Αναφορά αξιολόγησης για συστήματα επεξεργασίας και εξόρυξης κειμένου χρησιμοποιώντας την Ανάλυση SWOT

Δημήτρης Ρουσίδης,[1,2] Πάνος Μπαλατσούκας,[3] Εμμανουήλ Γαρουφάλλου,[2]

[1] University of Alcala, Madrid, Spain, drousid@gmail.com
[2] Αλεξάνδρειο Τεχνολογικό Εκπαιδευτικό Ίδρυμα Θεσσαλονίκης, mgarou@libd.teithe.gr
[3] University of Manchester, United Kingdom, panagiotis.balatsoukas@manchester.ac.uk

Εισαγωγή

Τεράστιες ποσότητες νέων πληροφοριών και δεδομένων παράγονται καθημερινά, μέσω οικονομικών, ακαδημαϊκών και κοινωνικών δραστηριοτήτων [1]. Ο κατάλογος της βιβλιοθήκης για το Max Planck Society (MPG), για παράδειγμα, περιείχε 41.000 e-books σχεδόν στα τέλη του 2008. Ο Springer ξεκίνησε e-book πρόγραμμα το 2006 με 10.000 τίτλους, το 2009 ήταν περίπου 30.000 και 5.000 προστίθενται κάθε χρόνο [2]. Ο μέσος χρόνος που δαπανάται για να διαβαστεί ένα άρθρο περιοδικού ήταν περίπου 45-50 λεπτά μεταξύ του 1977 και τα μέσα της δεκαετίας του 1990, αλλά έκτοτε μειώθηκε σε μόλις πάνω από 30 λεπτά. Αυτό συνέβη παρά το γεγονός ότι το μέσο μέγεθος των άρθρων περιοδικών αυξήθηκε σημαντικά από τις 7,4 στις 12,4 σελίδες μεταξύ του 1975 και του 2001 [2].

Σύμφωνα με το MGI (McKinsey Global Institute) εκτιμάται ότι οι επιχειρήσεις αποθήκευσαν σε παγκόσμιο επίπεδο, περισσότερα από 7 exabytes νέων δεδομένων στις μονάδες δίσκων το 2010, ενώ οι καταναλωτές αποθήκευσαν πάνω από 6 exabytes νέων δεδομένων σε συσκευές όπως προσωπικούς και φορητούς υπολογιστές [3]. Για να αντιληφθούμε το απίστευτο αυτό μέγεθος θα πρέπει να γνωρίζουμε ότι ένα exabyte δεδομένων αντιστοιχεί σε 50.000 χρόνια συνεχόμενων κινηματογραφικών ψηφιακών ταινιών [4] και είναι επίσης ισοδύναμο με πάνω από 4.000 φορές τα στοιχεία που φυλάσσονται στη Βιβλιοθήκη του Κογκρέσου των ΗΠΑ [3]. Όπως είναι αναμενόμενο όσο αυξάνεται ο όγκος των δεδομένων τόσο αυξάνεται και η ανάγκη πόρων, κυρίως των ανθρώπινων αλλά και εφαρμογών που θα συνεισφέρουν στη διαχείριση. Για παράδειγμα, μόνο οι Ηνωμένες Πολιτείες αντιμετωπίζουν έλλειψη 140.000 με 190.000 ειδικού προσωπικού με βαθιά αναλυτικές ικανότητες, καθώς και 1,5 εκατ. διαχειριστών και αναλυτών για την ανάλυση μεγάλου όγκου δεδομένων (γνωστό ως Big Data) και τη λήψη αποφάσεων με βάση τα ευρήματά τους [3]. Υπάρχει επίσης η ανάγκη για συνεχή καινοτομία στις τεχνολογίες και τεχνικές που θα βοηθήσει τα άτομα και τις οργανώσεις να ενσωματώσουν, να αναλύσουν, να απεικονίσουν και να καταναλώσουν τον

αυξανόμενο «χείμαρρο» των μεγάλων δεδομένων [3]. Στο παραπάνω συμπέρασμα κατέληξαν και οι συντάκτες της έκθεσης JISC συμπεραίνοντας αυτή η θάλασσα δεδομένων, η οποία προβλέπεται να αυξάνεται με ρυθμό 40% ετησίως, έχει σημαντικές δυνατότητες οικονομικής και κοινωνικής αξίας. Τεχνικές, όπως η εξόρυξη κειμένου και δεδομένων και αναλύσεων απαιτούνται για την αξιοποίηση αυτού του δυναμικού [1].

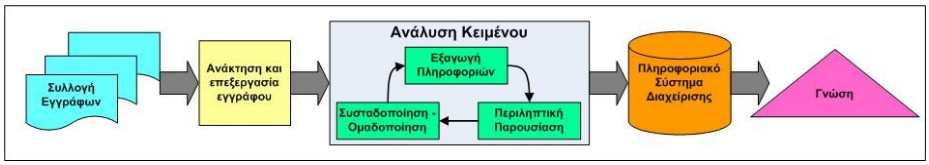
Η δομή του υπόλοιπου άρθρου έχει ως εξής: Αρχικά επεξηγούνται βασικοί όροι όπως η εξόρυξη κειμένου, η Ανάλυση SWOT και τα εργαλεία εξόρυξης κειμένου που αναλύθηκαν. Κατόπιν αναφέρεται η μεθοδολογία που ακολουθήθηκε και παρατίθενται τα εργαλεία με τα σημαντικότερα χαρακτηριστικά τους καθώς και τα αποτελέσματα της Ανάλυσης SWOT. Τέλος συζητούνται τα συμπεράσματα.

Βασικοί Όροι

Εξόρυξη Κειμένου

Εξόρυξη κειμένου είναι η ανακάλυψη προηγουμένως αγνώστων πληροφοριών ή ιδεών από αρχεία κειμένου με αυτόματη εξαγωγή πληροφοριών από διάφορες γραπτές πηγές χρησιμοποιώντας λογισμικό ηλεκτρονικών υπολογιστών [5].

Στην Εικόνα 1 φαίνεται ένα γενικό μοντέλο διαδικασίας για μια εφαρμογή εξόρυξης κειμένου. Ξεκινώντας με μια συλλογή εγγράφων το εργαλείο εξόρυξης κειμένου θα ανακτήσει κείμενο και θα το προ-επεξεργαστεί ελέγχοντας τη μορφή και το σύνολο των χαρακτήρων του. Στη συνέχεια θα περάσει στη φάση ανάλυσης του και χρησιμοποιώντας κάποιες από μια πλειάδα διαφόρων τεχνικών εξόρυξης δεδομένων θα εξάγει πληροφορίες. Τοποθετώντας τα αποτελέσματα της παραπάνω διαδικασίας σε ένα πληροφοριακό σύστημα διαχείρισης θα προκύψει μια αφθονία γνώσης για το χρήστη.



Εικόνα 1. Παράδειγμα εξόρυξης κειμένου (προσαρμογή από [6])

Στον Πίνακα 1 φαίνονται κάποιες από τις εφαρμογές της εξόρυξης κειμένου σε σημαντικούς τομείς όπως η Ιατρική, οι Επιχειρήσεις, η Κυβέρνηση και η Εκπαίδευση.

	εξαγωγή πληροφοριών	εντοπισμός θέματος	περίληψη	κατηγοριοποίηση	συσταδοποίηση-ομαδοποίηση	σύνδεση εννοιών	οπτικοποίηση πληροφοριών	απαντήσεις ερωτήσεων
Ιατρικά:								
Συχνές ερωτήσεις (FAQ's)	x			x		x		x
Σχεδιασμός φαρμάκων	x				x	x		
Νέες θεραπείες		x				x		
Επιχειρήσεις:								
Ανάλυση ανταγωνισμού		x	x					
Επιπτώσεις/ανάλυση των media		x						
Τρέχουσα αντίληψη		x						
Παραβίαση της πνευματικής ιδιοκτησίας	x	x			x			
Υποστήριξη πελατών για Συχνές Ερωτήσεις	x			x	x			x
Εντοπισμός κοινωνικών δικτύων							x	
Εξατομίκευση περιεχομένου		x			x			
Κυβέρνηση:								
Εσωτερική ασφάλεια: εντοπισμός τρομοκρατικών δικτύων	x	x			x	x	x	
Επιβολή Νόμου: εντοπισμός/πρόληψη εγκλήματος	x	x			x	x	x	
Εκπαίδευση:								
Έρευνα ενός θέματος		x	x	x				
Ανάλυση παραπομπών	x				x		x	
Συχνές ερωτήσεις (FAQ's)	x			x	x			x

Πίνακας 1. Παραδείγματα εφαρμογής εργαλείων εξόρυξης κειμένου σε διάφορα πεδία (προσαρμογή από [6])

SWOT Analysis

Η ανάλυση SWOT είναι ένα κλασικό στρατηγικό εργαλείο σχεδιασμού. Χρησιμοποιώντας ένα πλαίσιο εσωτερικών δυνάμεων (Strengths - S) και αδυναμιών (Weaknesses - W) αλλά και εξωτερικών ευκαιριών (Opportunities - O) και απειλών (Threats - T), παρέχει έναν απλό τρόπο αξιολόγησης για το πώς μια στρατηγική μπορεί να υλοποιηθεί καλύτερα. Το εργαλείο βοηθά τους σχεδιαστές να είναι ρεαλιστές σχετικά με το τι μπορούν να επιτύχουν, και που θα πρέπει να επικεντρωθούν [7]. Σε ένα μοντέλο SWOT Ανάλυσης φαίνεται η εσωτερική καταγωγή των ιδιοτήτων του οργανισμού S και W (Internal Origin – Attributes of the organization), η εξωτερική καταγωγή των ιδιοτήτων του περιβάλλοντος O και T (External Origin – Attributes of the environment). Οι ιδιότητες S και O είναι Χρήσιμες για την επίτευξη του σκοπού (Helpful to achieving the objective) ενώ οι ιδιότητες W και T είναι το αντίθετο, δηλαδή Επιζήμιες (Harmful to achieving the objective) [9]. Αν και ο κύριος σκοπός της Ανάλυσης SWOT είναι η αξιολόγηση project ή επιχειρηματικών εγχειρημάτων [8], [9] βρίσκει εφαρμογές και σε αξιολογήσεις λογισμικού [10], [11].

Παρόμοιες εργασίες - Μεθοδολογία

Έχουν γίνει διάφορες εργασίες σχετικά με τη σύγκριση συστημάτων εξόρυξης κειμένου όπως των Fan et al. [6] όπου συνέκριναν συστήματα μεγάλων εταιριών όπως η IBM, SPSS, SAP, SAS, κ.λπ., των Crowsey et al. οι οποίοι αξιολόγησαν λογισμικό εξόρυξης κειμένου αρκετών χιλιάδων δολαρίων [12], των Κέντρων Ελέγχου και Πρόληψης Νοσημάτων (www.cdc.gov) [13] και των Krallinger et al. όπου αξιολογήθηκαν συστήματα εξόρυξης κειμένου για υποβοήθηση στο επιστημονικό πεδίο της Βιολογίας [14]. Κατόπιν έρευνας δεν υπέπεσε στην αντίληψή μας αξιολόγηση συστημάτων που να παρέχονται δωρεάν (free download) και να είναι ελεύθερου πηγαίου κώδικα (open source).

Υπάρχει μια πληθώρα εργαλείων εξόρυξης κειμένου διαδικτυακά. Ο Jan van Gemert [15] το 2000 είχε κάνει μια επισκόπηση 71 εργαλείων εξόρυξης κειμένου. Από τότε πολλά από αυτά τα εργαλεία έχουν βελτιωθεί, άλλα είναι παρωχημένα και καινούρια έχουν δημιουργηθεί.

Στο ερώτημα ποια εργαλεία θα έπρεπε να αξιολογηθούν, η απάντηση θεωρήσαμε ότι θα έπρεπε να είναι η λιγότερο χρονοβόρα. Σύμφωνα με βραβευμένη έρευνα της εταιρίας Google το 2010 [16] μετά την πληκτρολόγηση ερωτήματος σε μηχανή αναζήτησης οι χρήστες συνήθως αξιολογούν τα αποτελέσματα γρήγορα πριν κάνουν κάποιο κλικ ή βελτιώσουν το ερώτημά τους (κατά μέσο όρο μέσα σε 7,78 δευτερόλεπτα). Επίσης οι συγγραφείς ανακάλυψαν ότι οι χρήστες περνούν περισσότερο χρόνο στις σελίδες αναζήτησης όταν αποτυγχάνουν να βρουν αυτό που επιδιώκουν, σε αντίθεση με όταν βρίσκουν αυτό που αναζητούν.

Η σχετικότητα μιας πηγής είναι μία από τις πιο θεμελιώδεις, αν όχι η πιο θεμελιώδης έννοια της θεωρίας της Ανάκτησης Πληροφοριών (IR) [17]. Επί του παρόντος, υπάρχουν δύο κύριες απόψεις σχετικά με την Ανάκτηση Πληροφοριών. Η πρώτη είναι το θέμα-καταλληλότητα, ή τοπικότητα, η οποία ασχολείται με το αν ένα κομμάτι πληροφορίας είναι σχετικό με ένα θέμα που έχει κάποια τοπική σχέση με την ανάγκη πληροφοριών που έχει εκφραστεί από τον χρήστη στο ερώτημα. Η δεύτερη άποψη είναι το χρήστης-χρησιμότητα, η οποία πραγματεύεται με την απόλυτη χρησιμότητα του κομματιού της πληροφορίες για το χρήστη που υπέβαλε το ερώτημα [17]. Επίσης, στην ίδια έρευνα ο Tombros και οι συνεργάτες του προσδιόρισαν πέντε κατηγορίες σχετικά με τα έγγραφα που αναζητεί κάποιος στο Διαδίκτυο. Αυτές είναι: i) το κείμενο, ii) η δομή, iii) η ποιότητα, iv) τα αντικείμενα που δεν είναι κείμενο και v) οι φυσικές ιδιότητες. Σύμφωνα με τα αποτελέσματα της έρευνας από τα 24 συνολικά χαρακτηριστικά αυτών των 5 κατηγοριών οι χρήστες χρειαζόταν κατά μέσο όρο 1,9 και σε αρκετές περιπτώσεις μόνο ένα χαρακτηριστικό για να αξιολογήσουν μια ιστοσελίδα [17]. Τέλος η επιρροή της εμπειρίας στο Διαδίκτυο (αν κάποιος χρήστης είναι ειδικός ή όχι) σχετικά με το μέτρο του χρόνου αναζήτησης είναι μάλλον αδύναμη, αλλά λιγότερο εξαρτημένη από το αν μια σελίδα που σχετίζεται με το ερώτημα σε μηχανή αναζήτησης προσπελάζεται ή όχι (έρευνα των Hoelscher & Strube [18]).

Η επιλογή των πέντε εργαλείων για αξιολόγηση έγινε βάσει μιας απλής αναζήτησης για δωρεάν και ανοιχτού κώδικα εργαλεία στο Google, όπου εισήχθη η φράση κλειδί “free text mining tools”. Στην πρώτη σελίδα των αποτελεσμάτων αναζήτησης το πρώτο στη λίστα αποτέλεσμα ήταν το λήμμα της Wikipedia71 σχετικά με το text mining, το δεύτερο μια λίστα της ιστοσελίδας KDnuggets72 όπου παρέθετε 66 εμπορικά (commercial) και 12 δωρεάν (free) εργαλεία και το τρίτο ένα άρθρο της Butler Analytics73 σχετικά με πέντε δωρεάν εργαλεία εξόρυξης δεδομένων. Καθώς τα αποτελέσματα της αναζήτησης που ήταν επιτυχή ήταν το δεύτερο και το τρίτο επιλέχτηκε αυτό που ταίριαζε περισσότερο στο σκοπό της έρευνας και ήταν η ιστοσελίδα με τα πέντε δωρεάν εργαλεία εξόρυξης. Η συγκεκριμένη αξιολόγηση έγινε από έναν υποψήφιο διδακτορικό φοιτητή ο οποίος βρίσκεται στα πρώτα στάδια της έρευνάς του πάνω στην εξόρυξη δεδομένων και κειμένου σε ψηφιακές ακαδημαϊκές εκδόσεις. Συνεπώς θα μπορούσε να θεωρηθεί ότι κατέχει επαρκείς γνώσεις γύρω από τις έννοιες της εξόρυξης δεδομένων και των εργαλείων της.

---

71 http://en.wikipedia.org/wiki/Text_mining, τελευταία πρόσβαση 08/102013

72 http://www.kdnuggets.com/software/text.html, τελευταία πρόσβαση 08/102013

73 http://butleranalytics.com/5-free-text-mining-tools/, τελευταία πρόσβαση 08/102013

Εργαλεία εξόρυξης κειμένου

Τα εργαλεία τα οποία χρησιμοποιήθηκαν για την αξιολόγηση παρουσιάζονται στον πίνακα 2:

Εταιρεία	Προϊόν	Έκδοση	Αρχείο Εγκατάστασης (MB)	Απαιτήσεις εγκατάστασης (MB)
GATE	GATE	7.1	359	566
KNIMEtech	KNIME Text Processing	2.8.1	151	271
University of Illionois at Chicago	LPU	1.0	1,43	4,75
PyCharm	Orange-text	2.7	103	269
Rapid-I	Rapid Miner	5.3.013	79,3	205

Πίνακας 2. Εργαλεία εξόρυξης κειμένου προς αξιολόγηση

GATE (General Architecture for Text Engineering)

Το GATE είναι ένα μεγάλο full-lifecycle ανοικτού κώδικα, εξόρυξης κειμένου λογισμικό με συστατικά:

- GATE Developer που είναι ένα ολοκληρωμένο περιβάλλον και αποτελείται από στοιχεία επεξεργασίας γλώσσας, που ενσωματώνουν το ευρέως χρησιμοποιούμενο σύστημα εξαγωγή πληροφοριών μαζί με άλλα plugins.
- GATE Teamware το οποίο παρέχει ένα περιβάλλον συνεργασίας για σχολιασμό εγγράφων. Αυτό είναι χτισμένο γύρω από ένα παράδειγμα ροής εργασίας.
- GATE Embedded που είναι μια βιβλιοθήκη αντικειμένων Java για να παρέχουν μια διεπαφή με άλλες εφαρμογές του οργανισμού.

Πρόκειται για ένα λογισμικό ανοικτού κώδικα, το οποίο είναι αποτέλεσμα ενός R&D προγράμματος χρηματοδοτούμενο με πολλά εκατομμυρίων ευρώ από σημαντικές εταιρείες από το 1995 και είναι ικανό να επιλύσει σχεδόν όλα τα προβλήματα επεξεργασίας κείμενου. Το GATE υποστηρίζεται από μια ώριμη και εκτεταμένη κοινότητα προγραμματιστών, χρηστών, εκπαιδευτών, φοιτητών και επιστημόνων. Εφαρμόζεται άμεσα και ενεργά σε όλα τα είδη και εφαρμογές επεξεργασίας κειμένου συμπεριλαμβανομένων: άποψη πελάτη, έρευνα για καρκίνο και φάρμακα, υποστήριξη λήψης αποφάσεων, επάνδρωση προσωπικού, web-mining, εξαγωγή πληροφοριών και σημασιολογία.

Οι προγραμματιστές του GATE βασίστηκαν στην τεχνολογία της Java για να δημιουργήσουν το εργαλείο. Τις λειτουργίες τους τις κατανέμησαν σε 4 Εφαρμογές (Applications), Πόροι Γλώσσας (Language Resources), Πόροι Επεξεργασίας (Processing resources) και Αποθήκευση Δεδομένων (Datastores). Έχουν ενσωματώσει 71 plugins που παρέχουν 190 βοηθητικά αντικείμενα για τις διεργασίες του προγράμματος.

KNIME Text Processing

Το πρόγραμμα επεξεργασία κειμένου KNIME είναι ένα plug-in για τη (δωρεάν) σουίτα εξόρυξης δεδομένων KNIME. Υποστηρίζει μια διαδικασία επεξεργασίας κειμένου έξι βημάτων που ξεκινά με την ανάγνωση και την ανάλυση του κειμένου, ακολουθούμενη από αναγνώριση οντοτήτων, φιλτράρισμα και χειραγώγηση, καταμέτρηση λέξεων και εξαγωγή λέξεων-κλειδί, τόξο και φορέα (bow and vector representation) απεικόνισης, και τέλος οπτικοποίηση αποτελεσμάτων.

Το KNIME επιτρέπει την ανάγνωση, επεξεργασία, εξόρυξη και απεικόνιση δεδομένων κειμένου με ένα βολικό και χρηστικό τρόπο. Παρέχει λειτουργικότητα επεξεργασίας

φυσικής γλώσσας (NLP), εξόρυξη γνώσης από κείμενα και ανάκτησης πληροφοριών. Για να γίνει η εγκατάσταση του KNIME απαραίτητη προϋπόθεση να είναι εγκατεστημένη στον Η/Υ η Java. Το εργαλείο προσφέρει συνολικά 250 λειτουργίες (functions) οι οποίες έχουν ομαδοποιηθεί σε 10 κατηγορίες:

Κατηγορίες	Αριθμός
Είσοδος-Έξοδος (IO)	21
Βάση Δεδομένων (Database)	11
Χειραγώγηση Δεδομένων (Data Manupulation)	82
Προβολές Δεδομένων (Data Views)	21
Στατιστικά (Statistics)	14
Εξόρυξη (Mining)	39
Μεταδεδομένα (Meta)	6
Έλεγχος Ροής (Flow Control)	37
Διάφορα (Misc)	9
Χρόνος (Time Series)	10

Πίνακας 3. Κατηγορίες και λειτουργίες του KNIME

LPU (Learning from Positive and Unlabeled examples)

Το LPU (που σημαίνει μάθηση από θετικά και χωρίς ετικέτα παραδείγματα) είναι ένα σύστημα μάθησης και ταξινόμησης κειμένου που χρησιμοποιεί τεχνικές υποστήριξης φορέων μηχανής (SVM) και EM (προσδοκία μεγιστοποίησης). Τρέχει σε ένα παράθυρο DOS (τεχνολογία εντελώς παρωχημένη). Αυτός ο τύπος μάθησης είναι διαφορετικός από το κλασική μάθηση/ταξινόμηση κειμένου. Συνολικά παρέχει 28 λειτουργίες και επιλογές, αριθμός ο οποίος θεωρείται πολύ μικρός. Για την εγκατάσταση του είναι προαπαιτούμενη η εγκατάσταση του προγράμματος SVM-light.

Επιλογές	Αριθμός
Ορίσματα (Arguments)	2
Γενικές επιλογές (General options)	2
Μάθησης (Learning options)	6
Εκτίμησης Απόδοσης (Performance estimation options)	3
Μετατροπής (Transduction options)	1
Πυρήνα (Kernel options)	6
Βελτιστοποίησης (Optimization options)	6
Εξόδου (Output options)	2

Πίνακας 4. Επιλογές και λειτουργίες του LPU

Orange-text

Το Orange-text είναι ένα πρόσθετο (add-in) της δωρεάν σουίτας εξόρυξης δεδομένων Orange. Λειτουργεί με τα οπτικά εργαλεία ανάλυσης που παρέχονται από την Orange και προσθέτει τη δυνατότητα να επεξεργάζονται αδόμητα δεδομένα. Επιπροσθέτως προσφέρει την οπτικοποίηση και την ανάλυση δεδομένων ανοικτού κώδικα για αρχάριους αλλά και έμπειρους χρήστες. Το εργαλείο παρέχει επίσης εξόρυξη δεδομένων μέσα από οπτικό προγραμματισμού ή γλώσσας δέσμης ενεργειών (scripting) της δυναμικής γλώσσας προγραμματισμού Python. Η γλώσσα Python είναι απαιτούμενη για την εγκατάσταση του Orange, αλλά η εγκατάσταση της γίνεται κατά την εκκίνηση εγκατάστασης του Orange. Τέλος επιπλέον λειτουργίες που διαθέτει το εργαλείο είναι συστατικά για μάθηση μηχανής (machine learning), πρόσθετα Βιοπληροφορικής και

εξόρυξης κειμένου καθώς και ολοκληρωμένα πακέτα για ανάλυση δεδομένων. Συνολικά παρέχει 125 λειτουργίες ομαδοποιημένες σε 9 κατηγορίες:

Κατηγορίες	Αριθμός
Δεδομένα (Data)	22
Οπτικοποίηση (Visualize)	13
Ταξινόμηση (Classify)	17
Παλινδρόμηση (Regression)	10
Αξιολόγηση (Evaluate)	6
Χωρίς Επιβλεψη (Unsupervized)	14
Συσχετισμός (Associate)	5
Οπτικοποίηση ποιότητας (VisualizeQt)	8
Πρωτότυπα (Prototypes)	30

Πίνακας 5. Επιλογές και λειτουργίες του LPU

Rapid-Miner

Σύμφωνα με την ίδια την εταιρεία το Rapid-Miner είναι αναμφισβήτητα το παγκοσμίως κορυφαίο open source σύστημα εξόρυξης δεδομένων. Είναι ελεύθερα διαθέσιμο ως ένα αυτόνομο πρόγραμμα για την εξόρυξη, ανάλυση, αξιολόγηση και απεικόνιση δεδομένων και ως μια μηχανή εξόρυξης δεδομένων για την ενσωμάτωση προσωπικών προϊόντων.

Το Rapid-Miner παρέχει τελεστές και λειτουργίες για ανάλυση στατιστικών κειμένου. Υποστηρίζει πολλές πηγές δεδομένων συμπεριλαμβανομένων απλού κειμένου, HTML και pdf, καθώς και ένα μεγάλο αριθμό από τεχνικές φιλτραρίσματος. Όλα τα παραπάνω είναι ενσωματωμένα εντός ενός γραφικού περιβάλλοντος και πολλές εργασίες μπορούν να ολοκληρωθούν μέσω της drag and drop λειτουργίας. Μερικά από τα κύρια χαρακτηριστικά του είναι ότι τρέχει σε κάθε σημαντική πλατφόρμα και λειτουργικό σύστημα, με διαισθητική διαδικασία σχεδιασμού, πολυεπίπεδη προβολή δεδομένων που εξασφαλίζει την αποτελεσματική διαχείριση τους, GUI λειτουργία, λειτουργία διακομιστή (γραμμή εντολών), ή πρόσβαση μέσω των API της Java. Προσφέρει πάνω από 700 λειτουργίες μεταξύ άλλων δυνατότητες ισχυρά υψηλής-διαστατικής αποτύπωσης, τυποποιημένη μορφή ανταλλαγής XML για διεργασίες, ενσωματωμένη τη βιβλιοθήκη μηχανής μάθησης WEKA και πρόσβαση σε αρχεία προέλευσης δεδομένων όπως Excel, Access, Oracle, IBM DB2, Microsoft SQL, Sybase, Ingres, MySQL, Postgres, SPSS, dBase, αρχεία κειμένου και πολλά περισσότερα. Τέλος προσφέρει την περιεκτικότερη λύση εξόρυξης δεδομένων όσον αφορά την ενοποίηση των δεδομένων, τη μετατροπή και την μοντελοποίηση μεθόδων.

Συνολικά παρέχει 714 λειτουργίες/χειρισμούς (operators) ομαδοποιημένες σε 12 κατηγορίες:

Κατηγορίες	Αριθμός
Διαδικασία ελέγχου (Process control)	39
Βοηθητικές λειτουργίες (Utility)	54
Πρόσβαση Αποθετηρίου (Repository Access)	6
Εισαγωγή (Import)	28
Εξαγωγή (Export)	19
Μετασχηματισμός δεδομένων (Data transformation)	115
Μοντελοποίηση (Modeling)	263
Αξιολόγηση (Evaluation)	32
Επεξεργασία κειμένου (Text processing)	51
Web εξόρυξη (web mining)	14
Σειρές (Series)	87
Αναφορές (Reporting)	6

Πίνακας 6. Κατηγορίες και λειτουργίες τους του Rapid-Miner

SWOT Analysis

Μετά από έρευνα και με τη βοήθεια εργασιών πάνω στην εφαρμογή Ανάλυσης SWOT σε λογισμικό [10], [11], [19], [20] αλλά και μέσω πηγών από το Διαδίκτυο [21] δημιουργήθηκε μια λίστα με 90 κριτήρια εκ των οποίων τα 38 αφορούν στις Δυνάμεις (S), 33 στις αδυναμίες (W), 11 στις Ευκαιρίες (O) και 8 στις Απειλές (T). Πάνω σε αυτά έγινε η γενική αξιολόγηση και Ανάλυση SWOT και δημιουργήθηκαν οι Πίνακες 7 έως 10 όπου φαίνεται αν το κάθε εργαλείο ικανοποιεί τα συγκεκριμένα κριτήρια. Για να δοθεί μια πιο ευκρινής εικόνα για την ποιότητα, τη λειτουργικότητα και τη χρησιμότητα του κάθε εργαλείου αποφασίστηκε να δίνεται +1 βαθμός για κάθε Δύναμη S και Opportunity O και -1 βαθμός για κάθε Αδυναμία W και για κάθε Απειλή T. Θεωρήθηκε ότι τα κριτήρια είναι ισοδύναμα μεταξύ τους και ότι δεν έχουν κάποια μεγαλύτερη βαρύτητα. Παρόλα αυτά δε προκύπτει μια συνολική βαθμολογία για κάθε εργαλείο, αλλά δύο συνολικές με την μια να ανταποκρίνεται στο άθροισμα για τους εσωτερικούς παράγοντες (S και W) και η δεύτερη για τους εξωτερικούς (T και O).

Strengths - Δυνάμεις - S	GATE	KNIME	LPU	Orange	Rapid Miner
Λειτουργικότητα	X	X		X	X
Ευκολία στη χρήση		X		X	
Εύκολη προσαρμοστικότητα από τελικό χρήστη χωρίς τεχνικές ικανότητες		X		X	
Παρέχει δομή σε μη έμπειρο προσωπικό			X		
Σχετικότητα	X	X	X	X	X
Φιλικό προς το χρήστη	X	X		X	X
Ανταγωνιστικό πλεονέκτημα	X	X		X	X
Αγορά στόχος	X	X	X	X	X
Καινοτόμο	X	X		X	X
Προσφέρει κάτι νέο					X
Καθήκον ευκολότερο να ολοκληρωθεί		X		X	
Ποιότητα	X	X		X	X
Λειτουργίες που παρέχονται	X	X		X	X
Ευελιξία με μορφή προσαρμογής, σχεδιασμού και ανάπτυξης	X	X		X	X
Προσανατολισμένο στην ανάγκη ανάπτυξης	X	X		X	X
Ορόσημα σαφώς καθορισμένα και κατανοητά		X		X	X
Διαχείριση του προγράμματος για παρακολούθηση εξέλιξης μέσω οροσήμων	X	X		X	X
Ορίζει απαιτήσεις σταθερότητας	X	X		X	X
Κάθε φάση έχει σαφώς καθορισμένες εισόδους και εξόδους δεδομένων	X	X	X	X	X
Οτιδήποτε παραδοτέο πρέπει να δοκιμαστεί	X	X		X	X
Πολύ ευέλικτο, καθώς οι αλλαγές στις απαιτήσεις μπορούν να διευθετηθούν πολύ πιο εύκολα με κάθε νέα αναθεώρηση και βελτίωση	X	X		X	X
Απροσδόκητες απαιτήσεις μπορούν να εξυπηρετηθούν					X
Σταθερό, ορατά σημάδια της παραγόμενης προόδου	X				X
Κατάρτιση για χρήστες με ελάχιστες απαιτήσεις			X		
Επικέντρωση/εστίαση μετακινείται από την τεκμηρίωση στον κώδικα (WYSIWYG)			X		
Χρήση μοντελοποιημένων εννοιών για την καταγραφή πληροφοριών σχετικά με την εφαρμογή, δεδομένα και διεργασίες	X	X		X	X
Αυξάνει την επαναχρησιμοποίηση των συστατικών/τμημάτων					X
Υψηλή τμηματοποίηση επιτυγχάνει ένα πιο ευέλικτο και διατηρήσιμο σύστημα.					
Γρήγορα αρχικά σχόλια εμφανίζονται	X	X		X	X
Ενθαρρύνει την ανατροφοδότηση των πελατών					X
Χρήση "διαίρει και βασίλευε" για την κατανομή καθηκόντων	X	X		X	X
Υποστήριξη					
online videos	X				X
FAQ's	X	X		X	X
εγχειρίδιο	X	X		X	X
tutorials	X	X		X	X
σεμινάρια					X
Ενημερωτικό δελτίο	X	X		X	X
ΣΥΝΟΛΟ:	24	26	6	26	30

Πίνακας 7. Δυνάμεις (Strengths) των 5 εργαλείων

Weaknesses - Αδυναμίες - W	GATE	KNIME	LPU	Orange	Rapid Miner
Δυσκολία στη χρήση			X		
Μη εμπορευσιμότητα			X		
Κόστος παραγωγής					
Έλλειψη ανταγωνιστικού πλεονεκτήματος					
Δεν παρέχει νέες λειτουργίες			X		
Απαιτήσεις συστήματος	X			X	
Είναι "εχθρικό" προς το χρήστη;			X		
Δεν το αποδέχονται οι κορυφαίοι παίκτες του κλάδου;			X		
Έλλειψη υποστήριξης			X		
Καμία ευθύνη και λογοδοσία					
Προηγμένες δεξιότητες που απαιτούνται	X			X	X
Κόστος για υποστήριξης στο εσωτερικό, ανάπτυξη, εκπαίδευση, διαχείριση των οργανωτικών αλλαγών					
Μη δυνατότητες ενσωμάτωσης			X		
Ευρύ αλλά όχι βαθύ					
Έλλειψη προηγμένων εργαλεία			X		
Καμία κινητικότητα			X		
Όχι σε απευθείας σύνδεση με το οικοσύστημα			X		
Αναστέλλει την ευελιξία			X		
Μεγαλύτερο χρονικό διάστημα παράδοσης απτών αποτελεσμάτων					
Δεν αντικατοπτρίζει επίλυση προβλημάτων ανάπτυξης λογισμικού φύσεως δηλαδή επαναλήψεις των φάσεων			X		
Μικρή ευκαιρία για τον πελάτη για να κάνετε προεπισκόπηση του συστήματος			X		
Ακατάλληλο για μεγάλα έργα και όπου οι απαιτήσεις δεν είναι σαφείς			X		
Δεν χειρίζεται άνετα ταυτόχρονα καθήκοντα			X		
Δεν χειρίζεται επαναλήψεις ή φάσεις					
Χρειάζεται άφθονους εξειδικευμένους πόρους	X	X		X	X
Δεν χειρίζεται εύκολα δυναμικές αλλαγές			X		
Συνολικά μπορεί να αγνοηθεί η συντηρησιμότητα			X		
Δύσκολο να χρησιμοποιηθεί με τα κληροδοτημένα συστήματα			X		
Απαιτεί ένα σύστημα που μπορεί να είναι διαμορφωμένο	X			X	
Παρωχημένο λειτουργικό σύστημα			X		
Ακατάλληλο για σύνθετα έργα			X		
Προαπαιτούμενο εγκατεστημένο λογισμικό	X			X	
Χρονοβόρα εγκατάσταση	X				
ΣΥΝΟΛΟ:	**6**	**1**	**20**	**5**	**2**

Πίνακας 8. Αδυναμίες (Weaknesses) των 5 εργαλείων

Opportunities - Δυνατότητες - O	GATE	KNIME	LPU	Orange	Rapid Miner
Εξωτερικοί παράγοντες μπορεί να επηρεάσουν το προϊόν	X	X		X	X
Τεχνολογική ανάπτυξη και καινοτομία	X	X		X	X
Τι συμβαίνει στη βιομηχανία;	X	X		X	X
Υπάρχουν οποιαδήποτε θέση αγορές-στόχοι για περαιτέρω μάρκετινγκ;	X	X		X	X
Τυχόν σημαντικές ανακαλύψεις, θέματα ή γεγονότα που καθιστούν το προϊόν να ξεχωρίζει;					X
Τρωτά σημεία των ανταγωνιστών	X	X		X	X
Τάσεις που αναπτύσσονται στη βιομηχανία	X	X	X	X	X
Δυσαρέσκεια με και δημιουργία αντίθετου πόλου σε σχέση με τους εμπορικούς πωλητές	X	X		X	X
Όταν αυτό μπορεί να τροποποιηθεί για να χειριστεί τις μεταβαλλόμενες απαιτήσεις πέρα από το στάδιο ανάλυσης	X	X		X	X
Απαιτήσεις είναι πολύπλοκες	X	X		X	X
Σημαντικές αλλαγές αναμένονται (έρευνας και αναζήτησης)	X	X	X	X	X
ΣΥΝΟΛΟ:	**10**	**10**	**2**	**10**	**11**

Πίνακας 9. Ευκαιρίες (Opportunities) των 5 εργαλείων

Threats - Απειλές - T	GATE	KNIME	LPU	Orange	Rapid Miner
Τυχόν εξωτερικοί παραγόντες που ενδέχεται να παρεμποδίσουν την πρόοδο ή την εμπορευσιμότητα					
Ζήτηση			X		
Δημογραφικά στοιχεία της ομάδας στόχου			X		
Νομικοί κανονισμοί					
Σκιώδες αγορά ανοιχτού κώδικα	X	X	X	X	X
Μοναδικές απαιτήσεις			X		
Εμπορικοί πωλητές δεν υποστηρίζουν την ενσωμάτωση με Open Source στοιχεία	X	X		X	X
Μη ανάπτυξη μπορεί να οδηγήσει σε λήθη			X		
ΣΥΝΟΛΟ:	2	2	5	2	2

Πίνακας 10. Απειλές (Threats) των 5 εργαλείων

Σύμφωνα με την προαναφερθείσα μεθοδολογία τα συνολικά σκορ των εργαλείων ανά εσωτερικούς και εξωτερικούς παράγονται φαίνονται στον πίνακα 11.

Εργαλείο	Εσωτερικοί παράγοντες (S+W)	Εξωτερικοί παράγοντες (O+T)
GATE	18	8
KNIME	25	8
LPU	-14	-3
Orange	21	8
Rapid Miner	28	9

Πίνακας 11. Συνολικά σκορ εσωτερικών και εξωτερικών παραγόντων

Σύμφωνα με τον παραπάνω πίνακα, το Rapid-Miner ξεχωρίζει, με το KNIME να ακολουθεί, με τρίτο το Orange, τέταρτο το GATE και απελπιστικά τελευταία με πολύ κακή βαθμολογία το LPU.

Συμπεράσματα – Μελλοντική εργασία

Ο Jinha το 2010 [22] εκτίμησε ότι ο συνολικός αριθμός των άρθρων που έχουν δημοσιευθεί σε περιοδικά μέχρι το 2009 ήταν πάνω από 50 εκατομμύρια. Μια μελέτη Καναδικού Πανεπιστημίου το 2012 [23] έδειξε μια εκθετική αύξηση του αριθμού των ηλεκτρονικών περιοδικών καθώς και λήψεων (downloads) ηλεκτρονικών άρθρων. Η εκθετικά αυξανόμενη αυτή τάση για ηλεκτρονικές εκδόσεις, αν συνυπολογιστεί και η αύξηση των ηλεκτρονικών βιβλίων (e-books) καθιστά αναγκαία την ύπαρξη μηχανισμών εξόρυξης κειμένου. Ο Clark [24] προτείνει 4 κύριους λόγους για τους οποίους θα πρέπει να γίνεται εξόρυξη κειμένου: i) για τον εμπλουτισμό του περιεχομένου ii) για την υποβοήθηση βιβλιογραφικής ανασκόπησης iii) για την ανακάλυψη και iv) για την έρευνας της Υπολογιστικής Γλωσσολογίας (computational linguistics).

Για αυτό τον λόγο επιλέχθησαν 5 εργαλεία εξόρυξης κειμένου τα οποία είναι ελεύθερης πρόσβασης και ανοιχτού κώδικα, πάνω στα οποία εφαρμόσθηκε η Ανάλυση SWOT για να προσδιοριστούν οι Δυνάμεις, οι Αδυναμίες, οι Ευκαιρίες και οι Απειλές των εργαλείων αυτών. Η επιλογή των εργαλείων έγινε βάσει της απλούστερης δυνατής αναζήτησης σε μηχανή αναζήτηση. Συγκεντρώθηκαν 90 κριτήρια και μελετήθηκε σε βάθος κατά πόσο ικανοποιούνται από τα εργαλεία αυτά. Από τα 5 εργαλεία το Rapid-Miner ξεχώρισε σε όλα τα επίπεδα καθώς από πλευράς δυνάμεων ικανοποιούσε το 78,95% αυτών, με τα KNIME και Orange-Text να ακολουθούν με 68,42%, να έπεται το GATE με 63,16% και τελευταίο το LPU με μόνο 15,79%. Από πλευράς αδυναμιών το KNIME εμφάνισε μόνο 1 στις 33 (ποσοστό 3,03%), με δεύτερο το Rapid-Miner να ακολουθεί με ποσοστό 6,06%, έπειτα το Orange-Text με 15,15%, τέταρτο το GATE με

18,18% και μακράν το πιο αδύναμο το LPU με 60,6%. Όσον αφορά στους εξωτερικούς παράγοντες οι ευκαιρίες που μπορεί να εκμεταλλευτεί το Rapid-Miner έφτασαν στο απόλυτο 100%, με τα GATE, KNIME και Orange-text να βρίσκονται στο 90,9% και το LPU στο 18,18%. Τέλος από πλευράς απειλών μεγαλύτερο ποσοστό εμφάνισε πάλι το LPU με 62,5%, ενώ τα 4 υπόλοιπα εργαλεία εμφάνισαν 25%.

Η παρούσα μελέτη ήδη έχει θέσει τις βάσεις για μια εργασία που έχει σχεδιαστεί να είναι η εκ βάθους νέα αξιολόγηση SWOT πάνω σε εξειδικευμένα θέματα εξόρυξης κειμένου και δεδομένων. Η νέα ανάλυση θα γίνει αποκλειστικά και μόνο με κριτήρια όχι απλής χρήσης των εργαλείων αλλά για παράδειγμα του τι ιδιαιτερότητες και καινοτομίες προσφέρουν σε άκρως έμπειρους και εξειδικευμένους χρήστες του τομέα της εξόρυξης δεδομένων. Σε αυτή την νέα ανάλυση τα κριτήρια τα οποία θα επιλεγούν δε θα θεωρηθούν ισοδύναμα μεταξύ τους, αλλά θα διεξαχθεί έρευνα, θα δημιουργηθούν ερωτηματολόγια τα οποία θα δοθούν σε ειδικούς του τομέα έτσι ώστε να προκύψουν διαφορετικά βάρη για κάθε κριτήριο.

Βιβλιογραφία

JISC, 2012. *Value and benefits of text mining*. [online] Διαθέσιμο στο: <http://www.jisc.ac.uk/publications/reports/2012/value-and-benefits-of-text-mining.aspx#a01> [Πρόσβαση 1 Σεπτεμβρίου 2013].

Ware, M. and Mabe, M., 2009. The stm report: An overview of scientific and scholarly journal publishing. [online] Διαθέσιμο στο: <http://www.stm-assoc.org/2009_10_13_MWC_STM_Report.pdf> [Πρόσβαση 9 Σεπτεμβρίου 2013]

Manyika, J. et al., 2011. Big data: The next frontier for innovation, competition, and productivity. [pdf] McKinsey Global Institute. Διαθέσιμο στο: <http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights and pubs/MGI/Research/Technology and Innovation/Big Data/MGI_big_data_full_report.ashx> [Πρόσβαση 13 Σεπτεμβρίου 2013]

ARMA International. Chucking daisies introduction. [pdf] Διαθέσιμο στο: <http://www.arma.org/docs/bookstore/arma-chuckingdaisies-intro.pdf>. [Πρόσβαση 12 Σεπτεμβρίου 2013]

Marti Hearst. "What is Text Mining?" 17 October 2003. [online] Διαθέσιμο στο: <http://www.ischool.berkeley.edu/~hearst/text-mining.html> [Πρόσβαση 12 Σεπτεμβρίου 2013]

Fan, W., Wallace, L., Rich, S. and Zhang, Z. Tapping into the Power of Text Mining. Communications of ACM, 2005

Start D. and Hovland I. Tools for Policy Impact – A Handbook for Researchers. The Overseas Development Institute (ODI). [pdf] Διαθέσιμο στο: <http://www.odi.org.uk/sites/odi.org.uk/files/odi-assets/publications-opinion-files/194.pdf> [Πρόσβαση 16 Σεπτεμβρίου 2013]

Kim, J., et al. Conceptual Model of Intelligent Program Management Information Systems (iPMIS) for Urban Renewal Mega Projects. Journal of Asian Architecture and Building Engineering, Vol. 8(1), pp. 57-64, 2009.

SWOT Analysis. Wikipedia. [online] Διαθέσιμο στο: <http://en.wikipedia.org/wiki/SWOT_analysis> [Πρόσβαση 23 Σεπτεμβρίου 2013]

Wajszczuk Karol, Wawrzynowicz Jacek. A SWOT analysis of software for agriculture in Poland. 3rd AGRIMBA-AVA Congress (2013).

Sasankar A. and Chavan V. SWOT Analysis of Software Development Process Models. IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 3, September 2011 ISSN (Online): 1694-0814

Crowsey, M. J., Ramstad, A. R., Gutierrez, D. H., Paladino, G. W., and White, K. P., Jr.: An Evaluation of Unstructured Text Mining Software. In: Systems and Information Engineering Design Symposium 2007, IEEE, 1-6 (2007)

Centers for Disease Control and Prevention. Evaluation of Open Source Text Mining Tools for Cancer Surveillance: Phase I: Understanding text mining and identifying tools. [pdf] Διαθέσιμο στο <www.cdc.gov/cancer/npcr/pdf/aerro/text_mining_tools.pdf> [Πρόσβαση 18 Σεπτεμβρίου 2013]

Krallinger M, Morgan A, Smith L, et al. Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. Genome Biol.2008; 9 Suppl. 2:S1.

Van Germet J. Text Mining Tools on the Internet. Intelligent Sensory Information Systems (ISIS) technical report series. September, 2000.

A. Aula, R.M. Khan, and Z. Guan. How does search behavior change as search becomes more difficult? In CHI 2010, pages 35–44. ACM.

Tombros, A., I. Ruthven, and J.M. Jose, How Users Access Web Pages for Information Seeking. Journal of the American Society for Information Science and Technology, 2005. 56(4): p.327-344.

Hoelscher, C., & Strube, G. (2000). Web search behavior of Internet experts and newbies. Computer Networks 33(1-6), 337-346.

Schmuhl H., Bergh B. Strengths, weaknesses and barriers to adoption of Open Source Software in a clinical setting as seen by hospital CIOs. EFMI Special Topic Conference, Moscow in April 2012. Διαθέσιμο στο <http://www.apfelkraut.org/download/OS_Survey_H_Schmuhl.pdf> [Πρόσβαση 25 Σεπτεμβρίου 2013]

Dai, Y., Kakkonen, T. & Sutinen, E. 2011. MinEDec: A decision-support model that combines text-mining technologies with two competitive intelligence analysis method. International Journal of Computer Information System and Industrial Management Applications, 3, 165-173.

Mayberyy Stephanie. SWOT Analysis of Software. [online] Διαθέσιμο στο: <http://www.ehow.com/about_5549050_swot-analysis-software.html#ixzz2hFfsz9Vu> [Πρόσβαση 25 Σεπτεμβρίου 2013]

Jinha, A., 2010. Article 50 million: an estimate of the number of scholarly articles in existence. LearnedPublishing, 23 (3), 258-263

Lamothe, A., 2012. Factors Influencing Usage of an Electronic Journal Collection at a Medium-Size University: An Eleven-Year Study. Partnership: the Canadian Journal of Library and Information Practice and Research, Volume 7, issue 1. ISSN: 1911-9593

Clark, J. Text Mining and Scholarly Publishing, presented at Publishing Research Consortium, Loosdrecht, The Netherlands, 2013.