του ιστού μπορούν να επωφεληθούν από ορισμένες από τις προόδους της NLP, αξιοποιώντας παράλληλα τεχνικές ταξινόμησης εικόνων, αναγνώρισης προσώπων, ανάλυσης βίντεο και αναγνώρισης φωνής, εκτός από την πιο πρόσφατα αναπτυχθείσα αναγνώριση συναισθήματος.
- Δεδομένου του κεντρικού ρόλου της επικοινωνίας στον τουρισμό, εκτός από την ανάλυση περιεχομένου, είναι δυνατόν να εφαρμοστούν τα λεγόμενα συστήματα TN που δημιουργούν τα πάντα για την υποστήριξη παραγωγής περιεχομένου. Το κείμενο, ο ήχος και οι εικόνες που παράγονται από αυτά τα συστήματα δεν αναγνωρίζονται εύκολα ως αυτόματα παραγόμενα (και απαιτούν νέες λύσεις για τα πνευματικά δικαιώματα). Όσον αφορά την ποιότητα του περιεχομένου, η τεχνητή νοημοσύνη μπορεί επίσης να χρησιμοποιήσει σύνολα δεδομένων για τον έλεγχο των γεγονότων και τη βαθιά ανίχνευση πλαστών δεδομένων.
- Τέλος, οι νέες και καινοτόμες εφαρμογές της τεχνητής νοημοσύνης και των μεγάλων δεδομένων μπορούν επίσης να ωφελήσουν τα παραδοσιακά (αναλογικά) έγγραφα κάθε είδους. Υπάρχουν μεγάλες ποσότητες εγγράφων που, όταν ψηφιοποιηθούν, θα μπορούσαν να ενσωματωθούν σε υφιστάμενα σύνολα δεδομένων, προσθέτοντας πολύτιμες γνώσεις για τουριστικούς προορισμούς, π.χ. όσον αφορά την πολιτιστική και ιστορική κληρονομιά τους. Οι πρωτοβουλίες Crowdsourcing, υπό αυτή την έννοια, θα μπορούσαν επίσης να βοηθήσουν στη δημιουργία κοινοτήτων, οι οποίες είναι καθοριστικές για τον τουρισμό.

Εν κατακλείδι, είναι σημαντικό να τονιστεί ότι ο θόρυβος γύρω από τη λήψη απροσδόκητων αποτελεσμάτων από την εξόρυξη μεγάλων συνόλων δεδομένων μέσω της TN είναι εμφανής σε ακαδημαϊκές και επαγγελματικές πηγές. Ωστόσο, τα αποτελέσματα αυτά απαιτούν σημαντικές επενδύσεις και τεχνογνωσία, καθώς και προσοχή στις κοινωνικές και πολιτικές επιπτώσεις σε έναν συστημικό και δικτυωμένο τομέα όπως ο τουρισμός (Baggio et al., 2010).

5 Εξόρυξη δεδομένων

Αναμφισβήτητα, οι πιο εκτεταμένες αλλαγές τις τελευταίες δεκαετίες οφείλονται στην ψηφιακή της κοινωνίας μας, η οποία έχει επίσης αλλάξει μαζικά την τουριστική βιομηχανία. Η Nora και η Minc (1980) μίλησαν επίσης για το "computerization of society" το 1978 (Egger, 2015). Προς το παρόν, η δημοτικότητα των κοινωνικών οι πλατφόρμες των μέσων ενημέρωσης, ειδικότερα, οδηγούν σε αλλαγές στις κοινωνικές δομές (Fuchs, 2021) και, ως εκ τούτου, τα ταχέως αναπτυσσόμενα σύνολα δεδομένων απαιτούν νέα μεθοδολογικά προσεγγίσεις για τη διερεύνησή τους.

"Τα μεγάλα δεδομένα είναι το νέο πετρέλαιο" είναι ένα συχνά αναφερόμενο σύνθημα, και μάλιστα αυτή η αναλογία φαίνεται να επιβεβαιώνεται σε διάφορα επίπεδα. Αν και τα δεδομένα είναι ανανεώσιμα, σε αντίθεση με ορυκτά καύσιμα, Thorp (2012) αναφέρει τον κίνδυνο "ρύπανσης δεδομένων" εκτός από τη τεράστια οικονομική και εμπορική σημασία των δεδομένων. Ο Corbett (2018) υποστηρίζει ότι στην περίπτωση των ορυκτών καυσίμων, οι μακροπρόθεσμες επιπτώσεις, οι δαπάνες που πραγματοποιήθηκαν και η δυσκολία και τα συναφή προβλήματα απομάκρυνσης από αυτά δεν έχουν αντιμετωπιστεί και το ίδιο φαίνεται να ισχύει και για τα δεδομένα. Έτσι, συνιστάται προσοχή ως προς το πώς παράγουμε, επεξεργαζόμαστε και χρησιμοποιούμε δεδομένα (για περισσότερες πληροφορίες, ανατρέξτε στο κεφάλαιο "δεδομένα Επιστήμη Και Ηθική"). Σύμφωνα με αυτό, νέες μέθοδοι και προσεγγίσεις ανάλυσης δεδομένων θεωρούνται ως διαταρακτικές καινοτομίες που δημιουργούν επιστημολογικές προκλήσεις.

Η κοινωνία μας εξαρτάται από τη δημιουργία, τη συλλογή, την επεξεργασία και τη χρήση των δεδομένων και για τον καθένα μας τα δεδομένα έχουν γίνει κεντρικό στοιχείο στην καθημερινή μας ζωή (Τζόνσον, 2014). Ενώ εμείς, ως παραλήπτες προσπαθούμε τουλάχιστον να επιλέξουμε δεδομένα σκόπιμα, είτε με έρευνα στον ιστό, περιήγηση σε κανάλια κοινωνικών μέσων ή λήψη περιεχόμενο πολυμέσων, ταυτόχρονα παράγουμε ασυνείδητα ένα πλήθος δεδομένων. Εμείς αφήστε τα ψηφιακά αποτυπώματα με βάση την ηλεκτρονική μας έρευνα, δημιουργώντας μετα-δεδομένα, όπως Συντεταγμένες GPS και ημερομηνίες, όταν ανεβάζουμε φωτογραφίες ή συνάγουμε συμπεράσματα σχετικά με τις δικές μας προτιμήσεις με likes και σχόλια σε αναρτήσεις στα μέσα κοινωνικής δικτύωσης. Εκτός αυτού, οι αισθητήρες, τα wearables και το internet of things (IoT) συλλέγουν και μοιράζονται επίσης δεδομένα αυτόματα (Gretzel et al., 2015α, 2015B * Li et al., 2018; Xiang & Fesenmaier, 2017α).

Ο τουρισμός ήταν πάντα μια βιομηχανία έντασης πληροφοριών και είναι σήμερα μετατροπή σε μια βιομηχανία που βασίζεται σε δεδομένα. Η ποσότητα και η ποικιλία των διαθέσιμων δεδομένων σήμερα επιτρέπουν αδιανόητες δυνατότητες ανάμειξης και συνδυασμού δεδομένων, με αποτέλεσμα Νέα εξαγόμενα χαρακτηριστικά που μπορούν να τροφοδοτηθούν σε μοντέλα μηχανικής μάθησης για να αποκαλύψετε μοτίβα που ήταν

προηγούμενως άγνωστα. Ως εκ τούτου, συλλέγονται δεδομένα, αναμειγνύονται , και αναλύονται για την καλύτερη κατανόηση των πελατών, των ανταγωνιστών και άλλων μετοχών. (Xiang & fesenmaier, 2017b). 68 R.Egger et al.

Βασικά, μπορεί να γίνει διάκριση μεταξύ στατικών και δυναμικών δεδομένων. Στον τουρισμό τα στατικά δεδομένα περιλαμβάνουν συντακτικά κείμενα όπως περιγραφές προορισμών και αξιοθέατα ή οποιοδήποτε μέσο μάρκετινγκ από τουριστικούς παρόχους. Συνήθως, τέτοια δεδομένα μπορεί να είναι ανακτήθηκαν από τους αντίστοιχους ιστότοπους ή από τις πλατφόρμες μεσαζόντων. Αυτά τα δεδομένα σπάνια αλλάζουν και είναι διαθέσιμο σε όλες τις μορφές μέσων, που κυμαίνονται από περιγραφή κειμένου- φωτογραφίες μέχρι και εικόνες και αρχεία ήχου και βίντεο, που συνήθως υπάρχουν σε πολύ αδόμητες και κατακερματισμένες μορφές. Στις περισσότερες περιπτώσεις, προκειμένου να ληφθούν αυτά τα δεδομένα για περαιτέρω επεξεργασία και ανάλυση, πρέπει πρώτα να ανιχνευτεί. Δυναμικά δεδομένα, από την άλλη πλευρά, αλλάζει συχνά και περιλαμβάνει δεδομένα όπως τιμές και διαθεσιμότητα, τα οποία, όπως ένας εμπειρικός κανόνας, μπορεί να ληφθεί μόνο μέσω απόξεσης ιστού. Ένας άλλος τύπος είναι τα δυναμικά δεδομένα όπου περιλαμβάνουν δεδομένα που έχουν ήδη προεπεξεργαστεί και προετοιμαστεί για περαιτέρω χρήση, για παράδειγμα, πληροφορίες σχετικά με τις τουριστικές αγορές ή στατιστικούς δείκτες (π. χ. αριθμός αφίξεων ή διανυκτερεύσεων) που διατίθενται από οργανισμούς ή κυβερνητικά ιδρύματα είτε με τη μορφή ανοικτών δεδομένων είτε μέσω προγράμματος εφαρμογής- διεπαφής *ming* (API). Με τη δημοτικότητα των καναλιών κοινωνικών μέσων, που δημιουργούνται από χρήστες το περιεχόμενο (UGC) έχει αναδειχθεί ως ένας ξεχωριστός τύπος δεδομένων που μπορεί να είναι στατικός ή δυναμικός (Pantano et al., 2017). Ανάλογα με το κανάλι κοινωνικών μέσων, το περιεχόμενο τείνει να είναι βαρύ κείμενο, εικόνα ή βίντεο και οι πλατφόρμες ρυθμίζουν την πρόσβαση σε αυτά τα δεδομένα. Ενώ τα δεδομένα του Twitter μπορούν να προσπελαστούν σχετικά εύκολα μέσω API, Facebook, Instagram ή Τα δεδομένα του TripAdvisor είναι ελάχιστα προσβάσιμα μέσω API, ώστε να διασφαλίζεται η προστασία των χρηστών απορρήτου.

Ανεξάρτητα από τον τύπο δεδομένων, μετα-δεδομένα, περιγράφοντας τα αρχικά δεδομένα με περισσότερες λεπτομέρειες, είναι σχεδόν πάντα παρόν. Για παράδειγμα, ένα ξενοδοχείο μπορεί να περιγραφεί λεπτομερέστερα με βάση σχετικά με τις γεωγραφικές συντεταγμένες και τις διευθύνσεις του, καθώς και τις ώρες λειτουργίας και την αξιολόγηση με αστέρια, ενώ οι πληροφορίες τιμών μπορούν να ισχύουν για συγκεκριμένες χρονικές περιόδους ή να ισχύουν μόνο για ορισμένες ομάδες πελατών. Όταν πρόκειται για αναρτήσεις ιστολογίου, αυτές επισημαίνονται με μετα-ετικέτες όπως το όνομα του συγγραφέα, την ημερομηνία δημιουργίας και, ενδεχομένως μια καθορισμένη κατηγορία. Αυτά τα μετα-δεδομένα συχνά επιτρέπουν τη σύνδεση συνόλων δεδομένων και, ως εκ τούτου, νέα ενδιαφέροντα και πλούσια σε πληροφορίες χαρακτηριστικά μπορεί να εξαχθούν. Δεδομένου ότι τα μετα-δεδομένα περιγράφουν περαιτέρω άλλα δεδομένα.

5.1 Ανοικτά Δεδομένα

Σήμερα, έχουμε περισσότερα δεδομένα από ποτέ στα χέρια μας. Η τάση προς τα ανοικτά δεδομένα συνεχίζει να επιταχύνει αυτή την έννοια, και στον τουρισμό, επίσης, όλο και περισσότεροι οργανισμοί, φορείς και κυβερνήσεις καθιστούν τα δεδομένα δημοσίως διαθέσιμα (Pesonen & Lampi, 2016). Οι Longhi, Titz και Viallis (2014) μαζί με τους Pesonen και Lampi (2016) σημειώνουν ότι η τουριστική βιομηχανία έχει αναλάβει πρωτοποριακό ρόλο στην παροχή και χρήση ανοικτών δεδομένων, τα οποία περιλαμβάνουν πληροφορίες για την κυκλοφορία, χωρικά δεδομένα, δεδομένα για εκδηλώσεις, στατιστικά στοιχεία και βασικά δεδομένα, καθώς και πληροφορίες για τον πολιτισμό και τα σημεία ενδιαφέροντος, για να αναφέρουμε μόνο μερικά από αυτά. Τα ανοικτά δεδομένα είναι ελεύθερα διαθέσιμα σε όλους και καθίστανται προσβάσιμα μέσω του Διαδικτύου, χωρίς να υπόκεινται σε περιοριστικούς όρους χρήσης (Fermoso et al., 2015- Longhi et al., 2014). Έτσι, η γενική πρόθεσή τους είναι να προωθήσουν την καινοτομία (Xiang & Fesenmaier, 2017b), να βελτιώσουν τη δημόσια ευημερία και να χρησιμοποιήσουν τους δημόσιους πόρους πιο αποτελεσματικά (Johnson et al., 2012). Καθώς ο τουρισμός είναι μια πολύ σύνθετη και αλληλένδετη βιομηχανία, η οποία θέτει υψηλές απαιτήσεις για επικοινωνία και συντονισμό μεταξύ όλων των ενδιαφερόμενων φορέων, τα ανοικτά δεδομένα μπορούν να συμβάλουν σε αυτό, καθιστώντας τα δεδομένα διαθέσιμα για την έρευνα στον τουρισμό και τη φιλοξενία, π.χ. για την κατανόηση των αναγκών των τουριστών και τη βελτιστοποίηση της ποιότητας των εμπειριών τους (Pantano et al, 2019). Ως εκ τούτου, η διαθεσιμότητα και η χρήση ανοικτών δεδομένων έχει συζητηθεί και υποστηριχθεί εντατικά, ιδίως στο πλαίσιο των έξυπνων πόλεων ή των έξυπνων προορισμών (Gretzel et al., 2015a, 2015b).

Για την αντιμετώπιση αυτού του προβλήματος, έχουν αρχίσει να δημιουργούνται σε όλο τον κόσμο πλατφόρμες που λειτουργούν από κυβερνήσεις, προορισμούς ή αρχές για τη διάθεση τουριστικών δεδομένων προς χρήση. Παραδείγματα στην Ευρώπη είναι το <https://data.europa.eu/en/highlights/open-data-tourism>, <https://info.datatourisme.gouv.fr/> από τη γαλλική κυβέρνηση ή <https://opendata.swiss/de/group/tourism> από την Ελβετία. Ένα άλλο παράδειγμα είναι το Experience Data Hub στην Αυστρία, το οποίο, μαζί με την τουριστική βιομηχανία, προσπαθεί να καθορίσει πρότυπα δεδομένων και να καταστήσει τα σχετικά δεδομένα από τρίτους παρόχους, όπως δεδομένα κινητικότητας, καιρού ή τηλεπικοινωνιών, διαθέσιμα με ομαδοποιημένο και απλό τρόπο. Με αυτόν τον τρόπο, τα δεδομένα μπορούν να ενσωματωθούν στο οικοσύστημα της ίδιας της εταιρείας και να αναπτυχθούν καινοτόμες υπηρεσίες για παράδειγμα μέσω νεοφυών επιχειρήσεων. Ωστόσο, οι μελέτες για τον τουρισμό χρησιμοποιούν επίσης σύνολα δεδομένων που παρέχονται σε πλατφόρμες όπως η Kaggle, δεδομένα που παρέχονται από εταιρείες για περαιτέρω ανάλυση, όπως το ακαδημαϊκό σύνολο δεδομένων Yelp (Guerreiro & Rita, 2020), ή πρωτοβουλίες όπως η Inside Airbnb, ένας ανεξάρτητος και μη εμπορικός ιστότοπος που παρακολουθεί τα δεδομένα της Airbnb και τα

δημοσιοποιεί. Για παράδειγμα, οι Yu κ.ά. (2020) χρησιμοποίησαν δεδομένα από το Inside Airbnb για να συγκρίνουν τις ανέσεις των καταχωρίσεων της Airbnb με τα ξενοδοχεία, και οι Tussyadiah και Park (2018) ανέλυσαν την περιγραφή και την εμπιστοσύνη των οικοδεσποτών στην Airbnb, ενώ οι Güçlü, Roche και Marimon (2020) διερεύνησαν τα χαρακτηριστικά των πόλεων στην Ευρώπη χρησιμοποιώντας αυτά τα δεδομένα. Ταυτόχρονα, ωστόσο, ο Alsudais (2021), διερευνώντας την ακρίβεια των συνόλων δεδομένων Inside Airbnb, προειδοποιεί κατά τη χρήση τέτοιων συνόλων δεδομένων. Επιπλέον, οι Marsden και Pingry (2018) προειδοποιούν για την ποιότητα των σημαντικών προβλημάτων αριθμητικών δεδομένων στη μη αντανakλαστική χρήση των δημοσίως παρεχόμενων δεδομένων.

5.2 API

Ένας άλλος τρόπος ανάκτησης δεδομένων είναι η χρήση διεπαφών προγραμματισμού εφαρμογών (API), τις οποίες παρέχουν πολλές εταιρείες για τη σύνδεση με το σύστημα. Με αυτόν τον τρόπο, τα κοινά δεδομένα μπορούν επίσης να χρησιμοποιηθούν από άλλους προγραμματιστές για να σχεδιάσουν πρόσθετες υπηρεσίες, αυξάνοντας τελικά την ελκυστικότητα του αρχικού συστήματος. Για πολλά API, πρέπει να υποβάλλετε ένα αίτημα, μαζί με μια περιγραφή, για ποια υπηρεσία χρειάζεστε τα δεδομένα, και οι εταιρείες αποφασίζουν στη συνέχεια ξεχωριστά αν θα δώσουν έγκριση ή όχι. Για ερευνητικά έργα, είναι επομένως αρκετά δύσκολο να αποκτήσετε πρόσβαση, ειδικά όταν πρόκειται για κανάλια κοινωνικών μέσων. Παρόλο που προσφέρουν τα δεδομένα τους μέσω APIs, πρόσφατα εφαρμόστηκαν μαζικοί περιορισμοί για την προστασία των δικαιωμάτων των χρηστών τους. Για παράδειγμα, ενώ εξακολουθεί να είναι σχετικά εύκολο να λαμβάνει κανείς tweets από το Twitter ή φωτογραφίες από το Flickr μέσω ενός API (Ainin et al., 2020), το Instagram και το Facebook δεν επιτρέπουν πλέον τη συλλογή δεδομένων χρηστών (Zhang et al., 2021). Εναλλακτικά, στο Github, για παράδειγμα, μπορούν μερικές φορές να προγραμματιστούν και να διατεθούν άτυπα API (Golenvaux et al., 2020).

5.3 Απόξεση Δεδομένων

Τέλος, υπάρχουν επίσης δεδομένα που είναι δημόσια διαθέσιμα στο Διαδίκτυο, αλλά δεν είναι διαμορφωμένα ως λήψη ή δεν μπορούν να ληφθούν μέσω ενός API. Συνήθως, αυτό το περιεχόμενο πρέπει να αντλείται από ιστότοπους ή κανάλια κοινωνικών μέσων και, επειδή είναι διαθέσιμο για εμφάνιση σε προγράμματα περιήγησης, αυτό περιλαμβάνει λογισμικό για την ανάκτηση και την αποθήκευση των ιστοσελίδων. Ως αποτέλεσμα αυτής της ανάκτησης σελίδων, το κείμενο από την πηγή HTML μετατρέπεται σε αρχείο απλού κειμένου. Ωστόσο, ένα πρόβλημα με την ανάκτηση δεδομένων με αυτόν τον τρόπο είναι ότι, εκτός από τα δεδομένα που πραγματικά απαιτούνται, το αρχείο συχνά περιέχει κώδικα που χρησιμοποιείται για την εμφάνιση ή την πλοήγηση στη σελίδα. Συνεπώς, τα δεδομένα πρέπει να εξακολουθούν να είναι

καθαρισμένα και απαλλαγμένα από κώδικα HTML.

Μια προσεκτικότερη ματιά στη βιβλιογραφία για τον τουρισμό αποκαλύπτει ότι τα δεδομένα του TripAdvisor, ειδικότερα, χρησιμοποιούνται αρκετά συχνά για την ανάλυση των τουριστικών προϊόντων ή της τουριστικής συμπεριφοράς. Για παράδειγμα, οι Taecharungroj και Mathayomchan (2019) χρησιμοποίησαν δεδομένα του TripAdvisor για να αναλύσουν τα τουριστικά αξιοθέατα στο Πουκέτ, ενώ οι Khorsand, Rafiee και Kayvanfar (2020) παρείχαν πληροφορίες για ξενοδοχεία στην Τεχεράνη και οι Banerjee και Chua (2016) ανέλυσαν τα πρότυπα αξιολόγησης των ταξιδιωτών σε διαδικτυακές κριτικές ξενοδοχείων. Ως εκ τούτου, χρησιμοποιούνται διαφορετικές προσεγγίσεις απόξεσης για την απόκτηση των απαιτούμενων δεδομένων. Οι Chang, Ku και Chen (2020), για παράδειγμα, προγραμματίσαν ένα C#-based Selenium crawler για τη διερεύνηση της βαθιάς μάθησης και της οπτικής ανάλυσης των κριτικών και των αποκρίσεων ξενοδοχείων. Επιπλέον, οι An, Ma, Du, Xiang και Fan (2020) ανέπτυξαν έναν ανιχνευτή Python για την ανάλυση φωτογραφιών που δημιουργούνται από χρήστες σε διαδικτυακές κριτικές ξενοδοχείων και οι Ganzaroli, Noni και van Baalen (2017) χρησιμοποίησαν το import.io για να διερευνήσουν την επιρροή του TripAdvisor στην ποιότητα των ποιότητα των εστιατορίων στη Βενετία.

Επομένως από τη μία πλευρά, μπορούν να χρησιμοποιηθούν διάφορες γλώσσες προγραμματισμού ή σεναρίων, με την Python και την R να θεωρούνται οι πιο σχετικές, ενώ, ταυτόχρονα, πολυάριθμα εργαλεία λογισμικού και λύσεις παρέχουν ένα γραφικό περιβάλλον εργασίας για την ανάπτυξη ενός ενιαίου προγράμματος ανίχνευσης ιστού. Συνήθως, μπορεί να δημιουργηθεί μια ροή εργασίας στην οποία να καθορίζεται ο τρόπος με τον οποίο ο ανιχνευτής θα πρέπει να περιηγηθεί στη σελίδα και ποια δεδομένα θα πρέπει να αποθηκευτούν, και στη συνέχεια να δημιουργηθεί ένα αρχείο CSV ή Excel με τα δεδομένα που ανιχνεύθηκαν κατεβάστηκαν.

5.4 Νομικές προοπτικές εξόρυξης κειμένου και δεδομένων

Η νομική προοπτική του εντοπισμού δεδομένων στο Διαδίκτυο είναι ένα συχνά συζητούμενο θέμα, καθώς είναι πολύ εύκολο να βρεθεί κανείς σε δύσκολη νομική κατάσταση, ιδίως αν ληφθεί υπόψη ότι οι διεθνείς νομικές συμβάσεις διαφέρουν σημαντικά. η νομική κατάσταση για την Ευρωπαϊκή Ένωση.

Η Ευρωπαϊκή Ένωση (ΕΕ) θέσπισε ένα νομικό πλαίσιο για την εξόρυξη κειμένου και δεδομένων (ΕΞΕΔ) με την οδηγία 2019/790 της ΕΕ, της 17ης Απριλίου 2019, σχετικά με το δικαίωμα του δημιουργού και τα συγγενικά δικαιώματα στην ψηφιακή ενιαία αγορά (οδηγία ΕΞΕΔ), η οποία έπρεπε να μεταφερθεί στο εθνικό δίκαιο από τα κράτη μέλη έως τις 7 Ιουνίου 2021. Επιπλέον, η Ευρωπαϊκή Επιτροπή είχε δηλώσει στο έγγραφο εργασίας των υπηρεσιών της Επιτροπής της 14ης Σεπτεμβρίου 2016 ότι ο κατακερματισμός της ενιαίας αγοράς είναι πιθανό να αυξηθεί ως αποτέλεσμα της υιοθέτησης εξαιρέσεων από την DSM από τα κράτη μέλη σε

εθνικό επίπεδο. Αυτό θα μπορούσε να βασίζεται σε διαφορετικές προϋποθέσεις, κάτι που είναι πιθανό να συμβεί ελλείψει παρέμβασης της ΕΕ.

Ένα παράδειγμα που αναφέρεται συχνά εκτός Ευρώπης είναι αυτό της HiQ έναντι του LinkedIn. Η HiQ αναζήτησε δεδομένα από δημόσια προφίλ χρηστών του LinkedIn χωρίς να συνδεθεί, οδηγώντας τους κατευθείαν στο δικαστήριο (Woollacott, 2019)- ωστόσο, υπάρχουν και αντιπαραδείγματα. Καθώς η διεθνής νομική κατάσταση μπορεί να διαφέρει σημαντικά από χώρα σε χώρα, ο Ng (2019) συνιστά να εξετάζετε προσεκτικά και να σέβεστε τους όρους χρήσης, να επικοινωνείτε με τον ιδιοκτήτη του ιστότοπου και να ελέγχετε το Robots.txt προτού ξεκινήσετε την απόξεση δεδομένων. Προτρέπει επίσης όλους τους ερευνητές να διεξάγουν τη διαδικασία απόξεσης 74 R. Egger κ.ά. συντηρητικά και να μην καταπονούν ή καταρρίπτουν τους διακομιστές, ώστε να διασφαλίζεται ότι τα δεδομένα που έχουν απόξεση διατηρούνται ασφαλή, δεν δίνονται αδιακρίτως και αδιακρίτως σε τρίτους ή/και δεν αντιγράφονται.

5.5 Τυπικές περιπτώσεις χρήσης απόξεσης ιστού στον τουρισμό

Μια πολύ τυπική εφαρμογή του Web scraping στον τομέα του τουρισμού είναι η εξαγωγή δεδομένων για κρατήσεις και αξιολόγηση προσφερόμενων διακοπών που δεν είναι συνήθως προσβάσιμες μέσω API, όπως η προαναφερθείσα πύλη TripAdvisor. Αυτές περιλαμβάνουν συνήθως περιγραφές των προσφορών και πληροφορίες τιμών που παρέχονται από τους φορείς εκμετάλλευσης της πύλης, καθώς και περιεχόμενο που δημιουργείται από χρήστες, όπως κριτικές. Ως εκ τούτου, τα δεδομένα μπορούν να χρησιμοποιηθούν όχι μόνο για την ανάλυση και τη σύγκριση της ποιότητας των προσφερόμενων διακοπών, αλλά φυσικά και για να παρακολουθούν τις τιμές, οι οποίες, με τη σειρά τους, επιτρέπουν την παρατήρηση και τη σύγκριση των γενικών τάσεων στον κλάδο μεταξύ διαφορετικών περιοχών. Μια συγκεκριμένη περίπτωση χρήσης, ειδικότερα, αφορά τη συλλογή δεδομένων για τις τιμές των αεροπορικών εισιτηρίων από ιστότοπους αεροπορικών εταιρειών και διαδικτυακές πύλες κρατήσεων πτήσεων. Εδώ, μια μοναδική πρόκληση έγκειται στη δυναμική τιμολόγηση αυτού του τμήματος, γεγονός που καθιστά αναγκαία την ανίχνευση με ανιχνευτές όχι μόνο σε συχνές χρονικές περιόδους αλλά και με διαφορετικές τελικές συσκευές και από διαφορετικές γεωγραφικές τοποθεσίες. Ωστόσο, εκτός από τις εφαρμογές που βασίζονται σε παγκόσμια δεδομένα, υπάρχουν και άλλες περιπτώσεις χρήσης που λειτουργούν σε πιο τοπικό επίπεδο, όπως η αξιολόγηση του μενού ενός εστιατορίου, οι οποίες είναι εξίσου σημαντικές.

Προτού χρησιμοποιηθούν τα δεδομένα και αναλυθούν με μια κατάλληλη μέθοδο, για παράδειγμα, παλινδρόμηση ή ταξινόμηση (βλ. τα σχετικά κεφάλαια) για την πρόβλεψη, πρέπει συνήθως να υποβληθούν σε ορισμένα στάδια προεπεξεργασίας. Εκτός από τα αριθμητικά δεδομένα (π.χ. τιμές), τα δεδομένα που έχουμε στη διάθεσή μας θα μπορούσαν δυνητικά να έχουν τη μορφή κειμένου, τα οποία όμως συχνά περιέχουν τμήματα που δεν χρειάζονται για την

ανάλυση. Η αναζήτηση και η εξαγωγή των επιθυμητών αποσπασμάτων κειμένου με τη χρήση λειτουργιών ανάλυσης κειμένου, όπως οι κανονικές εκφράσεις, είναι συνήθως μια πολύπλοκη και χρονοβόρα εργασία που είναι επίσης επιρρεπής σε σφάλματα. Ως εκ τούτου, εξειδικευμένα πακέτα όπως το "BeautifulSoup",¹ "Selenium",² ή το "Scrapy"³ παρέχουν μια λύση εδώ

5.6 BeautifulSoup

Το BeautifulSoup είναι μια βιβλιοθήκη Python που διαβάζει δεδομένα από αρχεία HTML και XML. Το προσφέρει αμέτρητες δυνατότητες πλοήγησης, αναζήτησης και αλλαγής δέντρων ανάλυσης, έτσι συνήθως εξοικονομώντας ώρες, ή ακόμα και ημέρες, εργασίας. Με την αντιμετώπιση ενός συγκεκριμένου HTML στοιχείο (χρησιμοποιώντας το όνομα κλάσης CSS), το πακέτο προγραμμάτων προσφέρει λειτουργίες που επιτρέπουν μέρη του κειμένου που πρέπει να βρεθούν και έχει επίσης λειτουργίες για την εξαγωγή του περιεχομένου ή χαρακτηριστικά του στοιχείου HTML. Μπορεί να χρησιμοποιηθεί για να βρει είτε μεμονωμένα στοιχεία είτε επίσης αποσπάσματα που πληρούν ορισμένα κριτήρια. Για να μπορέσετε να μετακινηθείτε γύρω από το έγγραφο, ένα όχι μόνο έχει τη δυνατότητα να το διασχίσει γραμμικά, αλλά μπορεί επίσης να πλοηγηθεί κατά μήκος από τις άκρες των κλαδιών που εκτείνονται από τα στοιχεία της σελίδας HTML. Για παράδειγμα, όλα οι υπερσύνδεσμοι που υπάρχουν σε μια συγκεκριμένη παράγραφο του εγγράφου μπορούν να είναι εξάγεται, πρώτα μεταβαίνοντας στην επιθυμητή παράγραφο και στη συνέχεια αναζητώντας τα χαρακτηριστικά των < a>-tags που εμφανίζονται εκεί.

Το δέντρο σελίδων του εγγράφου HTML ή XML μπορεί, ωστόσο, όχι μόνο να χρησιμοποιηθεί για πλοήγηση, αλλά μπορεί επίσης να αλλάξει η ίδια. Για παράδειγμα, τμήματα του το έγγραφο μπορεί να αφαιρεθεί με δομημένο τρόπο διαγράφοντας ένα δευτερεύον δέντρο ή / και κόμβους μπορεί να εισαχθεί σε ένα πολύ συγκεκριμένο σημείο του εγγράφου.

Αν και υπάρχουν διάφορα άλλα εργαλεία διαθέσιμα (βλ. πρακτική επίδειξη αργότερα, θα χρησιμοποιήσουμε το BeautifulSoup για να εξαγάγουμε το TripAdvisor κριτικές.

5.7 selenium

Το Selenium είναι ένα λογισμικό αυτοματοποίησης προγράμματος περιήγησης ιστού που χρησιμοποιείται συχνά για αυτοματοποιημένες δοκιμές, αλλά μπορεί επίσης να εφαρμοστεί για την απόξεση ιστού. Σε γενικές γραμμές, χρησιμοποιείται πάντα όταν η κλήση και η αξιολόγηση σελίδων μέσω αιτήσεων δεν είναι αρκετή. Εάν απαιτούνται ενέργειες του προγράμματος περιήγησης, όπως κλικ του ποντικιού, για την πρόσβαση σε πληροφορίες, τότε η δοκιμή του προγράμματος περιήγησης με εργαλεία όπως το Selenium είναι χρήσιμη. Επιπλέον, ένα πρόσθετο στον Chrome και τον Firefox μπορεί να καταγράφει και να αναπαράγει τις

αλληλεπιδράσεις του χρήστη με το πρόγραμμα περιήγησης, οι οποίες, με τη σειρά τους, μπορούν να χρησιμοποιηθούν για τη δημιουργία απλών σεναρίων. Εκτός από το δέντρο σελίδας HTML από το Tripadvisor.com 76 R. Egger et al. το πραγματικό εργαλείο απόξεσης, το Selenium προσφέρει επίσης ένα περιβάλλον ανάπτυξης για τη δημιουργία εφαρμογών απόξεσης, καθώς και μια πλατφόρμα που επιτρέπει την εκτέλεση εφαρμογών απόξεσης σε ένα δίκτυο διακομιστών, προκειμένου να ικανοποιούνται ακόμη και απαιτήσεις υψηλής απόδοσης.

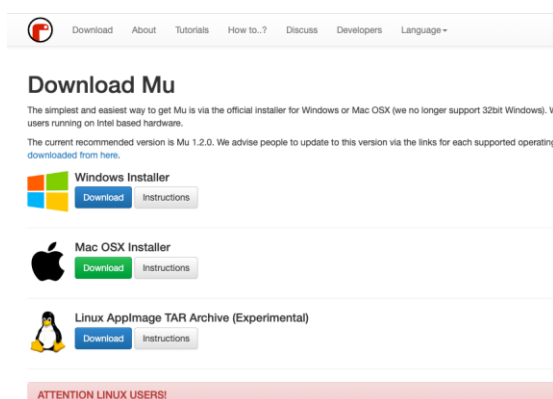
5.8 Scrapy

Το Scrapy είναι ένα πακέτο προγραμμάτων Python που μπορεί εύκολα να ενσωματωθεί σε άλλο λογισμικό που αναπτύσσεται με Python. Εκτός από αυτή τη λήψη, προσφέρεται επίσης εκτενής τεκμηρίωση. Η βιβλιοθήκη Scrapy δεν σας επιτρέπει μόνο να σαρώνετε σελίδες HTML και να εξαγάγετε τμήματα αυτών, αλλά, ως πλαίσιο, προσφέρει επιπλέον τη δυνατότητα να γράφετε πλήρεις εφαρμογές web crawler που διαβάζουν τα περιεχόμενα και ακολουθούν τους συνδέσμους που περιέχονται σε αυτές, προκειμένου να κάνουν διαθέσιμα τα περιεχόμενά τους. Επιπλέον, λαμβάνει επίσης την επεξεργασία και την αποθήκευση των δεδομένων σε βάσεις δεδομένων.

Παράρτημα – Οδηγίες Κώδικα

Οδηγίες Mu Editor – Python

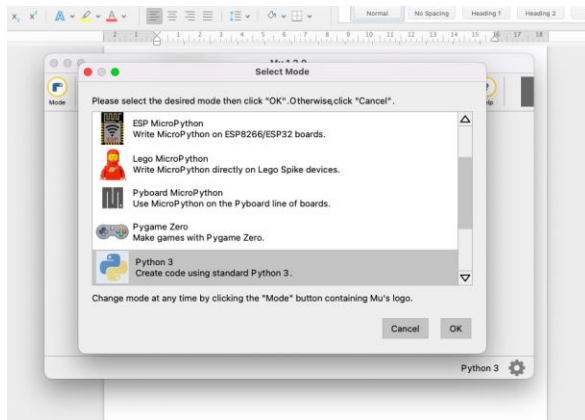
1. Κατεβάστε το από : <https://codewith.mu/en/download>



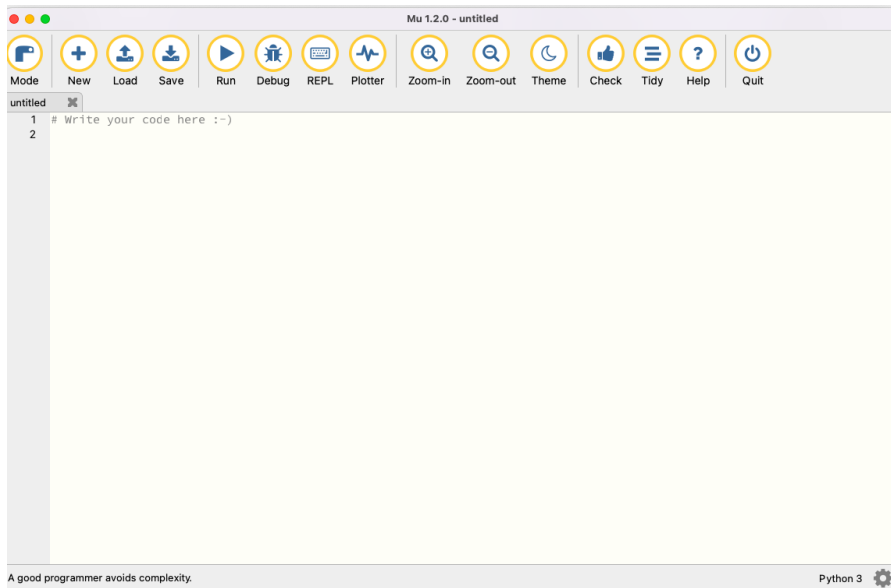
2. Πατήστε για εγκατάσταση και θα δείτε την παρακάτω οθόνη. Όταν όλα τελειώσουν επιτυχώς ανοίχτε την εφαρμογή:



3. Αφού το ξεκινήσετε θα δείτε την οθόνη: Επιλέξτε python 3 και μετά OK



4. Για να ρυθμίσουμε τα απαραίτητα πακέτα κάντε κλικ στο γρανάζι κάτω δεξιά που γράφει Python3

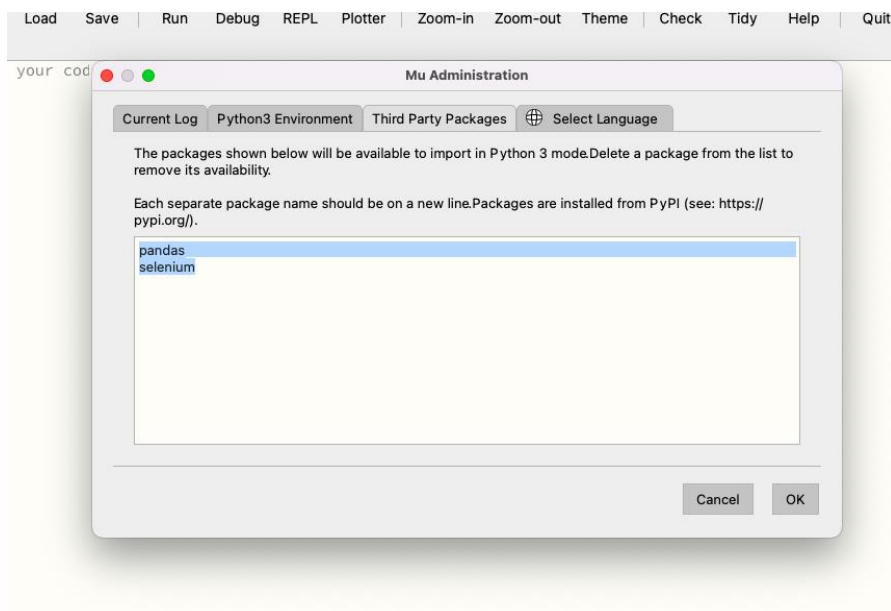


5. Πηγαίνετε στο 3^ο tab (Third Party Packages) και γράψτε δυο γραμμές:

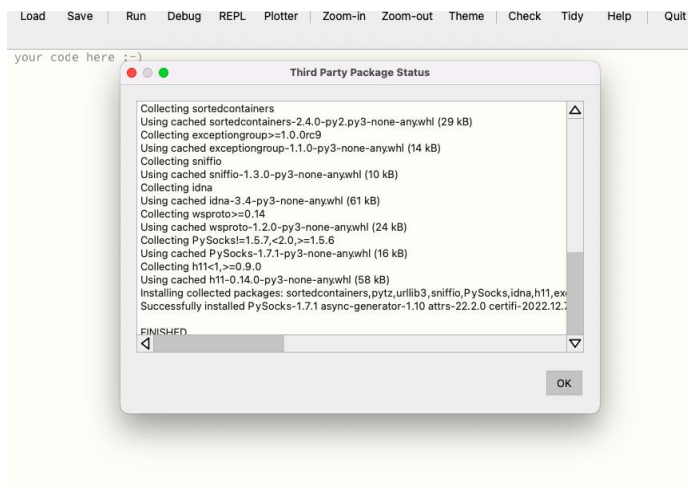
pandas : Για να έχουμε πίνακες

selenium : Για να ανοίξουμε τον browser προγραμματιστικά

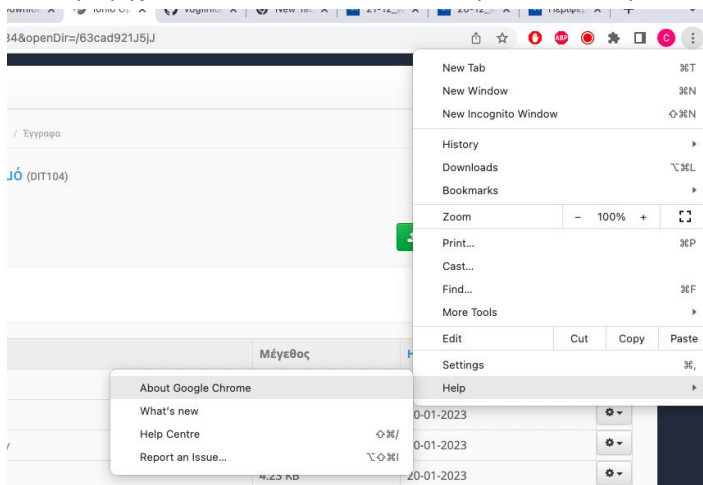
Μετά πατήστε OK

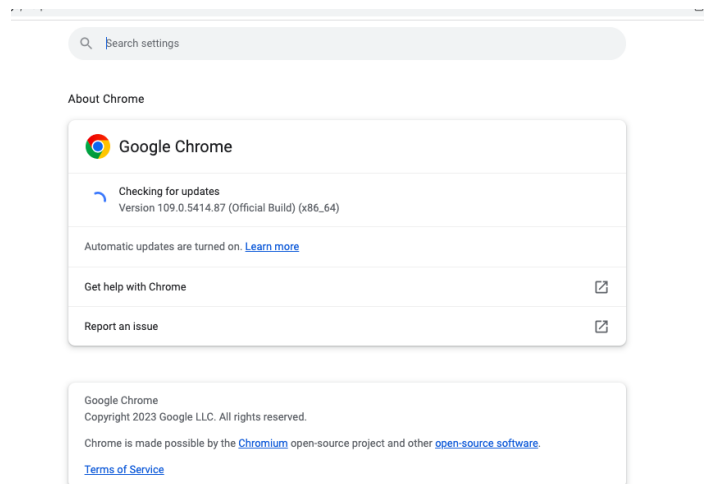


6. Περιμένετε λίγο να τα κατεβάσει και όταν γράψει FINISHED πατήστε OK



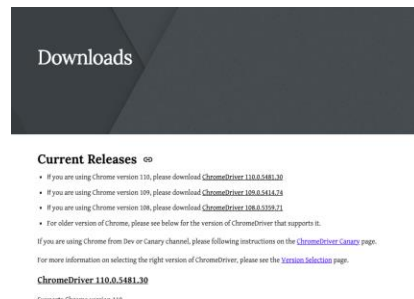
7. Μην ξεχάσετε το chromedriver. Αφού δείτε την έκδοση του chrome





Πηγαίνετε στην <https://chromedriver.chromium.org/downloads>

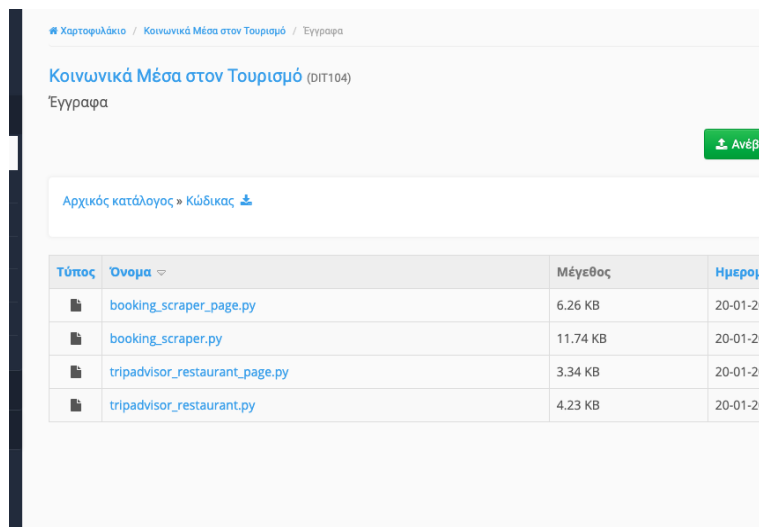
Index of /109.0.5414.74/



	Name	Last modified	Size	ETag
	Parent Directory		-	
	chromedriver_linux64.zip	2023-01-11 05:40:13	6.96MB	245747a6d5f2b7da02c465941f43e0e8
	chromedriver_mac64.zip	2023-01-11 05:40:17	8.72MB	5b5bab24847bb5a6d714985967f2a87a
	chromedriver_mac_arm64.zip	2023-01-11 05:40:20	7.95MB	d5e6a6e6ad4c3647d677b4350c7721
	chromedriver_win32.zip	2023-01-11 05:40:23	6.79MB	b447a4cb23af51da8272f6d867b9b1c7
	notes.txt	2023-01-11 05:40:30	0.00MB	c0c31f9a056a92blad304580f25cbf81

Κατεβάστε το zip chromedriver_win32.zip και αποσυμπιέστε το στα Downloads.

8. Είστε έτοιμοι! Ο κώδικας θα είναι διαθέσιμος στο opencourse. Πχ.



Κατεβάστε για δοκιμή το booking_scraper.py στο δίσκο εκεί που έχετε αποσυμπιέσει το chromedriver_win32.zip.

9. Ανοίξτε το αρχείο booking_scraper.py από το Mu για να δείτε τον κώδικά του

10. Τρέξτε με το Run

Booking

A. Φάση: Το booking κατεβάζει σε δύο φάσεις. Στην 1^η φάση (booking_scraper.py) απλά κατεβάζετε τα καταλύματα ένα ανά γραμμή

1. Σημείο αλλαγής: path_to_file που θα αποθηκεύσετε

```
6 from sklearn.metrics.pairwise_distances import pairwise_distances
7 import pandas as pd
8 from os.path import exists
9
10
11 # default path to file to store data
12 path_to_file = "booking_zakynthos_v3.csv"
13
14 if exists(path_to_file):
15     df = pd.read_csv(path_to_file, header=None)
16     list_ids = df[0].values.tolist()
17     print("Starting with " + len(df) + " properties" + len(list_ids) + " IDs")
```

2. Σημείο αλλαγής:

```
25 # default tripadvisor website of hotel or things to do (attraction/monument)
26
27 urls = [
28 "https://www.booking.com/searchresults.en-gb.html?label=genI73nr-1FCAEoggI46AdIM1gEaFyIAQGYAQm4AQfIAQzYAQHoAQH4AQuIAgGo
29 "https://www.booking.com/searchresults.en-gb.html?ss=Zakynthos%2C+Ionian+Islands%2C+Greece&ssne=Zakynthos&ssne_un touche
30 ]
31
32 pages = [
33 40,
34 10
35 ]
```

B. Φάση: Έχοντας κατεβάσει τα καταλύματα που σας ενδιαφέρουν από τη πρώτη φάση....

Ανοίχτε το αρχείο: booking_scraper_page.py

```

6 from selenium.webdriver.common.keys import Keys
7 from selenium.webdriver.common.by import By
8 import pandas as pd
9
10 from selenium.common.exceptions import WebDriverException
11
12 #
13 # default path to file to store data
14 listings_file = "booking_zakynthos_v3.csv"
15 df = pd.read_csv(listings_file, header=None)
16 df.columns = ['id', 'name', 'txt', 'rate', 'revs', 'nights', 'adults', 'money', 'addr', 'link']
17 urls = df.link.tolist()

```

ΠΗΓΕΣ

Alpaydin, E. (2021). Machine learning (Revised and Updated Edition). MIT Press Essential Knowledge Series.

Baggio, J. A., & Baggio, R. (2020). *Modelling and simulations for tourism and hospitality: An introduction*. Channel View Publications.

Balakrishnan, T., Chui, M., Hall, B., & Henke, N. (2020). The state of AI in 2020. McKinsey. <https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/global-survey-the-state-of-ai-in-2020>

Iafrate, F., & ISTE (Londyn), & John Wiley & Sons. (2018). Artificial intelligence and big data: The birth of a new intelligence. ISTE.

Ilyas, I. F., & Chu, X. (2019). Data cleaning. ACM Books <https://doi.org/10.1145/3310205>
<https://www.gartner.com/smarterwithgartner/7-digital-disruptions-you-might-not-see-coming-in-the-next-5-years> 12

L. Mich Laudon, K. C., & Laudon, J. P. (2020). Managing information systems: Managing the digital firm (16th ed.)

Pearson. O'Neil, C., & Schutt, R. (2013). Doing data science: Straight talk from the frontline. O'Reilly Media.

Russell, S. J., & Norvig, P. (2020). Artificial intelligence: A modern approach (4th ed.). Pearson.

Sigala, M., Rahimi, R., & Thelwall, M. (Eds.). (2019). Big data and innovation in tourism, travel, and hospitality: Managerial approaches, techniques, and applications (1st ed.). Springer Publishing.

Voulgaris, Z., & Bulut, Y. E. (2018). AI for data science: Artificial intelligence frameworks and functionality for deep learning, optimization, and beyond. Basking Ridge. Zhang, D., Mishra, S.,

Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B., Lyons, T., Manyika, J., Niebles, J. C., Sellitto, M., Shoham, Y., Clark, J., & Perrault, R. (2021). The AI index 2021 annual report. AI Index Steering Committee Human-Centered AI Institute, Stanford University
<https://hai.stanford.edu/research/ai-index-2021>

Baggio, R., Scott, N., & Cooper, C. (2010). Improving tourism destination governance: A complex science approach. *Tourism Review*, 65(4), 51–60.

Bartneck, C., Lütge, C., Wagner, A., & Welsh, S. (2021). Privacy issues of AI. In *An introduction to ethics in robotics and AI* (SpringerBriefs in ethics). Springer. https://doi.org/10.1007/978-3-030-51110-4_8

Boden, M. A. (1998). Creativity and artificial intelligence. *Artificial Intelligence*, 103(1–2), 347–356. [https://doi.org/10.1016/S0004-3702\(98\)00055-1](https://doi.org/10.1016/S0004-3702(98)00055-1)

Bollinger, T. (2021). Why AlphaFold is not like AlphaGo. *Academia Letters*, Article 728. <https://doi.org/10.20935/AL728> AI and Big Data in Tourism

Brownlee, J. (2019). Deep learning for computer vision. Image classification, object detection, and face recognition in python. <https://machinelearningmastery.com/deep-learning-for-computer-vision>

Chauhan, P., & Sood, M. (2021). Big data: Present and future. *Computer*, 54, 59–65. <https://doi.org/10.1109/MC.2021.3057442> European Economic and Social Committee. (2017). The ethics of Big Data: Balancing economic benefits and ethical questions of Big Data in the EU policy context

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., Lally, A., Murdock, J. W., Nyberg, E., Prager, J., Schlaefer, N., & Welty, C. (2010). The AI behind Watson—The technical

article. AI Magazine. <http://www.aaai.org/Magazine/Watson/watson.php>

Garigliano, R., & Mich, L. (2019). Looking inside the black box: Core semantics towards accountability of artificial intelligence. In M. ter Beek, A. Fantechi, & L. Semini (Eds.), *From software engineering to formal methods and tools, and back (LNCS)* (Vol. 11865). Springer. https://doi.org/10.1007/978-3-030-30985-5_16

Hsu, J. (2017, January). Meet the new AI challenging human poker pros. *IEEE Spectrum*. <https://spectrum.ieee.org/automaton/artificial-intelligence/machine-learning/meet-the-new-ai-challenging-human-poker-pros>

Ilyas, I. F. (2020, February). AI should not leave structured data behind! How AI can solve the notorious data cleaning and prep problems. *Towards Data Science*. <https://towardsdatascience.com/ai-should-not-leave-structured-data-behind-33474f9cd07a>

Jordan, M. I. (2019). Artificial intelligence—The revolution hasn't happened yet *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.f06c6e61>

Krensky, P., Idoine, C., Brethenoux, E., den Hamer, P., Choudhary, F., Jaffri, A., & Vashisth, S. (2021, March). Gartner magic quadrant for data science and machine learning platforms. Gartner. <https://www.gartner.com/en/documents/3998753>

Lakshmanan, V., Robinson, S., & Munn, M. (2020). *Machine learning design patterns: Solutions to common challenges in data preparation, model building, and MLOps*. O'Reilly.

Lane, H., Howard, C., & Hapke, H. M. (2019). *Natural language processing in action*. Manning.

Li, J., Xu, L., Tang, L., Wang, S., & Li, L. (2018). Big data in tourism research: A literature review. *Tourism Management*, 68, 301–323. <https://doi.org/10.1016/j.tourman.2018.03.009>

Lim, M. (2018). History of AI winters. *Actuaries Digital*. <https://www.actuaries.digital/2018/09/05/history-of-ai-winters> Marr, B. (2015). *Big data: Using smart big data, analytics and metrics to make better decisions and improve performance*. Wiley.

McCarthy, J. (2007, November). What is artificial intelligence? Stanford. <http://jmc.stanford.edu/articles/whatisai/whatisai.pdf>

Mich, L. (2020a). Artificial intelligence and machine learning. In Z. Xiang, M. Fuchs, U. Gretzel, & W. Höpken (Eds.), *Handbook of e-Tourism* (pp. 1–21). Springer. https://doi.org/10.1007/978-3-030-05324-6_25-1

Mich, L. (2020b). Systems and technologies for fluxes management. In M. Franch & R. Peretta (Eds.), *Tourism, fragilities and emergencies* (pp. 107–134). McGraw Hill.

Murtagh, F., & Devlin, K. (2018). The development of data science: Implications for education, employment, research, and the data revolution for sustainable development. *Big Data and Cognitive Computing*, 2(2), 14. <https://doi.org/10.3390/bdcc2020014>

O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. *New York Time*. Petrov, C. (2021, July). 25+ impressive big data statistics for 2021. Techjury. <https://techjury.net/blog/big-data-statistics/#gref>

Pierce, C. S. (1931–1958). In: C. Hartshorne & P. Weiss (Eds.), *Collected papers of Charles Sanders Peirce* (Vols. 1–8). Harvard University Press.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *Sn Computer Science*, 2, 160. <https://doi.org/10.1007/s42979-021-00592-x> 14 L.

Mich Sharda, R., Delen, D., & Turban, E. (2021). *Analytics, data science, & artificial intelligence: Systems for decision support*. Pearson. Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of go without human knowledge. *Nature*, 550, 354–359.

Sommerville, I. (2018). *Software engineering* (10th Global Ed.).

Pearson. Southerton, C. (2020). Datafication. In L. Schintler & C. McNeely (Eds.), *Encyclopedia of big data*. Springer. https://doi.org/10.1007/978-3-319-32001-4_332-1

Subirana, B. (2020). Call for a wake standard for artificial intelligence. *Communication of ACM*, 63(7), 32–35. <https://doi.org/10.1145/3402193>

Van Harmelen, F., Lifschitz, V., & Porter, B. (2010). *Handbook of knowledge representation*. Amsterdam Elsevier. Zeni,

N., Kiyavitskaya, N., Mich, L., Cordy, J. R., & Mylopoulos, J. (2015). GaiusT: Supporting the extraction of rights and obligations for regulatory compliance. *Requirements Engineering*, 20(1), 1–22.

Zhou, T., Shen, J., He, D., Vijayakumar, D., & Kumar, N. (2020). Human-in-the-loop-aided privacy-preserving scheme for smart healthcare. *IEEE Transactions on Emerging Topics in Computational Intelligence*. <https://doi.org/10.1109/TETCI.2020.2993841>

References WEB SCRAPPING

Ainin, S., Feizollah, A., Anuar, N. B., & Abdullah, N. A. (2020). Sentiment analyses of multilingual tweets on halal tourism. *Tourism Management Perspectives*, 34, 100658. <https://doi.org/10.1016/j.tourman.2020.100658>

1016/j.tmp.2020.100658

Alsudais, A. (2021). Incorrect data in the widely used Inside Airbnb dataset. *Decision Support Systems*, 141, 113453. <https://doi.org/10.1016/j.dss.2020.113453> An,

Q., Ma, Y., Du, Q., Xiang, Z., & Fan, W. (2020). Role of user-generated photos in online hotel reviews: An analytical approach. *Journal of Hospitality and Tourism Management*, 45, 633–640. <https://doi.org/10.1016/j.jhtm.2020.11.002>

Banerjee, S., & Chua, A. Y. (2016). In search of patterns among travellers' hotel ratings in TripAdvisor. *Tourism Management*, 53, 125–131. <https://doi.org/10.1016/j.tourman.2015.09.020>

Chang, Y.-C., Ku, C.-H., & Chen, C.-H. (2020). Using deep learning and visual analytics to explore hotel reviews and responses. *Tourism Management*, 80, 104129. <https://doi.org/10.1016/j.tourman.2020.104129>

Corbett, C. J. (2018). How sustainable is big data? *Production and Operations Management*, 27(9), 1685–1695. <https://doi.org/10.1111/poms.12837>

Egger, R. (2015). Die Welt wird phygital: Metamorphosen touristischer Räume. In R. Egger & K. Luger (Eds.), *Tourismus und mobile Freizeit: Lebensformen, Trends, Herausforderungen* (pp. 27–46). Books on Demand.

Egger, R., & Yu, C.-E. (2022). Epistemological challenges. In R. Egger (Ed.), *Tourism on the verge. Applied data science in Tourism: Interdisciplinary approaches, methodologies and applications* (pp. 17–34). Springer.

Fermoso, A. M., Mateos, M., Beato, M. E., & Berjón, R. (2015). Open linked data and mobile devices as e-tourism tools. A practical approach to collaborative e-learning. *Computers in Human Behavior*, 51, 618–626. <https://doi.org/10.1016/j.chb.2015.02.032> Fuchs, C. (2021). *Social media: A critical introduction* (3rd ed.).

Sage. Ganzaroli, A., de Noni, I., & van Baalen, P. (2017). Vicious advice: Analyzing the impact of TripAdvisor on the quality of restaurants as part of the cultural heritage of Venice. *Tourism Management*, 61, 501–510. <https://doi.org/10.1016/j.tourman.2017.03.019>

Golenvaux, N., Alvarez, P. G., Kiossou, H. S., & Schaus, P. (2020). An LSTM approach to Forecast Migration using Google Trends. Gretzel, U., Reino, S., Kopera, S., & Koo, C. (2015a). Smart tourism challenges. *Journal of Tourism*, 41–47.

Gretzel, U., Sigala, M., Xiang, Z., & Koo, C. (2015b). Smart tourism: Foundations and developments. *Electronic Markets*, 25(3), 179–188. <https://doi.org/10.1007/s12525-015-0196-8>

Güçlü, B., Roche, D., & Marimon, F. (2020). City characteristics that attract Airbnb Travellers: Evidence from Europe. 1800–6450. *International Journal for Quality Research*, 14(1), 271–290. <https://doi.org/10.24874/IJQR14.01-17>

Guerreiro, J., & Rita, P. (2020). How to predict explicit recommendations in online reviews using text mining and sentiment analysis. *Journal of Hospitality and Tourism Management*, 43, 269–272. <https://doi.org/10.1016/j.jhtm.2019.07.001> Jaspers, K. (1955). Vom Ursprung und Ziel der Geschichte. Kösel.

Johnson, J. A. (2014). From open data to information justice. *Ethics and Information Technology*, 16(4), 263–274.

Johnson, P. A., Sieber, R. E., Magnien, N., & Ariwi, J. (2012). Automated web harvesting to collect and analyse user-generated content for tourism. *Current Issues in Tourism*, 15(3), 293–299. <https://doi.org/10.1080/13683500.2011.555528>

Khorsand, R., Rafiee, M., & Kayvanfar, V. (2020). Insights into TripAdvisor’s online reviews: The case of Tehran’s hotels. *Tourism Management Perspectives*, 34, 100673. <https://doi.org/10.1016/j.tmp.2020.100673> [Web Scraping 81](#)

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data and Society*, 1(1), 205395171452848. <https://doi.org/10.1177/2053951714528481>

Li, J., Xu, L., Tang, L., Wang, S., & Li, L. (2018). Big data in tourism research: A literature review. *Tourism Management*, 68, 301–323. <https://doi.org/10.1016/j.tourman.2018.03.009>

Longhi, C., Titz, J. B., & Viallis, L. (2014). Open data: Challenges and opportunities for the tourism industry. In E. Kitanovska-Ristoska (Ed.), *Tourism management, marketing, and development* (pp. 15–40).

Marsden, J. R., & Pingry, D. E. (2018). Numerical data quality in IS research and the implications for replication. *Decision Support Systems*, 115, A1–A7. <https://doi.org/10.1016/j.dss.2018.10.007>

Marshall, M. (1962). *The Gutenberg Galaxy: The making of typographic man*. University of Toronto Press. <https://doi.org/10.4324/9780203992968-14>

Ng, A. (2019, July 18). Is web crawling legal? Towards Data Science. Towards Data Science. Retrieved from <https://towardsdatascience.com/is-web-crawling-legal-a758c8fcacde>

Nora, S., & Minc, A. (1980). Computerizing society. *Society*, 17(2), 25–30. <https://doi.org/10.1007/bf02700056>

- Pantano, E., Priporas, C.-V., & Stylos, N. (2017). 'You will like it!' using open data to predict tourists' response to a tourist attraction. *Tourism Management*, 60, 430–438. <https://doi.org/10.1016/j.tourman.2016.12.020>
- Pantano, E., Priporas, C.-V., Stylos, N., & Dennis, C. (2019). Facilitating tourists' decision making through open data analyses: A novel recommender system. *Tourism Management Perspectives*, 31, 323–331. <https://doi.org/10.1016/j.tmp.2019.06.003>
- Pesonen, J., & Lampi, M. (2016). Utilizing open data in tourism. In A. Inversini & R. Schegg (Chairs), ENTER 2016 conference on information and communication technologies in tourism.
- Taecharungroj, V., & Mathayomchan, B. (2019). Analysing TripAdvisor reviews of tourist attractions in Phuket, Thailand. *Tourism Management*, 75, 550–568. <https://doi.org/10.1016/j.tourman.2019.06.020>
- Thorp, J. (2012). Big data is not the new oil. Retrieved from <https://hbr.org/2012/11/data-humansand-the-new-oil>
- Tussyadiah, I. P., & Park, S. (2018). When guests trust hosts for their words: Host description and trust in sharing economy. *Tourism Management*, 67, 261–272. <https://doi.org/10.1016/j.tourman.2018.02.002>
- Woollacott, E. (2019, October 9). LinkedIn data scraping ruled legal. *Forbes*. Retrieved from <https://www.forbes.com/sites/emmawoollacott/2019/09/10/linkedin-data-scraping-ruled-legal/?sh%3e9917a31b54>
- Xiang, Z., & Fesenmaier, D. R. (2017a). Analytics in tourism design. In Z. Xiang & D. R. Fesenmaier (Eds.), *Analytics in smart tourism design* (pp. 1–12). Springer International.
- Xiang, Z., & Fesenmaier, D. R. (2017b). Big data analytics, tourism design and smart tourism. In *Analytics in smart tourism design* (pp. 299–307). Springer. https://doi.org/10.1007/978-3-319-44263-1_17
- Yu, J., & Egger, R. (2021). Color and engagement in touristic Instagram pictures: A machine learning approach. *Annals of Tourism Research*, 103, 204. <https://doi.org/10.1016/j.annals.2021.103204>
- Yu, M., Cheng, M., Yu, Z., Tan, J., & Li, Z. (2020). Investigating Airbnb listings' amenities relative to hotels. *Current Issues in Tourism*, 1–18. <https://doi.org/10.1080/13683500.2020.1733497>
- Zhang, H., van Berkel, D., Howe, P. D., Miller, Z. D., & Smith, J. W. (2021). Using social media to measure and map visitation to public lands in Utah. *Applied Geography*, 128, 102389. <https://doi.org/10.1016/j.apgeog.2021.102389>