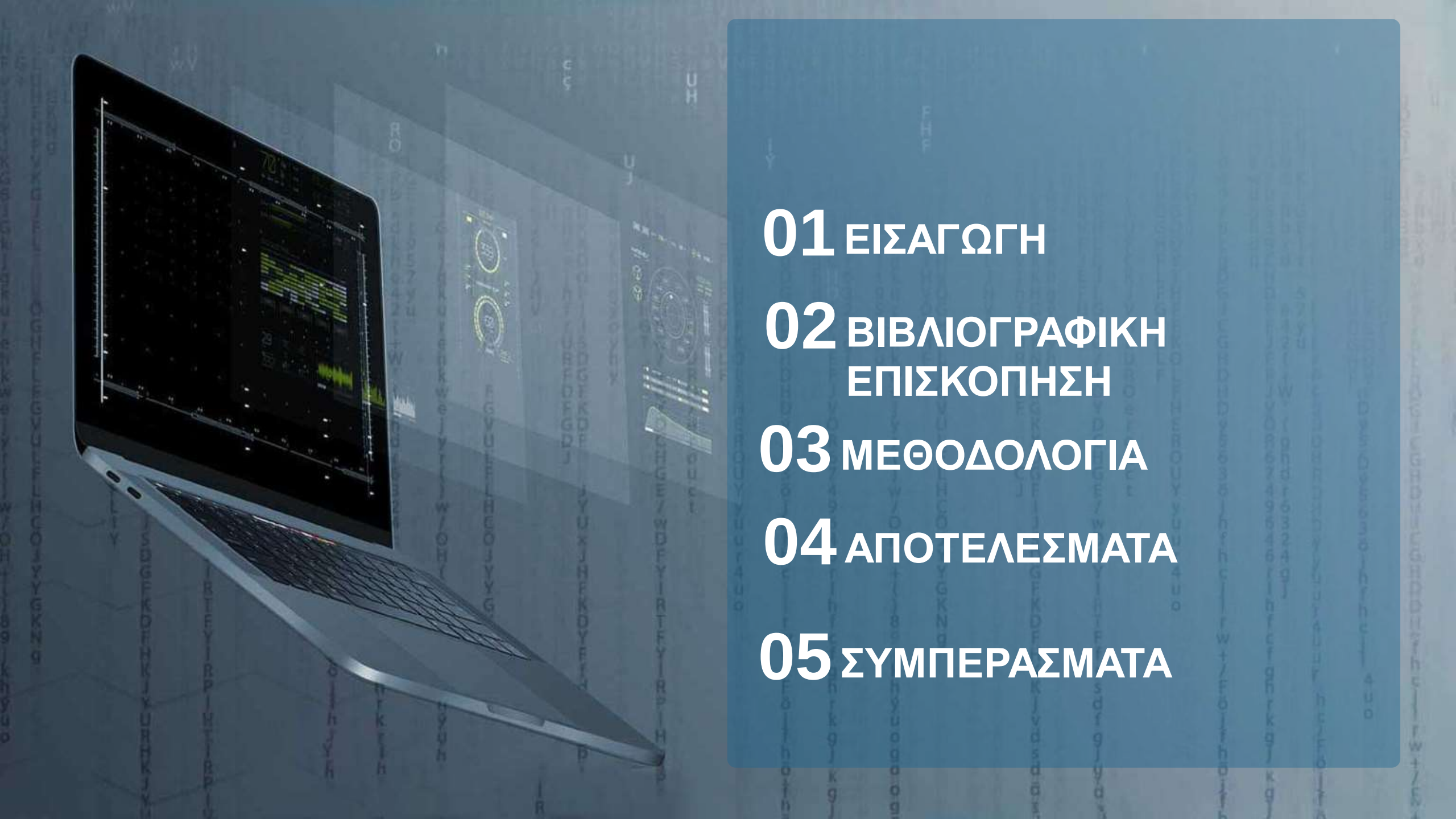
ΕΞΟΡΥΞΗ ΚΑΝΟΝΩΝ ΣΥΣΧΕΤΙΣΗΣ ΠΟΥ ΒΑΣΙΖΟΝΤΑΙ ΣΤΟ ΣΥΝΑΙΣΘΗΜΑ ΑΠΟ ΔΕΔΟΜΕΝΑ ΤΟΥ TWITTER

Μαρία Γεωργιάδου
mai20012@uom.gr
Επιβλέπουσα Καθηγήτρια: Γεωργία Κολωνιάρη
gkoloniari@uom.edu.gr



ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΑΠΟ ΔΕΔΟΜΕΝΑ

«Η εξόρυξη δεδομένων είναι μια δια-δραστική διαδικασία που περιλαμβάνει τη συγκέντρωση των δεδομένων σε μορφή ευνοϊκή για την ανάλυση. Μόλις διαμορφωθούν τα δεδομένα, πρέπει να καθαριστούν ελέγχοντας για προφανή σφάλματα ή ελαττώματα (όπως ένα στοιχείο που είναι εξαιρετικά ακραίο) και απλώς αφαιρώντας τα ».
(Forcht & Cochran, 1999)

"Οι αναζητήσεις εξόρυξης δεδομένων για κρυφές σχέσεις, μοτίβα, συσχετίσεις και αλληλεξαρτήσεις σε μεγάλες βάσεις δεδομένων που μπορεί να παραβλέψουν οι παραδοσιακές μέθοδοι συλλογής πληροφοριών".
(Gargano & Raggad, 1999)

«Η εξόρυξη δεδομένων είναι η διαδικασία εφαρμογής τεχνικών τεχνητής νοημοσύνης (όπως προηγμένη μοντελοποίηση και διέγερση κανόνων) σε ένα μεγάλο σύνολο δεδομένων για τον προσδιορισμό των προτύπων στα δεδομένα».
(Ma, 2000**)**

«Ως εξαγωγή γνώσεων, ανακάλυψη πληροφοριών, συλλογή πληροφοριών, διερευνητική ανάλυση δεδομένων, αρχαιολογία δεδομένων, επεξεργασία προτύπων δεδομένων και ανάλυση λειτουργικής εξάρτησης».
(Imberman, 2001)

«Η εξόρυξη δεδομένων είναι η διαδικασία εύρεσης συσχετίσεων ή μοτίβων ανάμεσα σε δεκάδες πεδία σε μεγάλες σχεσιακές βάσεις δεδομένων».
(Palace, 1996)

ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΑΠΟ ΔΕΔΟΜΕΝΑ

Ανάλυση αγοράς
Amazon (Ayres, 2007)

Εταιρική Ανάλυση

Ανίχνευση απάτης
πιστωτικές κάρτες
(Brause et al., (1999)

Ιατρική
Διάγνωση- Προδιάθεση
(Lavrač και Zupan (2005)

Τηλεόραση και ραδιόφωνο
Δημοτικότητα (Zhu, 2017)

Αστρονομία
προσδιορισμό των φυσικών παραμέτρων των γαλαξιών
(Wang, 2008)

ΜΟΝΤΕΛΑ

- Πολικότητα
- Συναισθήματα
- Ενδιαφέρον

ΜΕΘΟΔΟΙ

- Βαθμολόγηση με αστέρια
- Λεξικά
- Κατανομή Σχολίων

ΣΥΝΑΙΣΘΗΜΑΤΙΚΗ ΑΝΑΛΥΣΗ

Η συναισθηματική ανάλυση επιτυγχάνεται μέσα από την επεξεργασία φυσικής γλώσσας και προσδιορίζεται η συναισθηματική πολικότητα που υπάρχει στα κειμενικά δεδομένα.

ΣΤΟΧΟΣ ΕΡΕΥΝΑΣ

Στην παρούσα διπλωματική εργασία θα πραγματοποιηθεί, στο τομέα της εξόρυξης γνώσης, η δημιουργία κανόνων συσχέτισης, με δεδομένα κείμενου από το Twitter, με σκοπό την συσχέτιση μεταξύ λέξεων και συναισθημάτων. Λέξεις, που δεν έχουν συναισθηματική έννοια και φόρτιση και παρέχουν μια πληροφορία, θα ταυτιστούν με κάποιο συναίσθημα που αυτά προκαλούν στους χρήστες. Τα δεδομένα που θα χρησιμοποιήσουμε είναι μεγάλα σύνολα δεδομένων κειμένου από το Twitter, τα γνωστά σε όλους tweets, τα οποία έχουν ήδη ταξινομηθεί σε ένα συναίσθημα. Μέσα από επεξεργασία του συνόλου από το γενικό συναίσθημα για κάθε tweet,θα γίνει προσπάθεια άντλησης κανόνων συσχέτισης των συχνών λέξεων που εμφανίζονται στα κείμενα με την συναισθηματική φόρτιση που προκαλούν .

ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

(Pang και Lee, 2004), (Kennedy και Inkpen, 2006), (Moghaddam και Ester, 2010), (Pak et al, 2010), (Maas et al (2011), (Wang και Manning, 2012), (Narayanan et al (2013), (Amine et al, 2014), (HarishRao et al, 2017)

ΛΕΞΙΚΑ

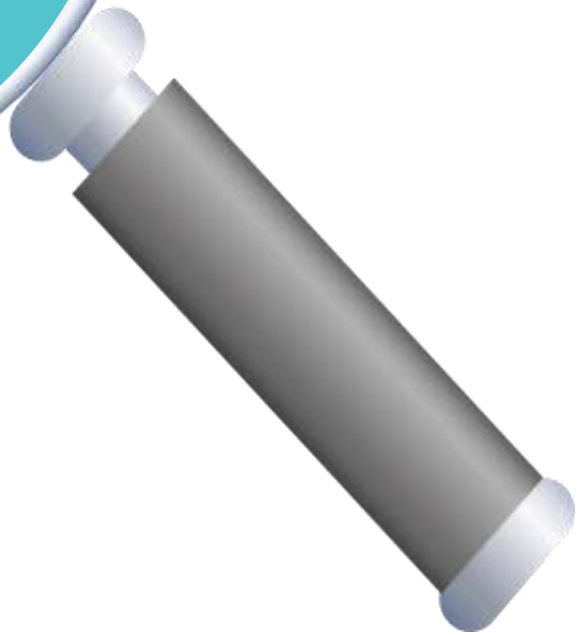
(Rajman και Besancon, 1997), (Zhuang et al, 2006), (Denecke, 2008), (Khan, 2011), (Kaur και Baghla, 2018)

ΜΕΤΡΑ- ΒΑΡΗ

(Benamara et al, 2006), (Liu et al, 2006), (Chen και Wu, 2008), (Abbasi et al, 2008)

ΚΑΤΑΤΑΞΗ-ΔΗΜΟΤΙΚΟΤΗΤΑ

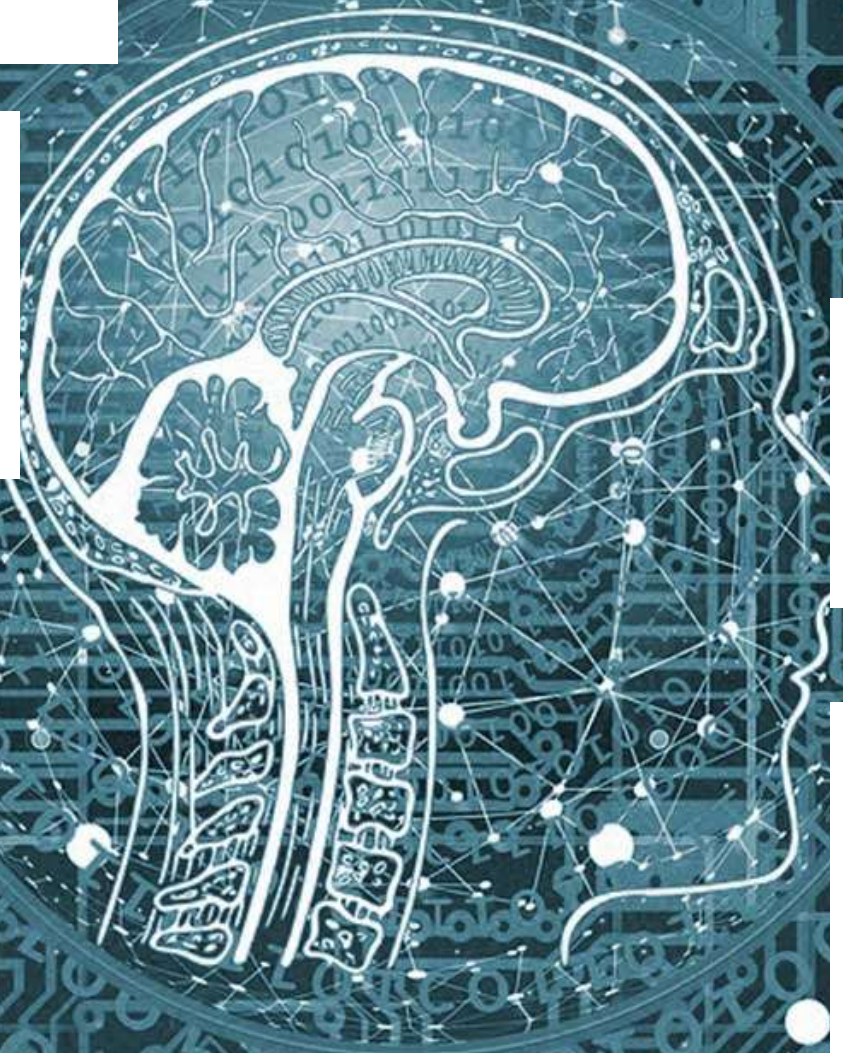
(Moghaddam και ο Ester, 2011), (Albornoz (2011), (Chen et al, 2013)



Επεξεργασία φυσικής γλώσσας (NLP) ασχολείται με τις αλληλεπιδράσεις μεταξύ των υπολογιστών
και των ανθρώπινων (φυσικών) γλωσσών (Βλαχάβας et al. , 2006)

ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ

A →B, με το A και το B να αποτελεί ένα στοιχείο από το σύνολο γνωρισμάτων I μέσα στα δεδομένα και $A \subseteq I$, $B \subseteq I$, $A \cap B = \emptyset$,



ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ

- Ανακάλυψη γνώσης από κείμενο (KDT)
- Εξόρυξη κειμένου (TM)

«Η ανακάλυψη γνώσης σε Κείμενο (KTD) είναι μια μη τετριμμένη διαδικασία ανακάλυψης έγκυρων, καινούργιων, δυνητικά χρήσιμων και τελικά κατανοητών προτύπων σε δεδομένα κειμένου» (Fayyad και Piatetsky – Shapiro, 1999)

ΟΡΙΣΜΟΙ

Στοιχειοσύνολο : «Ένα σύνολο από διακριτά στοιχεία ορίζεται ως $I=\{i_1, i_2, \ldots, i_k\}$, ενώ ένα στοιχειοσύνολο ορίζεται ένα υποσύνολο του I».

Συναλλαγές : « Ένα σύνολο από συναλλαγές παρουσιάζεται ως $T=\{t_1, t_2, \ldots, t_N\}$, όπου κάθε t_i είναι ένα στοιχειοσύνολο».

Υποστήριξη : «Υποστήριξη κανόνα-support (s) είναι το ποσοστό των συναλλαγών που περιέχουν το X και το Y ($X \cup Y$) με τύπο υπολογισμού $\sigma(X \cup Y)/ |T|$, όπου $\sigma(X \cup Y)$ η συχνότητα εμφάνισης του στοιχειοσυνόλου που περιέχει το X και το Y και $|T|$ ο αριθμός των δοσοληψιών».

Εμπιστοσύνη : «Εμπιστοσύνη – confidence (c) αναφέρει πόσες από τις συναλλαγές σε ποσοστό που περιέχουν και το X και το Y ($X \cup Y$) με τύπο $\sigma(X \cup Y)/\sigma(X)$».

Ελάχιστη υποστήριξη: «Ο κανόνας πρέπει να έχει υποστήριξη μεγαλύτερη από το όριο, που ονομάζεται ελάχιστη υποστήριξη (minsup)».

Ελάχιστη εμπιστοσύνη: «Ο κανόνας πρέπει να έχει εμπιστοσύνη μεγαλύτερη από το όριο, που ονομάζεται ελάχιστη εμπιστοσύνη (minconf)».

Συχνά στοιχειοσύνολα: «Τα στοιχειοσύνολα(itemsets) που έχουν υποστήριξη μεγαλύτερη από το minsup καλούνται συχνά ή μεγάλα».

Αρχή Apriori

«Αν ένα στοιχειοσύνολο είναι συχνό τότε όλα τα υποσύνολα του είναι συχνά»

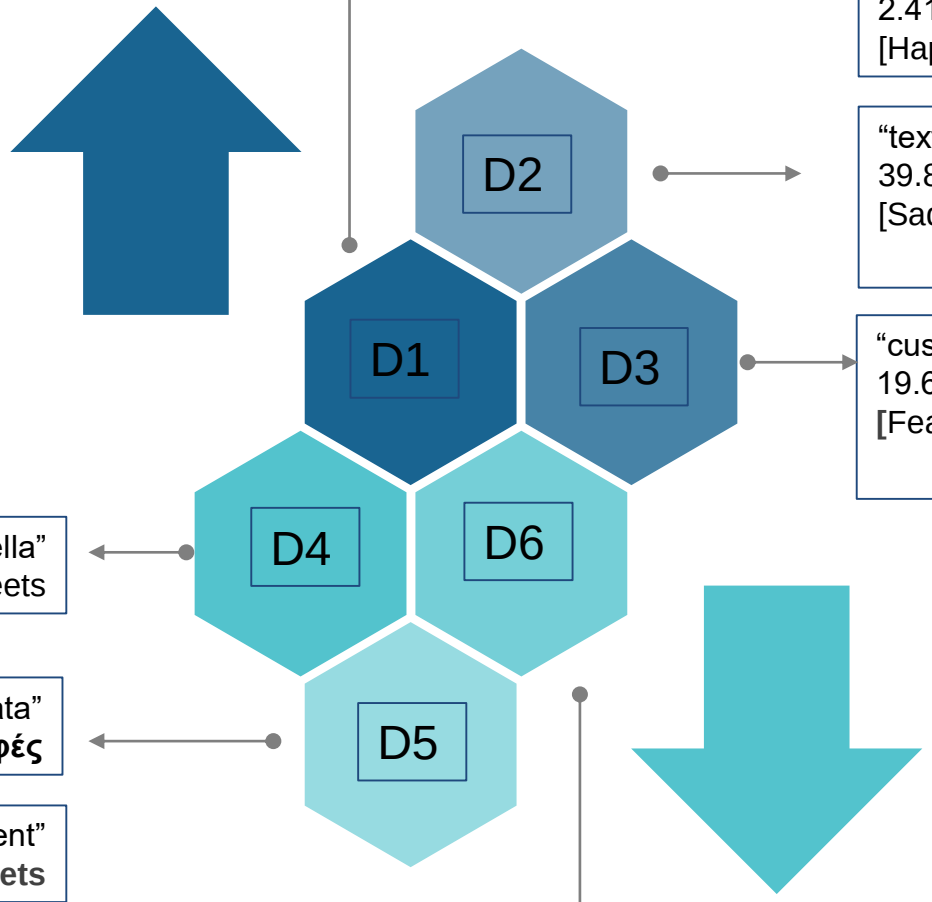
για κάθε X, Y ισχύει

$(X \subseteq Y) \Rightarrow S(X) \geq S(Y)$

«Αν ένα στοιχειοσύνολο είναι μη συχνό, τότε όλα τα υπερσύνολα του είναι μη συχνά».

ΔΕΔΟΜΕΝΑ

Στα σύνολα δεδομένων D4,D5 και D6 κάθε tweet συσχετίζεται με βάση την πολικότητα , δηλαδή θετικά/ αρνητικά



“emotion stimulus”
2.415 εγγραφές
[Happy, Sad, Anger, Fear, Shame]

“text emotion”
39.812 tweets
[Sad,Worry, Happy, Surprise, Love, Relief, Fun]

“customer support”
19.633 εγγραφές
[Fear, Anger, Surprise, Love, Joy, Sad]

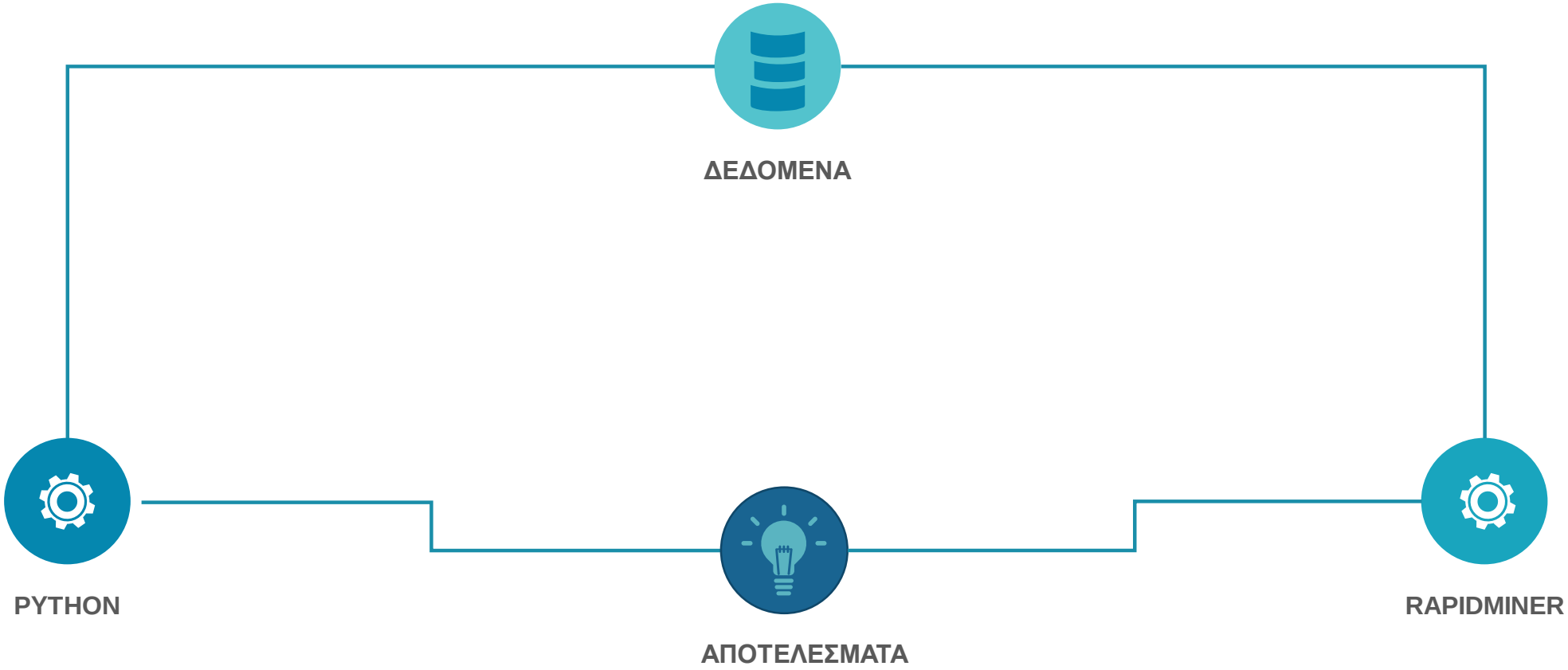
“Coachella”
4.173 tweets

“extra data”
12.502 **εγγραφές**

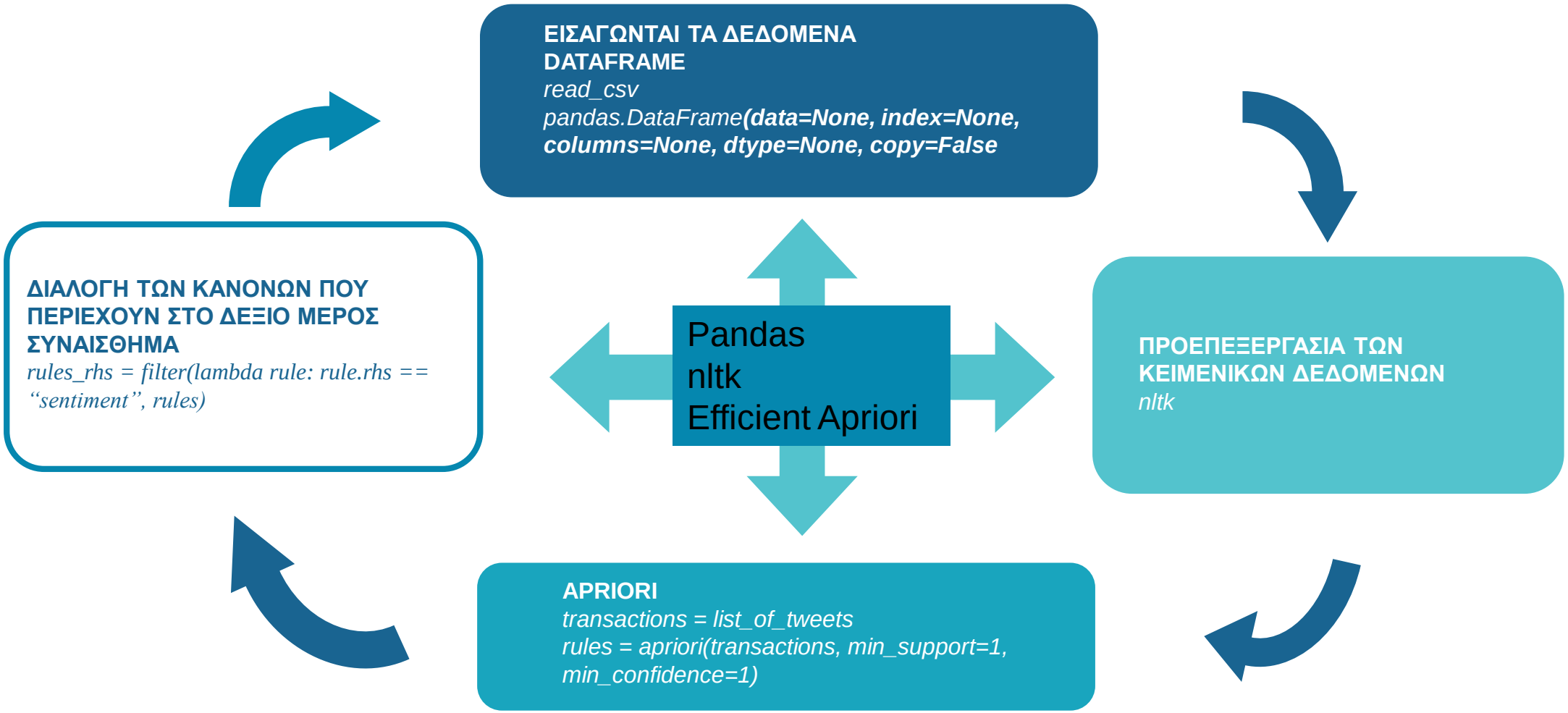
“twitter argument”
27.465 **tweets**

Τα αρχεία D1,D2 και D3 κατηγοριοποιούν κάθε εγγραφή σε ένα συναίσθημα

ΥΛΟΠΟΙΗΣΕΙΣ



PYTHON


ΕΙΣΑΓΩΝΤΑΙ ΤΑ ΔΕΔΟΜΕΝΑ
DATAFRAME
read_csv
pandas.DataFrame(data=None, index=None, columns=None, dtype=None, copy=False
ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΤΩΝ ΚΕΙΜΕΝΙΚΩΝ ΔΕΔΟΜΕΝΩΝ
nltk
APRIORI
transactions = list_of_tweets
rules = apriori(transactions, min_support=1, min_confidence=1)
ΔΙΑΛΟΓΗ ΤΩΝ ΚΑΝΟΝΩΝ ΠΟΥ ΠΕΡΙΕΧΟΥΝ ΣΤΟ ΔΕΞΙΟ ΜΕΡΟΣ ΣΥΝΑΙΣΘΗΜΑ
rules_rhs = filter(lambda rule: rule.rhs == "sentiment", rules)
Pandas
nltk
Efficient Apriori

ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

ΑΦΑΙΡΕΣΗ ΤΩΝ STOPWORDS

stopwords.words('english')
Π.χ. [happy, really, cheerful, groups, now]

ΛΗΜΜΑΤΟΠΟΙΗΣΗ (LEMMATIZATION)

lemma = WordNetLemmatizer()
lemma.lemmatize(word, pos = "n")
lemma.lemmatize(word, pos = "s")
Π.χ. [happy, really, cheerful, group, now]

ΑΠΟΚΑΤΑΛΗΞΗ (STEMMING)

PorterStemmer.stem(word)
SnowballStemmer.stem(word).
Π.χ. [happy, real, cheerful, groups, now]

ΛΕΞΙΚΟΦΡΑΦΙΚΗ ΑΝΑΛΥΣΗ (POS TAG)

postag(word)
.bigrams(postag)
Π.χ. [(happy, Adjective), (real, Adjective),
(cheerful, Adjective),
(groups, Noun), (now, Adverb)]

ΔΙΑΙΡΕΣΗ ΣΕ ΟΡΟΥΣ (TOKENIZATION)

.tokenize().
Π.χ. [happy, ",I, ',m, really, cheerful, in, the, GROUPS, NOW, !, ",]

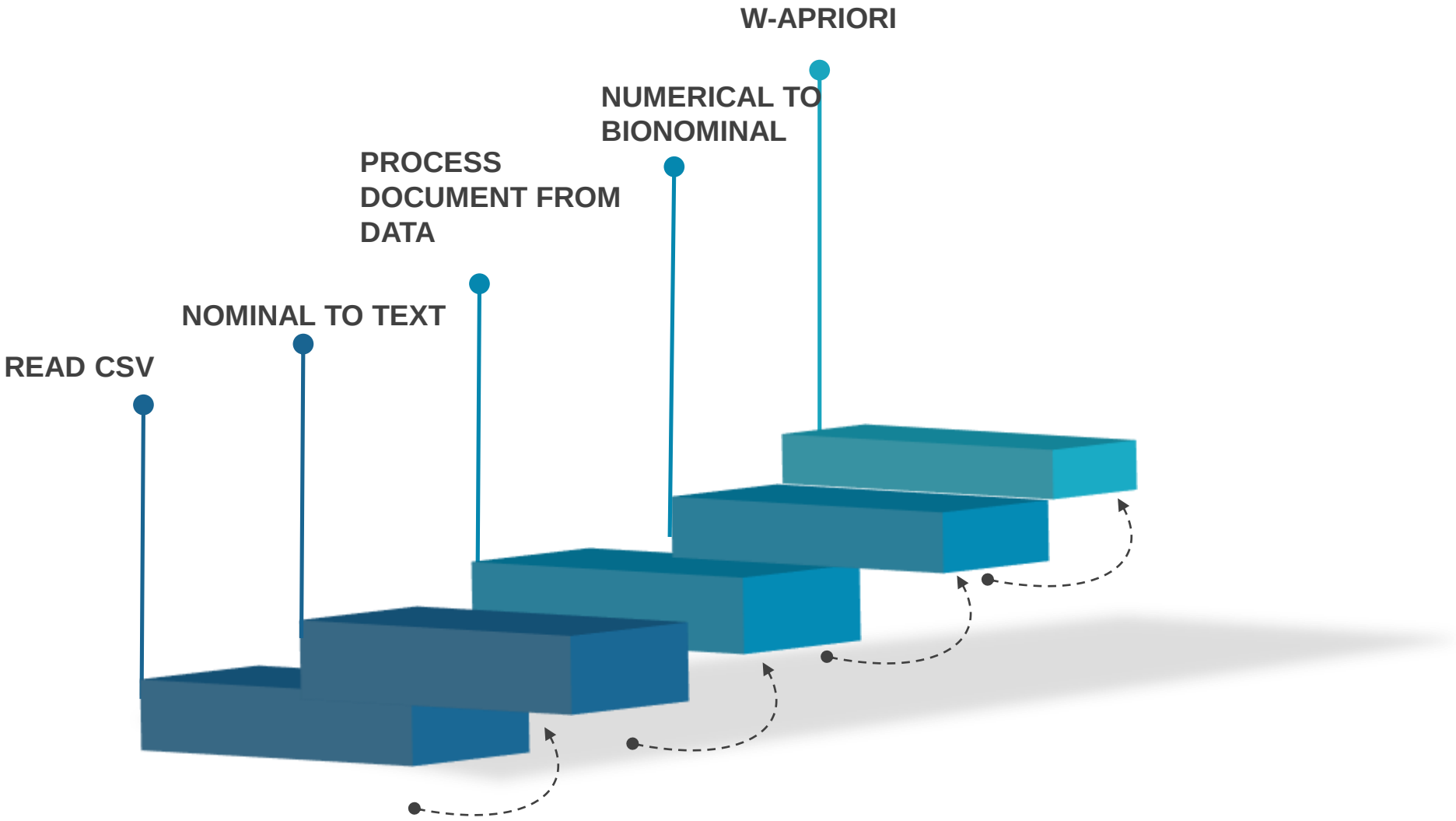
ΜΕΤΑΤΡΟΠΗ ΣΕ ΠΕΖΟΥΣ ΧΑΡΑΚΤΗΡΕΣ

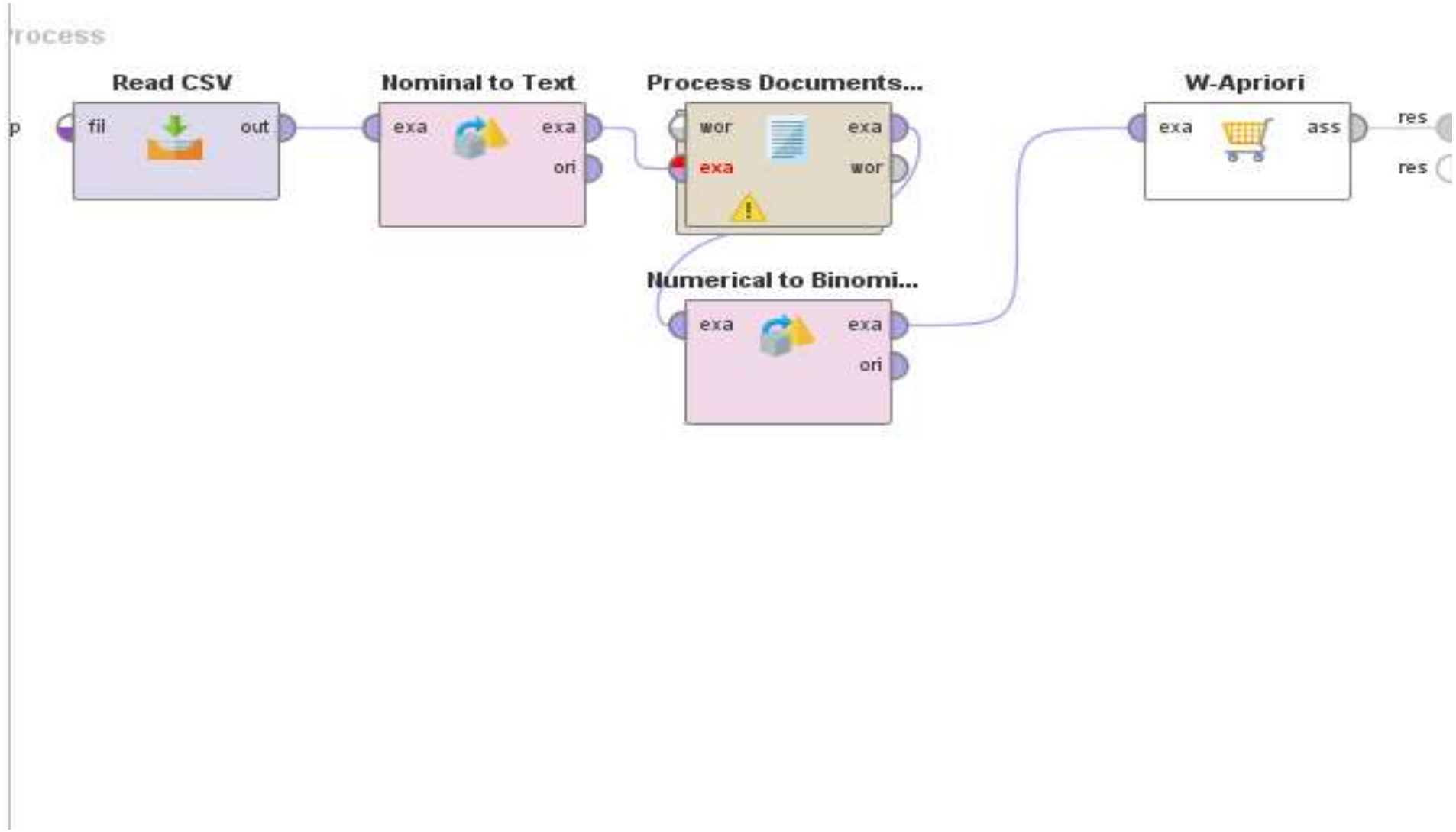
.lower().
Π.χ. [happy, ",I, ',m, really, cheerful, in, the, groups, now, !, ",]

ΑΦΑΙΡΕΣΗ ΑΡΙΘΜΩΝ ΚΑΙ ΣΥΜΒΟΛΩΝ

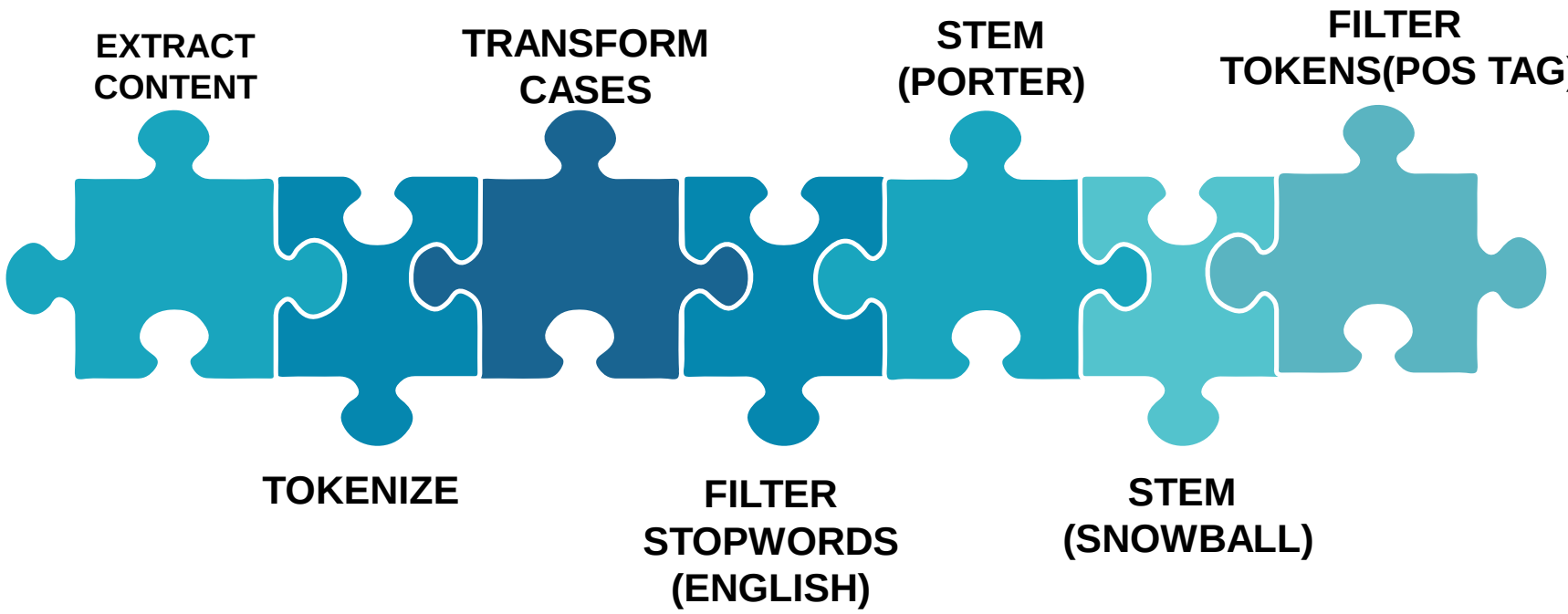
.isalpha(),
Π.χ. [happy, I, m, really, cheerful, in, the, groups, now]

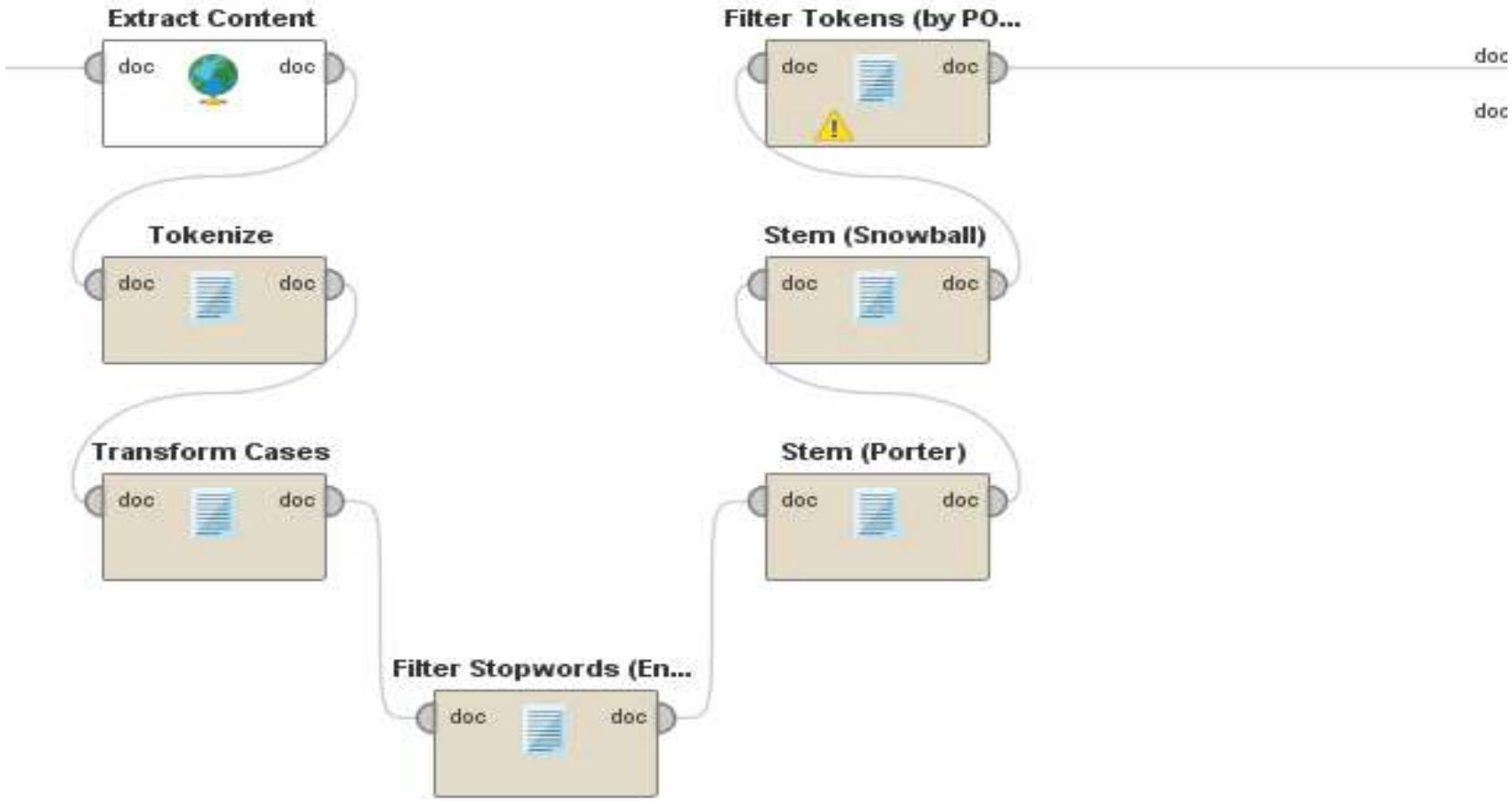
RAPIDMINER


W-APRIORI
NUMERICAL TO BIONOMINAL
PROCESS DOCUMENT FROM DATA
NOMINAL TO TEXT
READ CSV


rocess
Read CSV
fil
out
Nominal to Text
exa
exa
ori
Process Documents...
wor
exa
exa
wor
Numerical to Binomi...
exa
exa
ori
W-Apriori
exa
ass
res
res

ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ




Extract Content
doc
doc
Filter Tokens (by PO...
doc
doc
doc
doc
Tokenize
doc
doc
Stem (Snowball)
doc
doc
Transform Cases
doc
doc
Stem (Porter)
doc
doc
Filter Stopwords (En...
doc
doc

TL : διαίρεση του κειμένου σε όρους και μετατροπή των χαρακτήρων σε πεζούς

TLS: διαίρεση του κειμένου σε όρους, μετατροπή των χαρακτήρων σε πεζούς και αφαίρεση των Stopwords

TLSS: διαίρεση του κειμένου σε όρους, μετατροπή των χαρακτήρων σε πεζούς, αφαίρεση των Stopwords και αποκατάληξη

TLSL: διαίρεση του κειμένου σε όρους, μετατροπή των χαρακτήρων σε πεζούς, αφαίρεση των Stopwords και λημματοποίηση

TLSP: διαίρεση του κειμένου σε όρους, μετατροπή των χαρακτήρων σε πεζούς, αφαίρεση των Stopwords και λεξικογραφική ανάλυση

ΤΕΧΝΙΚΕΣ ΠΡΟΕΠΕΞΕΡΓΑΣΙΑΣ ΕΝ ΣΥΝΤΟΜΙΑ

	D1	D2	D3
min_sup=20% min_conf=20%	401	642	457
min_sup=30% min_conf=30%	119	67	182
min_sup=30% min_conf=37%	97	53	113
min_sup=30% min_conf=40%	82	22	102
min_sup=30% min_conf=45%	51	9	84

Πλήθος συχνών στοιχειοσυνόλων των D1-D3 με διάφορα ελάχιστα όρια

Πλήθος συχνών στοιχειοσυνόλων των D4-D6 με διάφορα ελάχιστα όρια

	D4	D5	D6
min_sup=30% min_conf=30%	556	811	749
min_sup=69% min_conf=69%	267	313	283
min_sup=80% min_conf=86%	89	115	102
min_sup=82% min_conf=89%	74	96	85

	D1	D2	D3	D4	D5	D6
min_sup=10% **min_conf=10%**	538	11.745	940	584	449	1.288
min_sup=20% **min_conf=20%**	199	385	185	370	403	519
min_sup=30% **min_conf=30%**	46	33	72	209	331	298
min_sup=30% **min_conf=37%**	42	14	44	188	239	221
min_sup=30% **min_conf=40%**	35	8	33	176	204	195
min_sup=30% **min_conf=45%**	9	3	19	142	180	157
min_sup=69% **min_conf=69%**	1	0	0	102	147	122
min_sup=80% **min_conf=86%**	0	0	0	67	79	71
min_sup=85% **min_conf=85%**	0	0	0	44	68	51
min_sup=82% **min_conf=89%**	0	0	0	31	49	43

Πλήθος κανόνων συσχέτισης με διάφορα ελάχιστα όρια

	D1	D2	D3
TL	89	71	96
TLS	42	14	44
TLSS	3	4	11
TLSL	0	2	8
TLSP(NOUN)	0	2	17
TLSP(ADJECTIVE)	0	1	15
TLSP(VERB)	0	0	10

Πλήθος κανόνων συσχέτισης με διάφορες τεχνικές προεπεξεργασίας στα D1-D3

Πλήθος κανόνων συσχέτισης με διάφορες τεχνικές προεπεξεργασίας στα D4-D6

	D4	D5	D6
TL	105	122	114
TLS	44	68	51
TLSS	1	33	1
TLSL	0	0	0
TLSP(NOUN)	3	2	18
TLSP(ADJECTIVE)	3	3	6
TLSP(VERB)	12	2	3

		TL	TLS	TLSS
1	STATE → HAPPY	59%/75%	69%/82%	55%/71%
2	VOICE → SHAME	37%/42%	40%/45%	35%/40%
3	WAR → FEAR	36%/43%	39%/44%	33%/40%
4	TIME → SAD	32%/47%	34%/48%	30%/30%
5	FACE → SAD	29%/43%	45%/51%	28%/33%

Κανόνες εξόρυξης για το αρχείο D1

		TL	TLS	TLSP (NOUN /ADJ/VERB)
1	STATE → HAPPY	59%/75%	69%/82%	29%/31%
2	ACCURATE →ANGER	30%/37%	31%/37%	22%/26%
3	TELL →FEAR	28%/34%	25%/31%	22%/27%

Κανόνες εξόρυξης
για το αρχείο D2

		TL	TLS	TLSS
1	STAR →HAPPY	54%/56%	49%/50%	50%/52%
2	WISH→WORRY	39%/47%	41%/48%	31%/39%
3	GREAT → WORRY	29%/39%	37%/39%	29%/37%
4	SUN → SAD	33%/37%	37%/40%	29%/36%
5	FRIEND → LOVE	26%/32%	34%/37%	26%/29%

		TL	TLS	TLSP (NOUN /ADJ/VERB)
1	STAR → HAPPY	54%/56%	49%/50%	21%/26%
2	ALTERNATIVE→SAD	31%/38%	28%/32%	30%/37%
3	CARE → LOVE	27%/30%	24%/29%	19%/22%

		TL	TLS	TLSS	TLSP (ADJECTIVE)
1	LITTLE→ LOVE	63%/67%	62%/65%	41%/43%	64%/69%
2	LAST → ANGER	54%/59%	54%/58%	37%/42%	58%/61%
3	HANDSOME→ JOY	50%/55%	47%/50%	37%/40%	51%/57%
4	FAMILIAR→ SUPRISE	48%/51%	47%/51%	35%/39%	48%/52%
5	NEXT → FEAR	45%/50%	44%/46%	32%/36%	46%/50%

		TL	TLS	TLSS	TLSP (NOUN)
1	SPEED→ FEAR	49%/52%	50%/53%	37%/40%	50%/53%
2	TEAM → SUPRISE	48%/50%	48%/51%	34%/38%	48%/50%
3	POSTER → HAPPY	43%/46%	48%/50%	30%.35%	48%/49%
4	TV → SAD	39%/40%	41%/45%	28%/30%	38%/42%
5	GOLD →JOY	35%/37%	36%/39%	22%/24%	36%/38%

Κανόνες εξόρυξης
για το αρχείο D3

Κανόνες εξόρυξης
για το αρχείο D4

		TL	TLS	TLSP (NOUN /ADJ)
1	BAG→POSITIVE	82%/83%	83%/83%	66%/68%
2	TREE→NEGATIVE	81%/83%	84%/85%	61%/64%
3	CIVIL→POSITIVE	75%/77%	75%/78%	59%/63%

		TL	TLS	TLSS	TLSP(VERB)
1	THINK →POSITIVE	93%/97%	95%/98%	79%/81%	97%/99%
2	HAVE → NEGATIVE	88%/92%	89%/95%	76%.79%	92%/96%
3	CAN → NEGATIVE	85%/90%	86%/90%	71%/74%	89%/91%
4	GO → POSITIVE	85%89%	88%/91%	67%/73%	86%/90%
5	GET → NEGATIVE	86%/87%	86%/86%	66%/70%	86%/89%

		TL	TLS	TLSS
1	CENTER →POSITIVE	90%/93%	93%/94%	87%/90%
2	TIME→ POSITIVE	89%/92%	90%/92%	89%/90%
3	LINE → NEGATIVE	90%/91%	93%/96%	88%/85%
4	SALE →NEGATIVE	87%/91%	88%/90%	86%/90%
5	TICKET → NEGATIVE	81%/87%	85%/89%	86%/88%

Κανόνες εξόρυξης
για το αρχείο D5

		TL	TLS	TLSP (NOUN /ADJ/VERB)
1	CENTER →POSITIVE	90%/93%	93%/94%	73%/77%
2	CURIOUS →POSITIVE	80%/83%	82%/86%	69%/74%
3	BAKE →NEGATIVE	78%/80%	80%/81%	66%/68%

		TL	TLS	TLSS	TLSP (NOUN)
1	WATCH →POSITIVE	95%/98%	96%/98%	78%/83%	98%/99%
2	ORANGE → POSITIVE	92%/94%	92%/95%	77%/80%	94%/95%
3	SERIES→ NEGATIVE	89%/92%	93%/95%	72%/79%	89%/90%
4	BOOK→POSITIVE	85%/88%	84%/89%	65%/68%	85%/90%
5	FOOD→ NEGATIVE	81%/83%	83%/87%	62%/63%	85%/88%

Κανόνες εξόρυξης
για το αρχείο D6

		TL	TLS	TLSP (ADJ/VERB)
1	EASY→POSITIVE	83%/88%	84%/88%	70%/72%
2	SING →POSITIVE	80%/84%	83%/85%	67%/70%
3	CLEAN →NEGATIVE	76%/78%	81%/83%	87%/88%

	Υποστήριξη/εμπιστοσύνη
STATE → HAPPY	69%/82%
LITTLE→ LOVE	62%/62%
LAST → ANGER	54%/58%
SPEED→ FEAR	50%/53%
STAR →HAPPY	49%/50%

Ισχυρότεροι κανόνες εξόρυξης με αφαίρεση stopwords

	Υποστήριξη/εμπιστοσύνη
WATCH →POSITIVE	96%/98%
THINK →POSITIVE	95%/98%
LINE → NEGATIVE	93%/96%
SERIES→ NEGATIVE	93%/95%
CENTER →POSITIVE	93%/94%

ΣΥΜΠΕΡΑΣΜΑΤΑ

- Τα σύνολα δεδομένων D1, D2 και D3 παρουσιάζουν κανόνες με ποσοστά υποστήριξης 35%-45%, ενώ τα D4, D5 και D6 περίπου στο 85%.
- Στα D1-D3 τα επιθυμητα όρια υποστήριξης και εμπιστοσύνης είναι 30%/37% και στα D4-D6 στο 85% και στα δύο
- Αποδοτικότερη τεχνική προεπεξεργασίας TLS
- Το D2 λόγω του μεγάλου πλήθους tweets, περιέχει τους λιγότερα ισχυρούς κανόνες και τα λιγότερα συχνά στοιχειοσύνολα
- Με κοινή θεματολογία tweets η μέθοδος προεπεξεργασίας TLSP βελτιώνει την υποστήριξη των κανόνων
- Η TLSS βελτιώνει ή αφήνει σταθερά τα αποτελέσματα στα σύνολα δεδομένων χωρίς κοινή θεματολογία

ΜΕΛΛΟΝΤΙΚΕΣ ΕΠΕΚΤΑΣΕΙΣ

- ΥΛΟΠΟΙΗΣΗ ΜΕ ΔΙΑΦΟΡΕΤΙΚΟΥΣ ΑΛΓΟΡΙΘΜΟΥΣ
- ΕΡΕΥΝΑ ΓΙΑ ΣΥΝΑΙΣΘΗΜΑΤΙΚΗ ΠΟΛΙΚΟΤΗΤΑ ΠΕΛΑΤΩΝ-ΕΤΑΙΡΙΕΣ
- ΔΙΑΦΟΡΕΤΙΚΕΣ ΠΗΓΕΣ ΔΕΔΟΜΕΝΩΝ
- ΑΝΤΙΜΕΤΩΠΙΣΗ ΔΙΑΔΙΚΤΥΑΚΟΥ ΕΚΦΟΒΙΣΜΟΥ


THANK YOU
QUESTIONS ?!?