Πανεπιστήμιο Πειραιώς – Τμήμα Πληροφορικής

Πρόγραμμα Μεταπτυχιακών Σπουδών

«Πληροφορική»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	**Θεματική Μοντελοποίηση σε σώμα ειδησεογραφικών κειμένων**
Ονοματεπώνυμο Φοιτητή	**Κωνσταντίνος Παπακωνσταντίνου**
Πατρώνυμο	**Νικόλαος Παπακωνσταντίνου**
Αριθμός Μητρώου	**ΜΠΣΠ 15064**
Επιβλέπων	**Γεώργιος Τσιχριντζής, Καθηγητής**

Ημερομηνία Παράδοσης **Οκτώβριος 2018**

Τριμελής Εξεταστική Επιτροπή

(υπογραφή)	(υπογραφή)	(υπογραφή)
Γεώργιος Τσιχριντζής Καθηγητής	Ευάγγελος Σακόπουλος Επίκουρος Καθηγητής	Διονύσιος Σωτηρόπουλος Δρ.

Θεματική μοντελοποίηση σε σώμα ειδησεογραφικών κειμένων

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέπων καθηγητή μου κ. Γεώργιο Τσιχριντζή για την δυνατότητα εκπόνησης της διπλωματικής μου εργασίας με θέμα στο πεδίο της Ανάλυσης Δεδομένων. Ευχαριστώ θερμά τον Δρ. κ. Διονύσιο Σωτηρόπουλο για την βοήθεια, την καθοδήγηση, την κατανόηση και την επίβλεψη της διπλωματικής μου εργασίας.

Θα ήθελα να εκφράσω την ευγνωμοσύνη μου και να πω ένα τεράστιο ευχαριστώ προς τους γονείς μου, Νίκο και Μαρία και την αδερφή μου Ιωάννα για την ανεξάντλητη και ανιδιοτελή αγάπη και υποστήριξη που μου προσφέρουν όλα αυτά τα χρόνια. Τέλος, θα ήθελα να ευχαριστήσω τους φίλους μου Νίκο, Φώντα, Ζαννή, Αλέξη, Αναστασία και Μαρία, που είναι πάντα δίπλα μου, με ανέχονται, με βοηθάνε, μου συμπαραστέκονται και αποτελούν το πιο σημαντικό ψυχολογικό μου στήριγμα.

Την συγκεκριμένη εργασία την αφιερώνω στην γιαγιά μου, που δεν είναι πια ανάμεσα μας, αλλά είναι πάντα δίπλα μου.

Περιεχόμενα

Περίληψη

Τα ειδησεογραφικά νέα αποτελούν μια τεράστια δομή ιστορικών εγγράφων. Αποτελούν έναν πολύτιμο πόρο για να μελετηθεί το παρελθόν. Η επεξεργασία φυσικής γλώσσας (Native Language Processing) αποτελεί έναν κλάδο της τεχνητής νοημοσύνης (Artificial Intelligence) που βοηθά τους υπολογιστές να κατανοούν, να ερμηνεύουν και να χειρίζονται την ανθρώπινη γλώσσα. Στα καθήκοντα κατανόησης της φυσικής γλώσσας, είναι να μπορούμε να εξαγάγουμε τη σημασία και το νόημα από λέξεις, προτάσεις, παραγράφους και έγγραφα. Σε επίπεδο εγγράφου, ένας από τους πιο χρήσιμους τρόπους κατανόησης του κειμένου είναι η ανάλυση των θεμάτων του. Το Topic Modelling (Μοντελοποίηση Θεμάτων) διαδραματίζει σημαντικό ρόλο στην ανάλυση των ιστορικών εγγράφων. Το Topic Modelling παρέχει έναν τρόπο ανάλυσης μεγάλου όγκου μη ταξινομημένου κειμένου. Ένα topic-θέμα περιέχει ένα σύνολο λέξεων που εμφανίζονται συχνά μαζί. Έχουμε προς υλοποίηση ένα πρόβλημα μάθησης χωρίς επίβλεψη. Στη μάθηση χωρίς επίβλεψη, το σύστημα πρέπει μόνο του να ανακαλύψει συσχετίσεις ή ομάδες σε ένα σύνολο δεδομένων, δημιουργώντας πρότυπα, χωρίς να είναι γνωστό αν υπάρχουν, πόσα και ποια είναι. Τα ιστορικά αρχεία πολλές φορές είναι περίπλοκα, είναι δύσκολα στην κατηγοριοποίηση και μπορεί να μην έχουν τυπική ορθογραφία και μορφοποίηση. Στην συγκεκριμένη εργασία από ένα σώμα τίτλων ειδησεογραφικών ειδήσεων από το 2003 έως το 2017, του ραδιοτηλεοπτικού φορέα ABC News, προσπαθούμε να υλοποιήσουμε πρότυπα επεξεργασίας, ώστε να καταφέρουμε να αντλούμε όλη την κατάλληλη πληροφορία. Στην περίπτωση μας έχουμε μόνο μη ετικετοποιημένα δεδομένα εισόδου και πρέπει να καθορίσουμε ενδογενώς τις κατηγορίες των θεμάτων. Συγκρίνουμε βασικούς αλγόριθμους υλοποίησης Topic Modelling, εντοπίζουμε γιατί δεν ισχύει η συνοχή θεμάτων στην περίπτωση μας και υλοποιούμε Μοντελοποίηση Θεμάτων. Αναλύουμε τα θέματα, βρίσκουμε την ανάθεση θεμάτων ανά έγγραφο και την εξέλιξη της ανάθεσης αυτής στο χρόνο. Στην συνέχεια παρουσιάζουμε την εξέλιξη των θεμάτων συναρτήσει του χρόνου και τέλος αφού προσθέτουμε νέες εγγραφές νέων ειδησεογραφικών τίτλων ειδήσεων, ο αλγόριθμος μας τις ταξινομεί στο πιο κατάλληλο θέμα.

Ετικέτες: Topic Modelling, LDA, Topic Coherence, Topic Over Time

Abstract

News articles is a huge structure of historical documents. They are a valuable resource to study the past. Native Language Processing is a field of Artificial Intelligence that helps computers understand, interpret and manipulate human language. The task of understanding natural language is to be able to extract meaning from words, sentences, paragraphs, and documents. At the document level, one of the most useful ways to understand the text is to analyze its subjects - topics. Topic Modeling is an important tool in analyzing historical documents. Topic Modeling provides a way to analyze a large volume of unclassified text. A topic contains a set of words that often appear together. We have an unsupervised learning problem to be implemented. In unsupervised learning, the system needs only to discover associations or groups in a set of data, creating patterns, without knowing anything about this. Historical documents are often complicated, difficult to categorize and may not have standard spelling and formatting. In this work from a corpus of news headlines from 2003 to 2017, of ABC News, we are trying to implement standard pattern works that we can get all the deep learning information. In our case, we only have unlabelled input data and we need to define endogenously the categories of topics. The modeling of topics is quite similar to a Clustering problem.

In this work, we compare the algorithms that implement Topic Modeling, examine why Topic coherency do not work in our case and implement Topic Modeling in a corpus of documents from news headlines. We analyze the Topics, and we find the dominant Topic per document and its evolution over time. Then we present the evolution of Topics over time and finally we add new headlines and how our algorithm classifies them in the most appropriate topic.

Tags: Topic Modelling, LDA, Topic Coherence, Topic Over Time

1. Εισαγωγή

1.1 ΔΙΑΤΥΠΩΣΗ ΤΟΥ ΠΡΟΒΛΗΜΑΤΟΣ

Τα ειδησεογραφικά νέα-άρθρα είναι μια συνεχόμενη ροή ιστοριών που ζουν και παρουσιάζουν τις στιγμές, τα γεγονότα και τα συναισθήματα που προκαλούν τα γεγονότα, στο παρόν τους. Οι ειδήσεις αποτελούν έναν πολύτιμο πόρο για να μελετηθεί το παρελθόν. Μπορείς να μελετήσεις πως αντιμετωπίστηκε από κάποιο ειδησεογραφικό πρακτορείο-εφημερίδα ή από κάποιον παγκόσμιο ή τοπικό δημοσιογραφικό τύπο η ανθρώπινη ιστορία. Δηλαδή πως παρουσιάστηκαν κάποια παγκόσμια γεγονότα, όπως πόλεμοι, ανακαλύψεις – εφευρέσεις, πανδημίες, αρρώστιες και οι θεραπείες τους, κάποια κοινωνικά θέματα, οικονομικές αλλαγές κτλ. Επίσης ποια ειδησεογραφικά πρακτορεία καλύπτουν περισσότερο τον επιστημονικό κλάδο, ποια τον πολιτικό-οικονομικό κλάδο, ποια τον αθλητικό κλάδο κτλ.

Οπότε στην πραγματικότητα τα ειδησεογραφικά νέα αποτελούν μια τεράστια δομή ιστορικών εγγράφων. Το Topic Modelling (Μοντελοποίηση Θεμάτων) διαδραματίζει σημαντικό ρόλο στην ανάλυση των ιστορικών εγγράφων[14]. Τα ιστορικά αρχεία πολλές φορές είναι περίπλοκα, είναι δύσκολα στην κατηγοριοποίηση και μπορεί να μην έχουν τυπική ορθογραφία και μορφοποίηση.

Το καθήκον ενός ιστορικού δεν είναι μόνο να απορροφήσει το περιεχόμενο των ιστορικών αρχείων, αλλά να γενικεύσει τις πληροφορίες και να βρει μοτίβα στα έγγραφα αυτά. Η Μοντελοποίηση Θεμάτων αντιμετωπίζει, σε μεγάλο βαθμό, αυτά τα ζητήματα. Αυτό την καθιστά πολύ χρήσιμη. Χαρακτηριστικά της Μοντελοποίησης Θεμάτων είναι ότι προσφέρει κλιμάκωση, είναι εύρωστη στη μεταβλητότητα και είναι σε θέση να γενικεύει ενώ παραμένει προσηλωμένη στην παρατήρηση. Μελετώντας την ιστορία συνήθως αντιμετωπίζουμε απροσδόκητα προβλήματα, σε γεγονότα που μας φαίνονται οικεία. Δεν γνωρίζουμε τον τρόπο με τον οποίο οι άνθρωποι στο παρελθόν μιλούσαν για συγκεκριμένα θέματα, τις ακριβείς αντιλήψεις τους και πώς οργάνωναν τη ζωή τους. Τις περισσότερες φορές υποθέτουμε ότι γνωρίζουμε αυτά τα πράγματα και θεωρούμε ότι οι πρόγονοί μας έβλεπαν τον κόσμο με τον ίδιο τρόπο που τον βλέπουμε εμείς. Η Μοντελοποίηση Θεμάτων δίνει μια άποψη, η οποία βασίζεται στα πρότυπα των εγγράφων κι όχι στις δικές μας αντιλήψεις για το πώς πρέπει να είναι τα πράγματα.

Ο χρόνος αποτελεί ίσως την πιο κρίσιμη μεταβλητή στη μελέτη των ιστορικών εγγράφων. Αν και πολλές σύγχρονες συλλογές έχουν μια σημαντική πτυχή της χρονικής διακύμανσης, ο χρόνος είναι ένα καθοριστικό στοιχείο της ιστορικής έρευνας. Συλλογές ιστορικών εγγράφων βρίσκονται αναγκαστικά σε χρόνο διαφορετικό από τον δικό μας, αλλά τείνουν να καλύπτουν μεγάλες περιόδους δεκαετίες ή και αιώνες. Ως αποτέλεσμα αυτού, σε τεράστιας χρονικής διάρκειας δεδομένα, πρέπει τα έγγραφα να οργανωθούν κατά μήκος ενός χρονικού άξονα.

Στο Topic Modelling δεν γνωρίζουμε εκ των προτέρων ποια είναι τα θέματα. Οπότε έχουμε προς υλοποίηση ένα πρόβλημα μάθησης χωρίς επίβλεψη. Στη μάθηση χωρίς επίβλεψη, το σύστημα πρέπει μόνο του να ανακαλύψει συσχετίσεις ή ομάδες σε ένα σύνολο δεδομένων, δημιουργώντας πρότυπα, χωρίς να είναι γνωστό αν υπάρχουν, πόσα και ποια είναι [9].

Στην περίπτωση μας έχουμε μόνο μη ετικετοποιημένα δεδομένα εισόδου και πρέπει να καθορίσουμε ενδογενώς τις κατηγορίες των θεμάτων. Η μοντελοποίηση των θεμάτων είναι αρκετά παρόμοια με ένα πρόβλημα Clustering – Ομαδοποίησης. Η βασική διαφορά τους είναι ότι στην μοντελοποίηση θεμάτων υλοποιείτε Clustering - Ομαδοποίηση σε κάποιο αφηρημένο χώρο λέξεων αντί για έναν πιο συμβατικό ευκλείδειο διανυσματικό χώρο.

Η σχετική ανάλυση ασχολείται ιδιαίτερα με το πώς η γλώσσα, όπως αντανακλάται στις συγκεντρώσεις θεμάτων και τα περιεχόμενα των θεμάτων, αλλάζει με την πάροδο του χρόνου. Αυτή η εργασία είναι οργανωμένη ώστε να επεξεργάζεται ιστορικά έγγραφα. Μια επαναλαμβανόμενη εστίαση είναι η επιθυμία να σχεδιαστούν τα γεγονότα και οι συζητήσεις ενάντια στον χρόνο. Χρησιμοποιούμε ως δεδομένα τους τίτλους από έναν ειδησεογραφικό οργανισμό, τα οποία είναι χρονικά κοντά μας, οπότε έχουμε μια πιο οικεία περίπτωση χρήσης στη μοντελοποίηση θέματος.

1.2 ΥΠΑΡΧΟΥΣΕΣ ΠΡΟΣΕΓΓΙΣΕΙΣ

Μέχρι σήμερα έχουν υπάρξει πολλές έρευνες και πολλά επιστημονικά άρθρα για το Topic Modelling. Ανάλογα με το είδος των δεδομένων, προσεγγίζουν διαφορετικά το θέμα. Σε αυτό το κομμάτι θα σας παρουσιάσουμε την εξέλιξη των θεματικών μοντέλων και τις μεθόδους που χρησιμοποιήθηκαν για την εξαγωγή των θεμάτων. Πολλοί ερευνητές πρότειναν πρόσφατα νέα μοντέλα θεμάτων, τα οποία θα μπορούσαν να μάθουν περισσότερα από ότι μία διάσταση θεμάτων.

Οι D.M. Blei, A.Y. Ng, M. I. Jordan, με το επιστημονικό τους άρθρο “Latent dirichlet allocation” [20], πρότειναν το Topic Modelling ως μία από τις πιο δημοφιλείς τεχνικές πιθανολογικής τεχνικής μοντελοποίησης κειμένων. Η συγκεκριμένη τεχνική θα μπορούσε αυτόματα να ταξινομήσει τα έγγραφα σε μια συλλογή από μια σειρά θεμάτων και να αντιπροσωπεύει κάθε έγγραφο με πολλαπλά θέματα και την αντίστοιχη κατανομή τους.

Οι H. D. Kim, D. H. Park, Y. Lu, C. Zhai, με το επιστημονικό τους άρθρο “Enriching text representation with frequent pattern mining for probabilistic topic modelling” [21], πρότειναν τα συχνότερα πρότυπα που δημιουργούνται από τα αρχικά έγγραφα, στη συνέχεια να εισάγονται στα πρωτότυπα έγγραφα ως μέρος της εισόδου σε ένα μοντέλο Topic Modelling όπως το LDA. Οι προκύπτουσες αναπαραστάσεις θέματος περιέχουν τόσο μεμονωμένες λέξεις όσο και προπαρασκευασμένα πρότυπα.

Οι X. Wang, A. McCallum, X. Wei, με την έρευνα τους “Topical n-grams: Phrase and topic discovery, with an application to information retrieval,” [22], πρότειναν το μοντέλο n-Gram που ανακαλύπτει θέματα και τοπικές συναφείς φράσεις αυτόματα και ταυτόχρονα. Έχει ενσωματωθεί άψογα στην γλωσσική μοντελοποίηση που βασίζεται στο IR, ενώ σε σύγκριση με την αναπαράσταση των λέξεων, οι φράσεις είναι πιο διακριτικές και φέρουν πιο συγκεκριμένη σημασιολογία. Δεδομένου ότι οι φράσεις είναι λιγότερο διφορούμενες από τις λέξεις, έχουν διερευνηθεί ευρέως ως αναπαράσταση κειμένου για την ανάκτηση κειμένου, αλλά λίγες μελέτες στον τομέα αυτό έχουν δείξει σημαντικές βελτιώσεις στην αποτελεσματικότητα.

Η Μοντελοποίηση Θεμάτων έχει ξεκινήσει με το “ Topic detection and Tracking (TDT) project [23, 24] και έχει μελετηθεί εκτενώς. Το TDT χρησιμοποιείται για την εύρεση και την παρακολούθηση του θέματος από μια ακολουθία ειδήσεων και η ακολουθούμενη τεχνική βασίζεται στο Clustering – Συσταδοποίηση. Αργότερα εμφανίστηκε η Probabilistic Latent Semantic Analysis (PLSA) [25], το Latent Dirichlet Allocation (LDA) και τα παράγωγά τους.

Ο Blei [26], πρότεινε το Dynamic Topic Model (DTM), το οποίο δίνει τη δυνατότητα να μοντελοποιηθεί ένα θέμα πάνω από την εξέλιξη του χρόνου. Το DTM αποτελεί μια βελτιωτική έκδοση του LDA. Το πλεονέκτημα του DTM σε σύγκριση με οποιοδήποτε άλλο πιθανολογικό αλγόριθμο μοντελοποίησης θέματος είναι ότι παρακολουθεί το θέμα σε μια χρονική περίοδο. Το πρόβλημα με το DTM είναι ότι έχει σταθερό αριθμό θεμάτων και είναι ξεχωριστή η έννοια του χρόνου.

Ο Blei [27], αργότερα, το 2003, εισήγαγε το hierarchical LDA – ιεραρχικό LDA, που αποτελεί την επέκταση του LDA. Το μοντέλο LDA είναι μια επίπεδη δομή θεμάτων, ενώ το Hierarchical LDA είναι ένα μοντέλο δέντρων των θεμάτων. Η ιεραρχική LDA χρησιμοποιεί μια non-parametric (μη παραμετρική) Bayesian προσέγγιση στις ιεραρχίες του μοντέλου. Το δέντρο του θέματος κατασκευάζεται ιεραρχικά από κόμβους με έναν αλγόριθμο. Στο μοντέλο δέντρου θέματος κάθε κόμβος αντιπροσωπεύεται από τυχαίο αριθμό και του έχει αντιστοιχιστεί η αντίστοιχη κατανομή λέξεων-θέματος. Το δέντρο μπορεί να διασχίζεται από τη ρίζα μέχρι τα φύλλα του[28].

2. Θεματική Μοντελοποίηση

2.1 NLP – ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ

Η επεξεργασία φυσικής γλώσσας (Native Language Processing) αποτελεί έναν κλάδο της τεχνητής νοημοσύνης (Artificial Intelligence) που βοηθά τους υπολογιστές να κατανοούν, να ερμηνεύουν και να χειρίζονται την ανθρώπινη γλώσσα. Η επεξεργασία της φυσικής γλώσσας χρησιμοποιείται από πολλούς επιστημονικούς κλάδους, συμπεριλαμβανομένης της επιστήμης της πληροφορικής και της υπολογιστικής γλωσσολογίας, προσπαθώντας να καλύψει το κενό που υπάρχει μεταξύ της ανθρώπινης επικοινωνίας και της κατανόησης των υπολογιστών.

Ενώ η επεξεργασία της φυσικής γλώσσας δεν αποτελεί μια νέα επιστήμη, η τεχνολογία προχωράει γρήγορα χάρη στο αυξημένο ενδιαφέρον για τις επικοινωνίες από άνθρωπο σε μηχανή, καθώς και τη διαθεσιμότητα μεγάλων δεδομένων, ισχυρών υπολογιστών και ενισχυμένων αλγορίθμων.

Η επεξεργασία φυσικής γλώσσας βοηθά τους υπολογιστές να επικοινωνούν με τους ανθρώπους στη γλώσσα τους και να κλιμακώνουν άλλες εργασίες που σχετίζονται με τη γλώσσα. Για παράδειγμα, μας παρέχει τη δυνατότητα να μπορούν οι υπολογιστές να διαβάζουν κείμενο, να ακούν και να κατανοούν τον λόγο μας, να ερμηνεύουν και να μετρούν το συναίσθημα μας, όπως και να καθορίζουν ποια μέρη του λόγου ή του κειμένου είναι τα σημαντικά.

Οι σημερινές μηχανές μπορούν να αναλύσουν περισσότερα δεδομένα βασισμένα στη γλώσσα από ό,τι οι άνθρωποι, με συνεπή και αμερόληπτο τρόπο. Λαμβάνοντας υπόψη την εκπληκτική ποσότητα των αδόμητων δεδομένων που παράγονται καθημερινά, από ιατρικά αρχεία μέχρι δημοσιεύσεις σε μέσα κοινωνικής δικτύωσης, οι τεχνικές αυτοματισμού αποτελούν κρίσιμο εργαλείο για την αποτελεσματική ανάλυση δεδομένων κειμένου και λόγου.

Η ανθρώπινη γλώσσα είναι εκπληκτικά πολύπλοκη και ποικίλει. Υπάρχουν άπειροι τρόποι, που μπορούμε να εκφραστούμε τόσο σε προφορικό όσο και σε γραπτό επίπεδο. Υπάρχουν εκατοντάδες γλώσσες, που κάθε μία διαθέτει και μια πληθώρα από διαλέκτους. Επιπλέον σε κάθε γλώσσα υπάρχει ένα μοναδικό σύνολο κανόνων γραμματικής, συντακτικού και ορολογίας. Στον γραπτό λόγο, συχνά κάνουμε λάθη, συντομεύουμε λέξεις και παραλείπουμε τα σημεία στίξης. Στον προφορικό λόγο, πολλοί έχουν τοπικές προφορές και χρησιμοποιούν τοπικές ή ξενόφερτες λέξεις. Επίσης στην ομιλία πρέπει να αντιμετωπιστούν προβλήματα προφοράς, βραδυγλωσσίας, παράλληλης ομιλίας ή ήχων-θορύβου. Οπότε για τη μοντελοποίηση της ανθρώπινης γλώσσας, υπάρχει ανάγκη για συντακτική και σημασιολογική κατανόηση και άριστη γνώση του συγκεκριμένου πεδίου.

Το NLP βοηθά στην επίλυση της ασάφειας στη γλώσσα και προσθέτει χρήσιμη αριθμητική δομή στα δεδομένα. Χρησιμοποιείται κυρίως για αναγνώριση ομιλίας και ανάλυση κειμένου.

Η επεξεργασία φυσικής γλώσσας περιλαμβάνει πολλές διαφορετικές τεχνικές για την ερμηνεία της ανθρώπινης γλώσσας, από μεθόδους στατιστικής και μηχανικής μάθησης έως προσεγγίσεις βασισμένες σε κανόνες και αλγορίθμους. Χρειάζεται μια ευρεία σειρά προσεγγίσεων, διότι τα δεδομένα που είναι φωνητικά ή είναι κείμενο ποικίλλουν, όπως και οι πρακτικές εφαρμογές τους.

Οι βασικές εργασίες NLP περιλαμβάνουν τις τεχνικές tokenization (χωρισμός σε λεκτικά - tokens), parsing, lemmatization / stemming και tagging σε μέρη του λόγου, την αναγνώριση γλώσσας και τον προσδιορισμό σημασιολογικών σχέσεων. Σε γενικές γραμμές, οι τεχνικές του NLP κομματιάζουν τη γλώσσα σε μικρότερα στοιχειώδη κομμάτια, προσπαθώντας να

κατανοήσουν τις σχέσεις μεταξύ των κομματιών και να διερευνήσουν πώς τα κομμάτια αυτά, συνεργάζονται ώστε να δημιουργούν νόημα.

Αυτά τα βασικά καθήκοντα χρησιμοποιούνται συχνά σε για επεξεργασία NLP υψηλότερου επιπέδου, όπως:

- Κατηγοριοποίηση περιεχομένου: Μια συλλογή δεδομένων βασισμένη σε γλωσσολογικές τεχνικές, όπως η αναζήτηση, η ευρετηρίαση και η ανίχνευση διπλοεγγραφών.
- Ανακάλυψη και μοντελοποίηση θεμάτων: Εύρεση με ακρίβεια του νοήματος και του θέματος σε συλλογές κειμένων και εφαρμογή προηγμένων αναλύσεων, όπως τεχνικές βελτιστοποίησης και πρόβλεψης.
- Συναφής εξαγωγή: Εξαγωγή δομημένων πληροφοριών από πηγές που βασίζονται σε κείμενο.
- Ανάλυση συναισθημάτων: Προσδιορισμός της διάθεσης και των υποκειμενικών απόψεων από μεγάλες ποσότητες κειμένου, συμπεριλαμβανομένου του μέσου συναισθήματος και της εξόρυξης γνώσης.
- Μετατροπή ομιλίας σε κείμενο και μετατροπή κειμένου σε ομιλία: Μετατροπή φωνητικών εντολών σε γραπτό κείμενο και αντίστροφα.
- Συνοπτική παρουσίαση εγγράφου: Δημιουργία αυτόματων συνόψεων μεγάλων κειμένων κειμένου.
- Μηχανική μετάφραση. Αυτόματη μετάφραση κειμένου ή ομιλίας από μια γλώσσα σε μία άλλη.

Σε όλες αυτές τις περιπτώσεις, ο πρωταρχικός στόχος είναι να ληφθούν ακατέργαστα δεδομένα μέσω της εισαγωγής κειμένου ή ήχου και να χρησιμοποιηθούν αλγόριθμοι ή τεχνικές γλωσσολογίας ώστε να μετατραπεί ή να εμπλουτιστεί ένα κείμενο με τέτοιο τρόπο ώστε να αποδίδει μεγαλύτερη αξία. Η επεξεργασία της φυσικής γλώσσας συμβαδίζει με την ανάλυση των κειμένων, η οποία μετράει, ομαδοποιεί και κατηγοριοποιεί τις λέξεις για να εξάγει τη δομή και τη σημασία τους από μεγάλες ποσότητες περιεχομένου.

Υπάρχουν πολλές κοινές και πρακτικές εφαρμογές του NLP στην καθημερινότητά μας. Μία εφαρμογή του είναι στο ηλεκτρονικό ταχυδρομείο μας, στο email, στο φάκελο ανεπιθύμητης αλληλογραφίας, όπου ίσως έχετε παρατηρήσει ομοιότητες στο τίτλο του θέματος του email. Το φιλτράρισμα των ανεπιθύμητων μηνυμάτων γίνεται με την Bayesian τεχνική, μια στατιστική τεχνική NLP που συγκρίνει τις λέξεις σε spam με έγκυρα μηνύματα ηλεκτρονικού ταχυδρομείου για τον εντοπισμό της ανεπιθύμητης αλληλογραφίας.

Ένα υποπεδίο του NLP είναι το NLU, που ονομάζεται φυσική γλώσσα κατανόησης, έχει αρχίσει να αυξάνεται σε δημοτικότητα λόγω των δυνατοτήτων του σε γνωστικές εφαρμογές και εφαρμογές τεχνητής νοημοσύνης. Το NLU υπερβαίνει τη δομική κατανόηση της γλώσσας, δημιουργεί λεξιλόγιο, και δημιουργεί μια καλά διαμορφωμένη ανθρώπινη γλώσσα από μόνο του. Οι αλγόριθμοι του NLU πρέπει να αντιμετωπίσουν το εξαιρετικά σύνθετο πρόβλημα της σημασιολογικής ερμηνείας, δηλαδή, να κατανοήσουν την προτεινόμενη έννοια του λεκτικού ή γραπτού λόγου, με όλα τα λεπτά στοιχεία και τα συμπεράσματα που μπορούμε να καταλάβουμε εμείς.

Η εξέλιξη του NLP προς την NLU έχει πολλές σημαντικές επιπτώσεις τόσο για τις επιχειρήσεις όσο και για τους καταναλωτές. Φανταστείτε τη δύναμη ενός αλγορίθμου που μπορεί να κατανοήσει τη σημασία και την απόχρωση της ανθρώπινης γλώσσας σε πολλά πλαίσια, από την ιατρική μέχρι την διδασκαλία.

2.2 ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

Η προ-επεξεργασία του κειμένου και των δεδομένων αποτελεί ουσιαστικό μέρος κάθε συστήματος NLP, καθώς οι χαρακτήρες, οι λέξεις και οι φράσεις που προσδιορίζονται σε αυτό το στάδιο είναι οι θεμελιώδεις μονάδες που μεταφέρονται σε όλα τα υπόλοιπα στάδια επεξεργασίας που ακολουθούν.

Δυστυχώς, οι λέξεις που εμφανίζονται σε έγγραφα και κείμενα έχουν πολλές δομικές παραλλαγές. Έτσι, πριν από την ανάκτηση πληροφοριών από τα κείμενα, οι τεχνικές προ-επεξεργασίας δεδομένων εφαρμόζονται στο στοχευόμενο σύνολο δεδομένων για να μειωθεί το μέγεθος του συνόλου δεδομένων κι έτσι να αυξηθεί η αποτελεσματικότητα του συστήματος.

Η προ-επεξεργασία περιλαμβάνει ένα σύνολο ενεργειών, οι οποίες προετοιμάζουν το κείμενο. Επειδή τα κείμενα περιέχουν συχνά μερικές συγκεκριμένες ειδικές μορφές, όπως αριθμούς, ημερομηνίες και τις συχνά χρησιμοποιούμενες λέξεις (όπως προθέσεις, άρθρα κτλ), που δεν βοηθούν στην εξόρυξη γνώσης από το κείμενο, πρέπει να εξαλειφθούν από αυτό.

Το πρώτο βήμα στην ανάλυση και εξόρυξη γνώσης από κείμενα, είναι να οριστεί σωστά το σώμα κειμένων που θα αναλυθεί. Εάν η φιλοδοξία είναι τα αποτελέσματα να γενικευθούν σε ένα μεγαλύτερο πληθυσμό εγγράφων, τότε εφαρμόζονται οι τυποποιημένοι κανόνες δειγματοληψίας. Δηλαδή τα έγγραφα θα πρέπει να επιλέγονται χρησιμοποιώντας κάποια τυχαία ή κάποια άλλη στρατηγική δειγματοληψίας και να είναι αντιπροσωπευτικά πάνω στο πεδίο που ασχολούμαστε. Μια πρόκληση που αντιμετωπίζεται μερικές φορές σε αυτό το στάδιο είναι η επανάληψη εγγραφών (διπλοεγγραφές), δηλαδή ότι μπορεί να υπάρχουν περισσότερες από μία περιπτώσεις του ίδιου εγγράφου σε ένα σώμα. Για παράδειγμα, στις βάσεις δεδομένων μεγάλων παγκόσμιων εφημερίδων, το ίδιο άρθρο μπορεί να υπάρχει περισσότερες από μία φορές, ίσως επειδή υπάρχουν διαφορετικές εκδόσεις μιας εφημερίδας σε διάφορες χώρες. Αυτή η αλληλοεπικάλυψη μπορεί να προκαλέσει στρέβλωση του συμπεράσματος με την υπερεκπροσώπηση ορισμένων εγγράφων. Η εύρεση και διαγραφή των διπλοεγγραφών μπορεί να πραγματοποιηθεί με διάφορους αυτοματοποιημένους (χρησιμοποιώντας κάποιον αλγόριθμο) ή μη τρόπους, ελέγχοντας το σώμα του κειμένου, ώστε να διασφαλιστεί ότι κάθε έγγραφο αποτελεί μια μοναδική εγγραφή.

Μόλις καθοριστεί το σώμα, το επόμενο βήμα είναι να μετασχηματιστεί σε μορφή που να μπορεί να υποβληθεί σε ανάλυση. Η διαδικασία αυτή είναι συχνά δυσκίνητη και χρονοβόρα. Για παράδειγμα, τα έγγραφα που αποθηκεύονται σε μορφή .pdf, ίσως χρειαστεί να μετατραπούν και να αποθηκευτούν ως αρχεία κειμένου (.txt). Επίσης, αν τα κείμενα υπάρχουν μόνο σε χαρτί, όπως πολλά αρχειακά αντικείμενα, τότε πρέπει να σαρωθούν και να μετατραπούν σε αρχεία .txt, χρησιμοποιώντας λογισμικό οπτικής αναγνώρισης χαρακτήρων.

Στην ουσία παίρνουμε το σώμα ενός κειμένου και εκτελούμε σε αυτό κάποιους βασικούς μετασχηματισμούς και αναλύσεις, ώστε να μείνει το σώμα του κειμένου με αντικείμενα που θα είναι πολύ πιο χρήσιμα για την εκτέλεση κάποιου περαιτέρω, πιο ουσιαστικού αναλυτικού έργου.

Υπάρχουν 3 κύριες συνιστώσες της προ-επεξεργασίας κειμένων:

- Tokenization – αναγνώριση λέξεων
- normalization - στελέχωση
- subsitution - αντικατάσταση

2.2.1 TOKENIZITION

Μόλις τα κείμενα μετασχηματιστούν σε κατάλληλη μορφή, το επόμενο βήμα είναι να χωριστούν τα έγγραφα σε λεκτικά (tokens) και προτάσεις. Στην εξόρυξη γνώσης από κείμενο, η αναπαράσταση του κειμένου σε λεκτικά ονομάζεται tokenizition και λέμε ότι δημιουργούμε μια τσάντα από λέξεις (bag of words).

Αυτό το βήμα περιλαμβάνει την αναπαράσταση του κειμένου ως μια λίστα λέξεων, αριθμών, σημείων στίξης και ενδεχομένως άλλων συμβόλων όπως νομίσματα ή σημάτων πνευματικής ιδιοκτησίας κτλ. Το Tokenization είναι ένα βήμα που χωρίζει τις μεγαλύτερα μέρη του κειμένου σε μικρότερα κομμάτια - λεκτικά. Μεγαλύτερα κομμάτια κειμένου μπορούν να διαχωριστούν σε προτάσεις, οι προτάσεις μπορούν να διαχωριστούν σε λέξεις κ.λ.π.

Περαιτέρω επεξεργασία γίνεται γενικά αφού ένα κομμάτι κειμένου έχει κατάλληλα διακριθεί. Το tokenization αναφέρεται επίσης ως κατακερματισμός κειμένου ή λεξικολογική ανάλυση. Μερικές φορές, η κατάτμηση - segmentation χρησιμοποιείται για να αναφερθεί στην κατανομή ενός μεγάλου τεμαχίου κειμένου σε τεμάχια μεγαλύτερα από τις λέξεις (π.χ. παραγράφους ή προτάσεις), ενώ το tokenization χρησιμοποιείτε για τη διαδικασία που σαν αποτέλεσμα έχει τον κατακερματισμό του κειμένου σε λέξεις.

Ανάλογα με τον όγκο των εγγραφών τόσο πιο χρονοβόρα και δύσκολη είναι κι αυτή η ενέργεια. Στο σημείο αυτό βασική πτυχή της εξόρυξης γνώσης από κείμενο είναι να μειωθεί το διαστασιολόγιο του bag-of-words, ώστε να εξαλειφθεί ο «θόρυβος» και να εξαπλωθεί το διακριτό περιεχόμενο των εγγράφων. Υπάρχουν αρκετές διαθέσιμες τεχνικές για την αντιμετώπιση των λέξεων που είναι περιττές ως προς το περιεχόμενο του σώματος.

2.2.2 NORMALIZATION

Πριν από την περαιτέρω επεξεργασία, το κείμενο πρέπει να κανονικοποιηθεί. Η κανονικοποίηση αναφέρεται γενικά σε μια σειρά σχετικών εργασιών που αποσκοπούν στην τροποποίηση όλου του σώματος του κειμένου σε ισότιμους όρους: μετατρέποντας όλο το κείμενο σε κεφαλαία ή πεζά, αφαιρώντας τη στίξη, μετατρέποντας τους αριθμούς σε ισοδύναμα λέξεων κλπ. Η κανονικοποίηση θέτει όλες τις λέξεις επί ίσοις όροις και επιτρέπει την ομοιόμορφη ανάλυση του κειμένου.

Η κανονικοποίηση του κειμένου μπορεί να σημαίνει την εκτέλεση πολλών εργασιών, αλλά η γενική προσέγγιση αποτελείτε από τρία διαφορετικά βήματα:

- το stemming,
- το lemmatization
- άλλες ενέργειες

2.2.2.1 Stemming

Το Stemming είναι η διαδικασία αναγωγής των όρων στις ρίζες του (επιθέματα, προθέματα, καταλήξεις. προσαυξήσεις) από μια λέξη, προκειμένου να κρατηθεί μόνο ο κορμός της λέξης. Για παράδειγμα η λέξη: *running → run*

2.2.2.2 Lemmatization

Το Lemmatization είναι μια ενέργεια κάπως σχετική με το Stemming και διαφέρει από το γεγονός ότι η λημματοποίηση θα υλοποιήσει περικοπή των κλιτικών καταλήξεων και αναγωγή παράγωγων μορφών μιας λέξης σε κοινή βασική μορφή.

Για παράδειγμα, η λέξη: *better → good*

Το Lemmatization αποτελεί εργασία δυσκολότερη και πιο χρονοβόρα σε σύγκριση με το Stemming.

2.2.2.3 Άλλες Ενέργειες

Το Stemming και το Lemmatization αποτελούν βασικά μέρη της προεπεξεργασίας κειμένων. Υπάρχουν, ωστόσο, πολυάριθμα άλλα βήματα που μπορούν να υλοποιηθούν, ώστε να βοηθήσουν να τοποθετηθούν τα κείμενα σε μία ίση βάση. Αυτές οι ενέργειες υλοποιούν λεπτομερείς και ξεχωριστούς γραμματικούς και συντακτικούς κανόνες. Μερικά από αυτά είναι:

- Μετατροπή όλα των χαρακτήρων σε πεζά γράμματα
- Απομάκρυνση των αριθμών ή μετατροπή τους στην λεκτική τους αναπαράσταση
- Απομάκρυνση των σημείων στίξης

- Αφαίρεση των κενών διαστημάτων
- Αφαίρεση των λέξεων “stopwords”

Οι “stopwords” λέξεις είναι οι λέξεις εκείνες που συμβάλλουν ελάχιστα στο γενικό νόημα, δεδομένου ότι είναι γενικά οι συνηθέστερες λέξεις σε μια γλώσσα. Γενικά πρόκειται για άρθρα, προθέσεις, ή λέξεις όπως το "είναι", το "όπως" κτλ, οι οποίες δεν συμβάλλουν γενικά στην κατανόηση του περιεχομένου. Έτσι φιλτράρονται πριν από την περαιτέρω επεξεργασία του κειμένου.

2.2.3 Αφαίρεση του «θορύβου» στα δεδομένα

Η αφαίρεση του θορύβου συνεχίζει τις εργασίες τροποποίησης και καθορισμού των δεδομένων. Ενώ τα πρώτα 2 βήματα το tokenization και το normalization εφαρμόζονται ως επί τον πλείστον σε όλα σχεδόν τα κομμάτια κειμένου, η απομάκρυνση του θορύβου από τα δεδομένα αφορά ένα πολύ πιο συγκεκριμένο τμήμα.

Δεν έχουμε να κάνουμε με μια γραμμική διαδικασία, τα βήματα της οποίας πρέπει να εφαρμόζονται αποκλειστικά με συγκεκριμένη σειρά. Συνεπώς, η απομάκρυνση του θορύβου μπορεί να συμβεί πρίν ή μετά των δύο προηγούμενων βημάτων ή σε κάποιο σημείο μεταξύ τους.

Για παράδειγμα, αν έχουμε αποκτήσει την συλλογή κειμένων μας από κάποια ιστοσελίδα στον παγκόσμιο ιστό, τότε υπάρχει μεγάλη πιθανότητα τα κείμενα μας να περιέχουν ετικέτες HTML ή XML. Επίσης μπορεί στο σώμα των δεδομένων να συλλέγουμε και τα metadata της ιστοσελίδας. Αυτό εξαρτάται από τον τρόπο απόκτησης και συλλογής των δεδομένων. Δεδομένου ότι έχουμε τον έλεγχο της διαδικασίας συλλογής και επεξεργασίας των δεδομένων, αυτό καθιστά σχετικά εύκολη υπόθεση τον εντοπισμό και την αντιμετώπιση αυτού του θορύβου. Ο εντοπισμός τους γίνεται ελέγχοντας στο σώμα των κειμένων για κάποια ακριβή ταιριαστά μοτίβα.

- Αρχεία κείμενου κεφαλίδας(header) και υποσέλιδου(footer)
- Ετικέτες HTML, XML και metadata
- Εξαγωγή δεδομένων από άλλες μορφές όπως JSON ή από βάση δεδομένων

Πολλές φορές η απομάκρυνση του θορύβου πρέπει να υλοποιηθεί πριν από τα άλλα βήματα προ-επεξεργασίας. Για παράδειγμα, οποιοδήποτε κείμενο λαμβάνεται από μια δομή JSON θα πρέπει προφανώς να καθαριστεί πριν από το tokenization.

2.3 TOPIC MODELLING

Στα καθήκοντα κατανόησης της φυσικής γλώσσας (NLU), πρέπει να μπορούμε να εξαγάγουμε νόημα από λέξεις σε προτάσεις, σε παραγράφους και σε έγγραφα. Σε επίπεδο εγγράφου, ένας από τους πιο χρήσιμους τρόπους κατανόησης του κειμένου είναι η ανάλυση των θεμάτων του. Η διαδικασία εκμάθησης, αναγνώρισης και εξαγωγής αυτών των θεμάτων σε μια συλλογή εγγράφων ονομάζεται Topic Modelling - μοντελοποίηση θέματος.

Το Topic Modelling παρέχει έναν τρόπο ανάλυσης μεγάλου όγκου μη ταξινομημένου κειμένου. Ένα topic-θέμα περιέχει ένα σύνολο λέξεων που εμφανίζονται συχνά μαζί. Το Topic Modelling μπορεί να συνδέσει λέξεις με παρόμοιες έννοιες και να διακρίνει την χρήση των λέξεων με πολλαπλές έννοιες.

Όλα τα μοντέλα θεμάτων βασίζονται στην ίδια βασική υπόθεση:

- Κάθε έγγραφο αποτελείται από ένα μείγμα θεμάτων
- Κάθε θέμα αποτελείται από μια συλλογή λέξεων

Δηλαδή, το Topic Modelling βασίζεται στην ιδέα ότι η σημασιολογία ενός εγγράφου διέπεται από ορισμένες κρυφές ή λανθάνουσες «latent» μεταβλητές που δεν παρατηρούμε. Ως

αποτέλεσμα αυτού, στόχος του Topic Modelling είναι να αποκαλυφθούν αυτά τα θέματα, τα οποία διαμορφώνουν την έννοια του εγγράφου και του συλλογής εγγράφων.

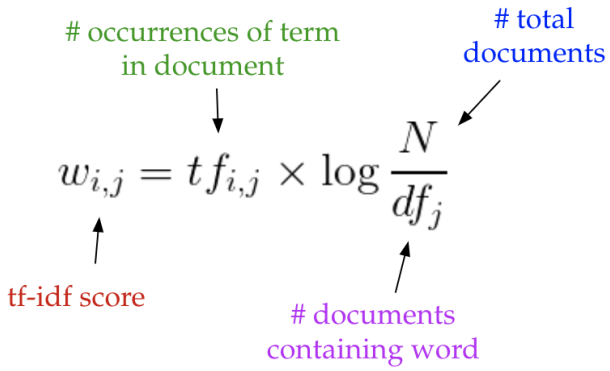
Οι βασικοί και πιο χρησιμοποιημένοι μέθοδοι για το Topic Modelling είναι τέσσερις. Αυτές οι μέθοδοι είναι η Λανθάνουσα Σημασιολογική Ανάλυση (LSA), η Πιθανότητα Λανθάνουσας Σημασιολογικής Ανάλυσης (PLSA), η Λανθασμένη Κατανομή Dirichlet (LDA) και το lda2vec βασισμένο στο deep learning. Υπάρχουν και μέθοδοι που ανήκουν στο πεδίο του Model Topic Evolution. Σε αυτή την κατηγορία υπάρχουν διάφορα μοντέλα, όπως το Topic Over Time (TOT), τα μοντέλα δυναμικών θεμάτων (DTM), η Multiscale Topic Tomography - τομογραφία θεμάτων πολλών επιπέδων, Detecting Topic Evolutionin scientific literatures - η ανίχνευση θεμάτων εξέλιξης σε επιστημονικές βιβλιογραφίες κλπ.

2.3.1 LSA - LATENT SEMANTIC ANALYSIS

Η Λανθάνουσα Σημασιολογική Ανάλυση, ή LSA, είναι μία από τις βασικές τεχνικές στη μοντελοποίηση θέματος. Η βασική ιδέα είναι να πάρουμε μια μήτρα εγγράφων και όρων και να την αποσυνθέσουμε σε ένα ξεχωριστό document-topic-matrix (έγγραφο-θέμα πίνακα) και ένα topic-term matrix(πίνακα θέμα-όρος).

Το πρώτο βήμα είναι η δημιουργία ενός πίνακα εγγράφου-όρου. Λαμβάνοντας τα m έγγραφα και τις n λέξεις στο λεξιλόγιό μας, μπορούμε να κατασκευάσουμε έναν πίνακα A διαστάσεων m × n, στον οποίο κάθε σειρά αντιπροσωπεύει ένα έγγραφο και κάθε στήλη αντιπροσωπεύει μια λέξη. Στην απλούστερη έκδοση του LSA, κάθε καταχώρηση μπορεί απλώς να είναι μια ακατέργαστη μέτρηση του αριθμού των φορών που η λέξη j εμφανίστηκε στο v-οστό έγγραφο. Στην πράξη, ωστόσο, οι πρώτες μετρήσεις δεν λειτουργούν ιδιαίτερα καλά επειδή δεν λαμβάνουν υπόψη τη σημασία κάθε λέξης στο έγγραφο. Για παράδειγμα, η λέξη "πυρηνική" πιθανότατα μας ενημερώνει περισσότερα για το θέμα (τα θέματα) ενός συγκεκριμένου εγγράφου από τη λέξη "δοκιμή - τεστ".

Συνεπώς, τα μοντέλα LSA συνήθως αντικαθιστούν τις ακατέργαστες μετρήσεις στον πίνακα του εγγράφου με ένα σκορ tf-idf. Το Tf-idf (συχνότητα όρου - αντίστροφη συχνότητα εγγράφου), εκχωρεί ένα βάρος για τον όρο j στο έγγραφο i ως εξής:



Γενικά ισχύει ότι όσο πιο συχνά ο όρος εμφανίζεται στο έγγραφο, τόσο μικρότερο είναι το βάρος του και όσο πιο σπάνια εμφανίζεται σε όλο το σώμα, τόσο μεγαλύτερο είναι το βάρος του.

Αφού αποκτήσουμε τον πίνακα αρχείων εγγράφων A, μπορούμε να αρχίσουμε να σκεφτόμαστε τα latent topics - λανθάνοντα θέματα μας. Κατά πάσα πιθανότητα, ο πίνακας εγγράφων A θα είναι πολύ αραιός, πολύ θορυβώδης και πολύ περιττός σε πολλές διαστάσεις του. Ως αποτέλεσμα, για να βρούμε τα λίγα λανθάνοντα θέματα που καταγράφουν τις σχέσεις μεταξύ των λέξεων και των εγγράφων, πρέπει να μειώσουμε την διάσταση του πίνακα A.

Αυτή η μείωση διαστάσεων μπορεί να πραγματοποιηθεί με τη χρήση της SVD. Η SVD (singular value decomposition - αποσύνθεση μοναδικής αξίας) είναι μια τεχνική γραμμικής άλγεβρας που παραγοντοποιεί οποιαδήποτε μήτρα M σε παράγωγο 3 ξεχωριστών μητρών:

- M = U * S * V, όπου S είναι μια διαγώνιος μήτρα των μοναδικών τιμών του M.

Το SVD μειώνει τη διάσταση του πίνακα επιλέγοντας μόνο τις t μεγαλύτερες μοναδικές τιμές και διατηρώντας μόνο τις πρώτες στήλες t των U και V. Στην περίπτωση αυτή, το t είναι ένα στοιχείο υπερπαράμετρος που μπορούμε να επιλέξουμε και να προσαρμόσουμε ώστε να αντικατοπτρίζει τον αριθμό των θεμάτων που θέλουμε να βρούμε.

Σε αυτήν την περίπτωση, το U ∈ ℝ ^ (m × t) αναδύεται ως η μήτρα του εγγράφου-θέματος και το
V ∈ ℝ ^ (n × t) γίνεται ο πίνακας όρων-θέματος. Και στις δύο U και V, οι στήλες αντιστοιχούν σε ένα από τα θέματα μας t. Στο U, οι σειρές αντιπροσωπεύουν τους όρους του εγγράφου που εκφράζονται με όρους των θεμάτων, ενώ στο V, οι σειρές αντιπροσωπεύουν τους διανυσματικούς όρους που εκφράζονται με όρους θεμάτων.

Με αυτούς τους φορείς εγγράφων και τα διανύσματα όρων, μπορούμε να εφαρμόσουμε εύκολα μέτρα όπως η ομοιότητα συνημιτόνων για να αξιολογήσουμε:

- την ομοιότητα διαφόρων εγγράφων
- την ομοιότητα διαφόρων λέξεων
- την ομοιότητα των όρων (ή των "ερωτημάτων") και των εγγράφων (η οποία γίνεται χρήσιμη στην ανάκτηση πληροφοριών, όταν θέλουμε να ανακτήσουμε αποσπάσματα τα οποία είναι πιο συναφή με το ερώτημα αναζήτησης).

Το LSA είναι γρήγορο και αποδοτικό στη χρήση, αλλά έχει μερικά βασικά μειονεκτήματα:

- έλλειψη ενσωματωμένων ερμηνευτικών (δεν γνωρίζουμε ποια είναι τα θέματα και τα συστατικά μπορεί να είναι αυθαίρετα θετικά ή αρνητικά)
- για να υπάρχουν ακριβή αποτελέσματα πρέπει να υπάρχει μεγάλο σύνολο εγγραφών και λεξιλογίου
- λιγότερο αποδοτική αναπαράσταση

2.3.2 PLSA - LATENT DIRICHLET ALGORITH

Το PLSA ή Probabilistic Latent Semantic Analysis, χρησιμοποιεί μια πιθανολογική μέθοδο αντί για το SVD για την αντιμετώπιση του προβλήματος. Η βασική ιδέα είναι να βρεθεί ένα πιθανολογικό μοντέλο με λανθάνοντα θέματα που μπορούν να δημιουργήσουν τα δεδομένα που υπάρχουν στη μήτρα του εγγράφου. Συγκεκριμένα, θέλουμε ένα μοντέλο P (D, W) τέτοιο ώστε για οποιοδήποτε έγγραφο d και λέξη w, ένα P(d, w) να αντιστοιχεί σε εκείνη την καταχώρηση στη μήτρα εγγράφου-όρου.

Η βασική παραδοχή του Topic Modelling είναι ότι κάθε έγγραφο αποτελείται από ένα μείγμα θεμάτων και κάθε θέμα αποτελείται από μια συλλογή λέξεων. Το pLSA προσθέτει μια πιθανή περιστροφή στις υποθέσεις αυτές:

- Δοθέντος ενός εγγράφου d, ένα θέμα z υπάρχει στο έγγραφο αυτό με πιθανότητα P (z | d)
- Δοθέντος ενός θέματος z, η λέξη w προέρχεται από το θέμα z με πιθανότητα P (w | z)

Η s συλλογική πιθανότητα να δούμε ένα συγκεκριμένο έγγραφο και μια λέξη μαζί είναι:

$$P(D,W) = P(D)\sum_{Z} P(Z|D)P(W|Z)$$

Στην εξίσωση αυτή, η δεξιά πλευρά μας δείχνει την πιθανότητα να εμφανιστεί κάποιο έγγραφο και στη συνέχεια βασισμένο στην κατανομή των θεμάτων αυτού του εγγράφου, πόσο πιθανό είναι να βρεθεί μια συγκεκριμένη λέξη μέσα σε αυτό το έγγραφο.

Οι όροι P (D), P (Z | D) και P (W | Z) είναι οι παράμετροι του μοντέλου μας. Το P (D) μπορεί να προσδιοριστεί απευθείας από το σώμα των κειμένων. Το P (Z | D) και το P (W | Z) σχεδιάζονται ως πολυωνυμικές κατανομές και μπορούν να εκπαιδευτούν χρησιμοποιώντας τον αλγόριθμο μεγιστοποίησης αναμονής (EM – expectation maximization algorithm). Χωρίς την

πλήρη μαθηματική αντιμετώπιση του αλγορίθμου, η EM είναι μια μέθοδος εύρεσης πιθανών εκτιμήσεων των παραμέτρων, για ένα μοντέλο που εξαρτάται από τις μη παρατηρημένες, λανθάνουσες μεταβλητές (τα θέματα).

Παρόλο που φαίνεται πολύ διαφορετικό και προσεγγίζει το πρόβλημα με έναν πολύ διαφορετικό τρόπο, το pLSA προσθέτει απλώς μια πιθανολογική αντιμετώπιση των θεμάτων και των λέξεων, πάνω από το LSA. Πρόκειται για ένα πολύ πιο ευέλικτο μοντέλο, αλλά εξακολουθεί να έχει μερικά προβλήματα. Συγκεκριμένα:

- Επειδή δεν έχουμε παραμέτρους για το μοντέλο P (D), δεν γνωρίζουμε πώς θα εκχωρηθούν οι πιθανότητες σε νέα έγγραφα.
- Ο αριθμός των παραμέτρων για το pLSA αυξάνεται γραμμικά με τον αριθμό των εγγράφων που διαθέτουμε, επομένως είναι επιρρεπής σε υπερεξειδίκευση – overfitting.

Το μοντέλο pLSA χρησιμοποιείται σπάνια μόνο του. Σε γενικές γραμμές, όταν αναζητούμε έναν αλγόριθμο για Topic Modelling από τις βασικές επιδόσεις που δίνει ο LSA, στρεφόμαστε στον LDA. Ο LDA, είναι ο πιο συνηθισμένος τύπος για Topic Modelling, κι επεκτείνει το PLSA ώστε να αντιμετωπίσει τα ζητήματα που έχει.

2.3.3 LDA - LATENT DIRICHLET ALGORITHM

Το LDA σημαίνει Latent Dirichlet Allocation. Το LDA είναι μια Bayesian έκδοση του pLSA. Συγκεκριμένα, χρησιμοποιεί προηγμένες τεχνικές Dirichlet για το θέμα του εγγράφου και τις κατανομές λέξεων-θέματος, προσδίδοντάς του την καλύτερη γενίκευση.

Ως μια επισκόπηση το dirichlet είναι "ως κατανομή πάνω από κατανομές". Στην ουσία, απαντά στην ερώτηση: "δοθέντος του τύπου κατανομής, ποιες είναι ορισμένες πιθανολογικές κατανομές που πιθανόν να βρεθούν;"

Αν υποθέσουμε ότι το σώμα κειμένων έχει έγγραφα από 3 πολύ διαφορετικές θεματικές περιοχές. Αν θέλουμε να διαμορφώσουμε αυτό το μοντέλο, ο τύπος της κατανομής θα είναι αυτός που θα δίνει μεγάλο βάρος σε ένα συγκεκριμένο θέμα και δεν θα δίνει πολύ βάρος σε όλα τα υπόλοιπα.

Αυτό που παρέχει μια διανομή dirichlet είναι ένας τρόπος δειγματοληψίας πιθανολογικών κατανομών ενός συγκεκριμένου τύπου.

Στο pLSA, δειγματίζεται ένα έγγραφο, στη συνέχεια ένα θέμα ταυτίζεται σε αυτό το έγγραφο και στη συνέχεια μια λέξη ταυτίζεται σε αυτό το θέμα.

Σε μια κατανομή Dirichlet Dir (α), σχεδιάζουμε ένα τυχαίο δείγμα που αντιπροσωπεύει την κατανομή θεμάτων ή το μείγμα θεμάτων ενός συγκεκριμένου εγγράφου. Αυτή η κατανομή θεμάτων είναι θ. Από το θ, επιλέγουμε ένα συγκεκριμένο θέμα Z βάσει της κατανομής.

Στη συνέχεια, από μια άλλη κατανομή dirichlet Dir (β), επιλέγουμε ένα τυχαίο δείγμα που αντιπροσωπεύει την κατανομή λέξεων του θέματος Z. Αυτή η κατανομή λέξεων είναι φ. Από το φ, επιλέγουμε τη λέξη w.

Τυπικά, η διαδικασία για τη δημιουργία κάθε λέξης από ένα έγγραφο έχει ως εξής:

1. Choose $\theta_i \sim \mathrm{Dir}(\alpha)$ (where $i = 1, \ldots, M$; $\theta_i \in \Delta_K$)
   - $\theta_{i,k}$ = probability that document $i \in \{1, \ldots, M\}$ has topic $k \in \{1, \ldots, K\}$.
2. Choose $\phi_k \sim \mathrm{Dir}(\beta)$ (where $k = 1, \ldots, K$; $\phi_k \in \Delta_V$)
   - $\phi_{k,v}$ = probability of word $v \in \{1, \ldots, V\}$ in topic $k \in \{1, \ldots, K\}$.
3. Choose $c_{i,j} \sim \mathrm{Polynomial}(\theta_i)$ (where $c_{i,j} \in \{1, \ldots, K\}$)
4. Choose $w_{i,j} \sim \mathrm{Polynomial}(\phi_{c_{i,j}})$ (where $w_{i,j} \in \{1, \ldots, V\}$)

Το LDA συνήθως λειτουργεί καλύτερα από το pLSA επειδή μπορεί εύκολα να γενικευθεί σε νέα έγγραφα. Στο pLSA, η πιθανότητα του εγγράφου είναι ένα σταθερό σημείο στο σύνολο δεδομένων. Αν δεν έχουμε δει ένα έγγραφο, δεν έχουμε αυτό το σημείο δεδομένων. Στο LDA, το σύνολο των δεδομένων χρησιμεύει ως δεδομένο εκπαίδευσης για την κατανομή dirichlet των κατανομών εγγράφων-θέματος. Εάν δεν έχουμε δει ένα έγγραφο, μπορούμε εύκολα να το δειγματήσουμε σε μια κατανομή dirichlet και να προχωρήσουμε την υλοποίηση από εκεί.

Με το LDA, μπορούμε να εξαγάγουμε θέματα που ερμηνεύονται από τον άνθρωπο μέσα από ένα σώμα εγγράφων, όπου κάθε θέμα χαρακτηρίζεται από τις λέξεις με τις οποίες συνδέεται περισσότερο. Για παράδειγμα, το θέμα 2 θα μπορούσε να χαρακτηριστεί από όρους όπως "πετρέλαιο, αέριο, γεωτρήσεις, σωλήνες, Keystone, ενέργεια," κλπ. Επιπλέον, δοθέντος ενός νέου εγγράφου, μπορούμε να αποκτήσουμε ένα διάνυσμα που αντιπροσωπεύει το μείγμα θεμάτων, π.χ. 5% θέμα 1, 70% θέμα 2, 10% θέμα 3, κλπ. Αυτά τα διανύσματα είναι συχνά πολύ χρήσιμα για downstream εφαρμογές.

2.3.4 lda2vec

Σε επίπεδο εγγράφου, παρουσιάσαμε πως μπορεί ένα κείμενο να εκπροσωπηθεί ως ένα μείγμα θεμάτων.

Σε επίπεδο λέξης, συνήθως χρησιμοποιείται το word2vec για να λάβουμε διανυσματικές αναπαραστάσεις. Το lda2vec είναι μια επέκταση του word2vec και του LDA που μαθαίνουν από κοινού τα διανύσματα των λέξεων, των εγγράφων και των θέμάτων.

Το lda2vec βασίζεται ειδικά στο μοντέλο skip-gram του word2vec για τη δημιουργία διανυσματικών λέξεων. Ουσιαστικά είναι ένα νευρωνικό δίχτυ που μαθαίνει μια λέξη που ενσωματώνει, προσπαθώντας να χρησιμοποιήσει τη λέξη εισόδου για να προβλέψει τις περιβάλλουσες συναφείς λέξεις.

Με το lda2vec, αντί να χρησιμοποιούμε απευθείας το διάνυσμα λέξεων για να προβλέψουμε τις συναφείς λέξεις, χρησιμοποιούμε ένα συναφή διάνυσμα για να κάνουμε τις προβλέψεις. Αυτό το διάνυσμα δημιουργείται ως το άθροισμα δύο άλλων διανυσμάτων: το διάνυσμα λέξεων και το διάνυσμα εγγράφων.

Το διάνυσμα λέξης δημιουργείται από το ίδιο μοντέλο word2vec του skip-gram. Το διάνυσμα εγγράφων είναι πιο ενδιαφέρον. Είναι πραγματικά ένας σταθμισμένος συνδυασμός δύο άλλων στοιχείων:

- το διάνυσμα βάρους του εγγράφου, που αντιπροσωπεύει το "βάρος" κάθε θέματος στο έγγραφο
- το θέμα μήτρα, που αντιπροσωπεύει κάθε θέμα και την αντίστοιχη ενσωμάτωση του διανύσματος

Μαζί, το διάνυσμα εγγράφων και το διάνυσμα λέξεων δημιουργούν διανύσματα "περιβάλλοντος" για κάθε λέξη στο έγγραφο. Η δύναμη του lda2vec έγκειται στο γεγονός ότι δεν μαθαίνει μόνο τις ενσωματωμένες λέξεις (και τις ενσωματώσεις των διάνυσμα πλαισίων) για λέξεις, μαθαίνει ταυτόχρονα επίσης τις αναπαραστάσεις του θέματος και τις αναπαραστάσεις εγγράφων.

2.3.5 Topic Coherence - Συνοχή Θεμάτων

Το Topic Modeling παρέχει έναν τρόπο ανάλυσης μεγάλου όγκου μη ταξινομημένου κειμένου, αλλά δεν παρέχει καμία εγγύηση ότι τα αποτελέσματα είναι ερμηνεύσιμα. Παρέχει μεθόδους για την οργάνωση, την κατανόηση και την σύνοψη μεγάλων συλλογών κειμένων. Τα θέματα, τα οποία συνήθως αντιπροσωπεύονται ως σύνολα σημαντικών λέξεων που περιέχονται στο σώμα των κειμένων, τα μαθαίνει αυτόματα από μη επισημασμένα έγγραφα με έναν χωρίς επιτήρηση τρόπο. Αποτελεί μια ελκυστική μέθοδος για να δοθεί δομή σε μη δομημένα δεδομένα κειμένου, αλλά τα θέματα δεν εγγυούνται ότι σωστά ερμηνεύσιμα. Για αυτό τον λόγο έχουν προταθεί μέτρα συνοχής, ώστε να επιτευχθεί διάκριση μεταξύ καλών και κακών θεμάτων.

Το πλαίσιο για την συνοχή θεμάτων – Coherence Framework, ομαδοποιείται στις 4 ακόλουθες διαστάσεις:

- Κατάτμηση - Segmentation: χωρισμός των εγγράφων σε λέξεις, έτσι ώστε κάθε εγγραφή να είναι διαφορετική
- Εκτίμηση πιθανοτήτων - Probability Estimation: ποσοτική μέτρηση των λέξεων
- Μέτρο επιβεβαίωσης - Confirmation Measure: προσδιορισμός της ποιότητας σύμφωνα με ορισμένα προκαθορισμένα πρότυπα (π.χ. % συμμόρφωση)
- Συσσωμάτωση - Aggregation: ενιαίος βαθμός συνολικής ποιότητας

Από τεχνική άποψη, το πλαίσιο συνοχής – Coherence Framework, αντιπροσωπεύεται ως σύνθεση των μερών που μπορούν να συνδυαστούν. Τα τμήματα ομαδοποιούνται σε διαστάσεις που καλύπτουν το χώρο διαμόρφωσης των μέτρων συνοχής. Κάθε διάσταση χαρακτηρίζεται από ένα σύνολο ανταλλάξιμων συστατικών.

Το σύνολο λέξεων t είναι κατακερματισμένο σε ένα σύνολο από ζεύγη υποσυνόλων λέξεων S. Δεύτερον, οι πιθανότητες των λέξεων P, υπολογίζονται με βάση ένα δεδομένο σώμα αναφοράς. Τόσο το σύνολο των υποσυνόλων λέξεων S όσο και οι υπολογισμένες πιθανότητες P καταναλώνονται από το μέτρο επιβεβαίωσης για να υπολογιστούν οι συμφωνίες Φ των ζευγαριών του S. Τέλος, αυτές οι τιμές συγκεντρώνονται σε μια ενιαία τιμή συνοχής c. Υπάρχουν δύο μέτρα στο Topic Coherence:

- Εγγενές Μέτρο – Intrinsic Measure: Παριστάνεται ως UMass. Μετράει για να συγκρίνει μια λέξη μόνο με τις προηγούμενες και τις επόμενες λέξεις, οπότε χρειαζόμαστε ένα ταξινομημένο σύνολο λέξεων. Χρησιμοποιεί κατά ζεύγη την βαθμολογία της συνάρτησης, η οποία είναι η εμπειρική εξαρτώμενη από όρους λογαριθμική πιθανότητα με ομαλή αρίμθηση για να αποφευχθεί ο υπολογισμός του λογαρίθμου μηδέν.
- Εξωγενές Μέτρο – Extrinsic Measure: Παριστάνεται ως UCI. Στο μέτρο του UCI, κάθε λέξη συνδυάζεται με κάθε άλλη λέξη. Η UCI Coherence χρησιμοποιεί κατά σημεία αμοιβαίες πληροφορίες (PMI).

Τόσο το Εγγενές Μέτρο όσο και το Εξωγενές Μέτρο υπολογίζουν τη βαθμολογία συνοχής c (ως άθροισμα κατά ζεύγη των βαθμολογιών στις λέξεις w1, ..., wn, που χρησιμοποιούνται για την περιγραφή του θέματος.

2.4 ΕΞΕΛΙΞΗ ΤΩΝ TOPICS ΣΤΗΝ ΠΟΡΕΙΑ ΤΟΥ ΧΡΟΝΟΥ

Οι αλγόριθμοι Μοντελοποίησης Θέματος, όπως η λανθάνουσα σημασιολογική ανάλυση (LSA), η λανθάνουσα κατανομή Dirichlet (LDA) και οι απόγονοί τους έχουν προσφέρει έναν ισχυρό τρόπο εξερεύνησης τεράστιων συλλογών κειμένων που είναι δύσκολο να υλοποιηθούν από κάποιον άνθρωπο χωρίς κάποια βοήθεια.

Χρησιμοποιώντας αυτούς τους αλγορίθμους εκτός από την ανεύρεση θεμάτων από το σώμα των κειμένων, ο αλγόριθμος μοντελοποίησης θέματος έχει επίσης χρησιμοποιηθεί για να μοντελοποιήσει την εξέλιξη των θεμάτων με την πάροδο του χρόνου, καθώς και τις συνδέσεις / ιεραρχίες των θεμάτων, επεξεργάζοντας τα έγγραφα ως δεδομένα χρονοσειρών. Στην συγκεκριμένη εργασία, μας ενδιαφέρει ιδιαίτερα η εφαρμογή της μεθόδου μοντελοποίησης θέματος για να εξερευνήσετε τη δυναμική στην εξέλιξη των θεμάτων ειδησεογραφικών νέων μέσα στο πέρασμα 13 χρόνων. Τα περιεχόμενα ενός ειδησεογραφικού τίτλου, συγκεκριμένα οι λέξεις ή οι ορολογίες που χρησιμοποιούνται, υποδεικνύουν τα θέματα στα οποία εστιάζει αυτό το άρθρο. Η τοποθέτηση των θεμάτων των ειδησεογραφικών τίτλων σε μια χρονολογική σειρά θα μας παρουσιάσει τις διακυμάνσεις των νέων και των θεμάτων μέσα στο χρόνο. Θα μας αποκαλύψει συσχετίσεις μεταξύ τους αλλά και κάποιου είδους αλληλουχία.

Στην πραγματικότητα τα Topic Over Time - Θέματα κατά τη διάρκεια του χρόνου (TOT) είναι μια μοντελοποίηση θέματος με την βοήθεια του αλγόριθμου LDA, η οποία εξειδικεύει την συνεμφάνιση των λέξεων από κοινού με το χρόνο. Με άλλα λόγια, το Topic Over Time λαμβάνει τόσο τη δομή δεδομένων όσο και την εξέλιξη της δομής αυτής με την πάροδο του χρόνου. Αυτή η μέθοδος υποκινείται από την διαπίστωση ότι η δομή στα δεδομένα (σε αυτή την περίπτωση τα

μοντέλα συν-εμφάνισης των θεμάτων) δεν είναι στατική αλλά δυναμική. Τα μοτίβα που υπάρχουν στο αρχικό μέρος των δεδομένων ενδέχεται να μην ισχύουν αργότερα. Τα θέματα γεννιούνται, αυξάνονται και μειώνονται, διαιρούνται και συγχωνεύονται, και αλλάζουν συσχετισμούς με την πάροδο του χρόνου. Στο Topic Over Time, κάθε παραγόμενο έγγραφο έχει ένα μείγμα μοντέλων θεμάτων, τα οποία επηρεάζονται από κοινού κι από τις συνεμφανίσεις λέξεων αλλά και από τη χρονική στιγμή εμφάνισης του εγγράφου.

2.5 ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ

Sentiment Analysis – Ανάλυση Συναισθημάτων[39] είναι η αυτοματοποιημένη διαδικασία κατανόησης μιας γνώμης ή του συναισθήματος της σχετικά με ένα δεδομένο θέμα από γραπτό ή προφορικό γλώσσα. Η ανάλυση συναισθημάτων έχει γίνει ένα βασικό εργαλείο για την κατανόηση των τόσο πολλών δεδομένων που παράγονται καθημερινά.

Το Sentiment Analysis - Ανάλυση Συναισθημάτων, γνωστή και ως Opinion Mining – Εξόρυξη Γνώμης, είναι ένα πεδίο στο πλαίσιο της Επεξεργασίας Φυσικής Γλώσσας (NLP) που αναπτύσσει συστήματα που προσπαθούν να εντοπίσουν και να εξάγουν απόψεις μέσα στο κείμενο. Συνήθως, εκτός από την αναγνώριση της γνώμης, τα συστήματα αυτά εξάγουν χαρακτηριστικά της έκφρασης π.χ.

- Πολικότητα: Εάν ο ομιλητής εκφράσει θετική ή αρνητική γνώμη,
- Θέμα: το θέμα για το οποίο γίνεται λόγος,
- Κάτοχος γνώμης: το πρόσωπο ή η οντότητα που εκφράζει τη γνώμη.

Η ανάλυση συναισθημάτων είναι ένα θέμα ιδιαίτερου ενδιαφέροντος και ανάπτυξης, καθώς έχει πολλές πρακτικές εφαρμογές. Δεδομένου ότι οι πληροφορίες που διατίθενται δημόσια και ιδιωτικά μέσω του διαδικτύου αυξάνονται διαρκώς, μεγάλος αριθμός κειμένων που εκφράζουν απόψεις διατίθενται στους ιστοσελίδες κριτικών, στα φόρουμ, στα blog και στα κοινωνικά μέσα.

Με τη βοήθεια συστημάτων ανάλυσης αισθήσεων, αυτές οι αδόμητες πληροφορίες θα μπορούσαν να μετατραπούν αυτομάτως σε δομημένα δεδομένα, που θα παρουσίαζαν τις απόψεις του κοινού για προϊόντα, υπηρεσίες, εμπορικά σήματα, πολιτική ή οποιοδήποτε θέμα που οι άνθρωποι μπορούν να εκφράσουν τις απόψεις τους. Αυτά τα δεδομένα μπορούν να είναι πολύ χρήσιμα για εμπορικές εφαρμογές, όπως η ανάλυση μάρκετινγκ, οι δημόσιες σχέσεις, οι αναθεωρήσεις προϊόντων, η βαθμολόγηση του καθαρού υποκινητή, η ανατροφοδότηση προϊόντων και η εξυπηρέτηση των πελατών.

Οι πληροφορίες κειμένου μπορούν ευρέως να ταξινομηθούν σε δύο βασικούς τύπους: γεγονότα και απόψεις. Τα γεγονότα είναι αντικειμενικές εκφράσεις για κάτι. Οι απόψεις είναι συνήθως υποκειμενικές εκφράσεις που περιγράφουν τα συναισθήματα, τις εκτιμήσεις και τα συναισθήματα των ανθρώπων προς ένα θέμα.

Η ανάλυση συναισθημάτων, όπως και πολλά άλλα προβλήματα NLP, μπορεί να διαμορφωθεί ως πρόβλημα ταξινόμησης όπου πρέπει να επιλυθούν δύο υπο-προβλήματα:

- Ταξινόμηση μιας φράσης ως υποκειμενική ή αντικειμενική, γνωστή ως ταξινόμηση υποκειμενικότητας.
- Η ταξινόμηση μιας φράσης ως μια θετική, αρνητική ή ουδέτερη άποψη, γνωστή ως ταξινόμηση πολικότητας.

Υπάρχουν δύο είδη απόψεων: οι άμεσες και οι συγκριτικές. Οι άμεσες απόψεις δίνουν άμεσα μια γνώμη σχετικά με την άποψη ή το συναίσθημα, ενώ στις συγκριτικές γνώμες, η γνώμη εκφράζεται συγκρίνοντας μια οντότητα με μια άλλη, Συνήθως, οι συγκριτικές απόψεις εκφράζουν ομοιότητες ή διαφορές μεταξύ δύο ή περισσοτέρων οντοτήτων χρησιμοποιώντας μια συγκριτική μορφή ενός επίθετου ή επιρρήματος.

Μια σαφής άποψη για ένα θέμα είναι μια άποψη που εκφράζεται ρητά σε μια υποκειμενική φράση. Μια υπονοούμενη γνώμη για ένα θέμα είναι μια άποψη που εμπεριέχεται σε μια αντικειμενική πρόταση. Εντός των σιωπηρών απόψεων θα μπορούσαμε να συμπεριλάβουμε

μεταφορές που μπορεί να είναι οι δυσκολότεροι τύποι απόψεων που πρέπει να αναλυθούν, καθώς περιλαμβάνουν πολλές σημασιολογικές πληροφορίες.

Υπάρχουν πολλοί τύποι ανάλυσης συναισθημάτων και τα εργαλεία για αυτή την εργασία κυμαίνονται από συστήματα που επικεντρώνονται στην πολικότητα (θετικά, αρνητικά, ουδέτερα), σε συστήματα που ανιχνεύουν αισθήματα και συναισθήματα (θυμωμένος, χαρούμενος, λυπημένος κ.λπ.) ή εντοπίζουν προθέσεις (π.χ. δεν ενδιαφέρομαι).

3. Συλλογή Δεδομένων

3.1 ΕΙΔΗΣΕΟΓΡΑΦΙΚΑ ΑΡΘΡΑ

Για οποιαδήποτε γλώσσα στον κόσμο, οι τίτλοι των εφημερίδων είναι πάντα σημαντικοί και με την ανάγνωση των τίτλων μπορούμε να έχουμε ιδέα ολόκληρων ειδήσεων χωρίς πλήρη ανάγνωση των άρθρων - ειδήσεων[15]. Σαν αναγνώστης είναι αρκετά εύκολο να δούμε ένα ειδησεογραφικό άρθρο ή έναν ειδησεογραφικό τίτλο και να καταλάβουμε και να ονομάσουμε τα θέματα που καλύπτονται, τα συναισθήματα που το διέπουν και την αντιμετώπιση (θετική – αρνητική – ουδέτερη) που έχει από το εν λόγω μέσο. Μπορεί στο άρθρο να χρησιμοποιούνται διαφορετικές λέξεις, αλλά είναι εύκολο να κατανοήσουμε την έννοια του - ακόμη και αν το θέμα δεν αναπαρίσταται ρητά στο κείμενο. Για παράδειγμα ένα άρθρο που μπορεί να αφορά την εκπαίδευση, να μην αναφέρει ποτέ την ίδια την λέξη «εκπαίδευση».

Στην σημερινή ψηφιακή εποχή, όπου υπάρχει ένα διευρυνόμενο χάσμα μεταξύ των απεριόριστων μέσων ενημέρωσης και της περιορισμένης προσοχής, οι τίτλοι των ειδήσεων έχουν ιδιαίτερο ρόλο ώστε να προσελκύσουν το κοινό[16]. Πολλές είναι πλέον οι ιστοσελίδες ειδησεογραφικών νέων, οι οποίες εξάγουν τους τίτλους ειδήσεων από τις εφημερίδες στο διαδίκτυο ώστε να ενημερώσουν τους διαδικτυακούς χρήστες τους.

Η κύρια λειτουργία τους είναι να προσελκύσουν την προσοχή και να λειτουργήσουν ως το οπτικό σημείο εισόδου στο ηλεκτρονικό ψηφιακό περιεχόμενο. Αυτό γίνεται εντονότερο τα τελευταία χρόνια, που όλο και περισσότεροι χρήστες ενημερώνονται οπτικά για γεγονότα, μέσω των δημοσιεύσεων των νέων στις σελίδες κοινωνικής δικτύωσης των ειδησεογραφικών φορέων, όπου σχεδόν πάντα οι τίτλοι είναι το μόνο ορατό κομμάτι του κύριου περιεχομένου. Μελέτες έχουν δείξει ότι σε σύγκριση με τα έντυπα μέσα ενημέρωσης οι ψηφιακοί αναγνώστες περνούν περισσότερο χρόνο στην περιήγηση, στην σάρωση και στον εντοπισμό των λέξεων-κλειδιών. Διάφορες μελέτες που πραγματοποιήθηκαν από τη Chartbeat διαπίστωσαν ότι το 38% των χρηστών εγκαταλείπουν μια ιστοσελίδα αμέσως μετά την πρόσβαση σε αυτήν και ότι ένας μέσος αναγνώστης θα δαπανήσει μόνο 15 δευτερόλεπτα σε μια ιστοσελίδα. Μια μελέτη του Αμερικανικού Ινστιτούτου Τύπου διαπίστωσε ότι περίπου έξι στους δέκα ανθρώπους, ελέγχουν μόνο τον τίτλο της είδησης και δεν διαβάζουν ολόκληρο το άρθρο.

Επομένως, απαιτείται αυτόματη επεξεργασία των τίτλων για να διευκολυνθεί η επιλογή και η ιεράρχηση μεγάλου όγκου ψηφιακού περιεχομένου. Η αυτόματη εξαγωγή των ειδήσεων από τίτλους μπορεί να αποτελέσει κεντρικό εργαλείο για μια σειρά εφαρμογών.

Οι αυτόματες εκτιμήσεις των τιμών των ειδήσεων μπορούν να συσχετιστούν με τις μετρήσεις της προσοχής στο διαδίκτυο, όπως οι προβολές σελίδας, για να διερευνήσουν ποιες επικεφαλίδες επηρεάζουν την δημοτικότητα του διαδικτύου. Αυτό απαιτεί προηγμένη επεξεργασία κειμένου για τον υπολογισμό των κατάλληλων.

3.2 TOPIC MODELING ΣΕ ΕΙΔΗΣΕΟΓΡΑΦΙΚΑ ΝΕΑ

Το Topic Modelling (Μοντελοποίηση Θέματος) αποτελεί έναν αρκετά αποτελεσματικό τρόπο ώστε να επιτύχουμε να φέρουμε σε δομή ένα αδόμητο κείμενο και τα δεδομένα του. Στο πλαίσιο των ειδήσεων, η Μοντελοποίηση Θέματος δεν είναι κάτι που μπορούμε απλά να εισάγουμε στο κείμενο και να αναμένουμε ότι αυτοματοποιημένα θα αντικαταστήσει τον χρόνο που δαπανούν οι δημοσιογράφοι για την ορθή ταξινόμηση και κατηγοριοποίηση των άρθρων. Ορισμένοι λόγοι για τους οποίους συμβαίνει αυτό είναι οι ακόλουθοι:

- Τα θέματα στη μοντελοποίηση θέματος δεν είναι τα ίδια με τα θέματα της ανθρώπινης κατανόησης. Για παράδειγμα το "Brexit" είναι ένα θέμα για εμάς, αλλά για τον LDA ή για οποιονδήποτε άλλο αλγόριθμο μοντελοποίησης θέματος σημαίνει μια κατανομή λεκτικών, πιθανότατα η εμφάνιση των λεκτικών "UK", "Europe", "EU", ίσως περιέχει και ονόματα ορισμένων πολιτικών που έλαβαν σημαντικό μέρος στην εξέλιξη του θέματος. Επιπλέον μπορεί να περιέχει και αφηρημένους όρους όπως "μετανάστευση" και "οικονομία". Ένα θέμα σε αυτό το πλαίσιο δεν έχει μια ετικέτα, αλλά η ετικέτα ενός θέματος είναι αυτή που συνδέει ένα άρθρο με ένα άλλο, το οποίο μεταφέρει το νόημα από τους κείμενο στους αναγνώστες. Για να εκχωρήσουμε ετικέτες σε θέματα θα χρειαστεί να εργαστούμε διαφορετικά, μιας και θα το πρόβλημα μας μετατρέπεται σε εποπτευόμενη μέθοδο εκμάθησης μηχανών.
- Η Μοντελοποίηση Θέματος έχει εφαρμοστεί σε διάφορους τομείς της επιστήμης και της γλωσσολογίας, από ιστορικά κείμενα έως γονιδιακή ανάλυση. Το κλειδί στις συγκεκριμένες αυτές περιπτώσεις είναι ότι το λεξιλόγιο του σώματος των κειμένων είναι σχετικά στατικό. Σε αντίθεση με τις ειδήσεις, που από τη φύση τους αλλάζουν συνεχώς. Οι φυσικές καταστροφές, οι πολιτικές αλλαγές, τα πολιτικά πρόσωπα, τα εγκλήματα και οι πολιτιστικές εκδηλώσεις φέρνουν μαζί τους ένα νέο λεξιλόγιο. Για παράδειγμα οι λέξεις "Brexit" - "Grexit", δεν υπήρχαν πριν από δέκα χρόνια. Εκτός αυτού, το λεξιλόγιο που επιλέγουμε για να μιλήσουμε και να γράψουμε για ένα θέμα σήμερα, διαφέρει από το λεξιλόγιο που χρησιμοποιούταν πριν από 20 χρόνια ή θα χρησιμοποιηθεί σε 20 χρόνια στο μέλλον. Τα νέα όπως και το λεξικό τους είναι ρευστά. Για να το υπολογίσουμε αυτό, θα μπορούσαμε να επαναπροσδιορίζουμε ένα μοντέλο θέματος πολύ συχνά ή να χρησιμοποιήσουμε εξαρτώμενα από το χρόνο μοντέλα για να υπολογίζουμε τη μετατόπιση. Σε αυτό το κομμάτι υπάρχουν έρευνες σε εξέλιξη αυτήν την στιγμή κι ίσως σύντομα έχουμε εξελίξεις.
- Το δύσκολο κομμάτι με τη Μοντελοποίηση Θέματος είναι ότι δεν είναι τόσο αυτόματη όσο νομίζουμε. Πρέπει να κάνουμε μια σωστή επιλογή στον αριθμό των θεματικών ενοτήτων που επιθυμούμε να ταξινομήσουμε τα κείμενα μας. Επιθυμούμε απλώς να τα χωρίσουμε σε δύο μεγάλες ομάδες ή ψάχνουμε εκατοντάδες ή ίσως χιλιάδες υποομάδες; Το αποτέλεσμα και οι αποφάσεις που θα πάρουμε είναι άρρηκτα συνδεδεμένες με το λεξιλόγιο και τον αριθμό των άρθρων που δίνονται. Η επιλογή του αριθμού των θεμάτων είναι θέμα βαθμού και επιπέδου λεπτομέρειας που θέλουμε να κατηγοριοποιήσουμε τα άρθρα. Για παράδειγμα μας ενδιαφέρει απλά ένα θέμα να περιγράφει την "επιστήμη" ή θέλουμε να «σκάψουμε» βαθύτερα στα νέα ώστε να ξεχωρίσουμε τα άρθρα που αφορούν την "τεχνολογία", την "κβαντική φυσική", τις αποστολές στον "Άρη"; Στο σημείο αυτό μια συνεργασία με τους δημοσιογράφους ή μια αναλυτικότερη απεικόνιση του στόχου που θέλουμε να υλοποιήσουμε θα μας οδηγήσει πιθανότατα σε μια πιο λογική και ρεαλιστική λύση, από αυτήν που θα μας οδηγήσει η αλγοριθμική αναζήτηση του τοπικού ελάχιστου.
- Μπορούμε να εφαρμόσουμε μια Μοντελοποίηση Θέματος στο σώμα ειδησεογραφικών νέων και να τα χωρίσουμε σε έναν δοσμένο αριθμό ομάδων. Το πρόβλημα είναι ότι ποτέ δεν μπορούμε να είμαστε απόλυτα σίγουροι ότι τα αποτελέσματα που πήραμε έχουν νόημα, μιας κι είμαστε απόλυτα εξαρτημένοι από τις μεταβλητές που χρησιμοποιήσαμε. Η Μοντελοποίηση Θέματος δεν ελέγχεται από τη φύση της, μιας και είναι μία μη εποπτευόμενη μέθοδο εκμάθησης μηχανών, αλλά μπορεί να επωφεληθεί από μια επαναληπτική διαδικασία με την παρέμβαση της ανθρώπινης κρίσης. Στο πλαίσιο που εργαζόμαστε, δηλαδή στα ειδησεογραφικά νέα, θα μπορούσαμε να παρουσιάσουμε σε έναν δημοσιογράφο που εργάζεται σε ένα ειδησεογραφικό φορέα τις υλοποιήσεις και τα θέματα που προτείνει ο αλγόριθμος μας και τις λέξεις που αντιπροσωπεύουν τις πιο πιθανές ομάδες θεμάτων. Μία απλή παρέμβαση από έναν δημοσιογράφο για κάθε μία από αυτές τις προτάσεις, θα μπορούσε να αποτελέσει ένα πολύτιμο σύνολο δεδομένων για να βελτιώσουμε διαδοχικά το μοντέλο μας, να βελτιώσουμε το λεξιλόγιο μας και να αυξήσουμε ή να μειώσουμε τον αριθμό των θεμάτων.

3.3 ΤΟ ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ ΤΟΥ ABC NEWS

Το σύνολο δεδομένων που θα χρησιμοποιήσουμε για την εξερεύνηση του μέσω του Topic Modelling είναι η συλλογή ειδησεογραφικών τίτλων “A Million News Headlines”, που δημοσιεύτηκε από τον Αυστραλιανό ραδιοτηλεοπτικό φορέα ABC News. Πρόκειται για ένα σύνολο εγγράφων, που έχει πάνω από ένα εκατομμύριο τίτλους ειδήσεων. Η συλλογή περιέχει τους τίτλους, οι οποίοι συλλέχθηκαν μεταξύ του 2003 και του 2017.

Επικεντρωθήκαμε στην ταξινόμηση και μοντελοποίηση θεμάτων των ειδήσεων ABC News, διότι θέλαμε να μην υπάρχει άλλη ετικέτα και ταξινόμηση στις ειδήσεις, αλλά να ακολουθεί την πραγματική ροή και εμφάνισή τους στον κόσμο. Δηλαδή δεν επιθυμούσαμε να έχουν κάποια συντακτική – κατηγοριοποιημένη ταξινόμηση σε ειδησεογραφικά τμήματα, για παράδειγμα πολιτικά, αθλητικά, οικονομία, κόσμος κτλ..

Στην πρωτογενή μορφή τους, οι τίτλοι ειδήσεων είναι διαθέσιμοι απλώς ως μια σειρά string κειμένου, συνοδευόμενα από την ημερομηνία δημοσίευσης τους. Μια μικρή παρουσίαση της πρωτογενούς τους παρουσίασης δίνεται παρακάτω:

Συνηθέστερα, η μοντελοποίηση θεμάτων διεξάγεται σε δεδομένα κειμένου μεγαλύτερης χρονικής διάρκειας, μαζί με το κείμενο ολόκληρου του άρθρου ειδήσεων ή με χρήση του αποσπάσματος του. Πράγματι, τα δείγματα μεγαλύτερου κειμένου είναι προτιμότερα, αφού ένα κείμενο με περισσότερες λέξεις σημαίνει μια πιο πλούσια εικόνα του θέματος και του συναισθήματος του και μας οδηγεί σε ένα διαφορετικό λεξιλόγιο θεμάτων. Ωστόσο, χάρη στον σύντομο και διαυγή χαρακτήρα των τίτλων των ειδήσεων, μπορούμε να περιμένουμε έναν ισχυρό πυρήνα σημασιολογικού περιεχομένου και σε συνδυασμό με τον τεράστιο αριθμό διαθέσιμων δεδομένων, είναι απίθανο η ανάλυσή μας να υποστεί έλλειψη βάθους.

Το συγκεκριμένο σύνολο δεδομένων αποτελεί έναν συνοπτικό ιστορικό πίνακα γεγονότων στον πλανήτη, με μια πιο λεπτομερή εστίαση στην Αυστραλία, από την σκοπιά, τον τρόπο ζωής και τις αντιλήψεις των δημοσιογράφων του ραδιοτηλεοπτικού φορέα ABC News, από τις αρχές του 2003 έως τα τέλη του 2017. Περιλαμβάνει το σύνολο των άρθρων που δημοσιεύονται από την ιστοσελίδα ABC στο συγκεκριμένο χρονικό διάστημα. Με έναν όγκο περίπου 200 άρθρων την ημέρα και μια καλή εστίαση στις διεθνείς ειδήσεις, μπορούμε να είμαστε αρκετά σίγουροι ότι κάθε σημαντικό γεγονός έχει καταγραφεί εδώ.

Εστιάζοντας στις λέξεις-κλειδιά και στις λέξεις με τις περισσότερες εμφανίσεις μπορούμε να δούμε όλα τα σημαντικά γεγονότα που έλαβαν χώρα την τελευταία δεκαετία και πώς εξελίχθηκαν με την πάροδο του χρόνου και διαμόρφωσαν την τωρινή κατάσταση. Παραδείγματα αποτελούν η οικονομική κρίση, ο πόλεμος του Ιράκ, οι εκλογές στις ΗΠΑ, οι οικολογικές καταστροφές, η τρομοκρατία, διάσημοι άνθρωποι, τα εγκλήματα της Αυστραλίας κλπ.

Περιεχόμενο:

- Μορφή: CSV
- publish_date: Ημερομηνία δημοσίευσης του άρθρου σε μορφή yyyyMMdd
- headline_text: Κείμενο του τίτλου σε Ascii, Αγγλικά, πεζά
- Ημερομηνία έναρξης: 2003-02-19 Ημερομηνία λήξης: 2017-12-31
- Συνολικές εγγραφές: 1.103.665

3.4 ΥΛΟΠΟΙΗΣΗ ΤΗΣ ΑΝΑΛΥΣΗΣ

Για το καλύτερο αποτέλεσμα και την χρησιμοποίηση όλων των σύγχρονων εργαλείων στην ανάλυση δεδομένων την εργασία την υλοποιήσαμε στο περιβάλλον του Anaconda. Το Anaconda δημιουργείται από την Continuum Analytics και αποτελεί μια Python διανομή που έρχεται προεγκατεστημένη με πολλές χρήσιμες βιβλιοθήκες της Python για την επιστήμη των δεδομένων.

Το Anaconda είναι δημοφιλές πρόγραμμα, επειδή φέρνει πολλά από τα εργαλεία που χρησιμοποιούνται για την ανάλυση δεδομένων και την εκμάθηση μηχανών. Σε απαιτητικές εργασίες όπως ένα έργο ανάλυσης δεδομένων, είναι βέβαιο ότι θα χρειαστούν πολλά διαφορετικά πακέτα (numpy, scikit-learn, scipy, pandas κτλ), τα οποία εγκαθίστανται μέσω της

προεγκατεστημένης εγκατάστασης του Anaconda. Εάν χρειαστούν επιπρόσθετα πακέτα, με χρήση του διαχειριστή πακέτων της Anaconda, το conda ή το pip, μπορούμε να υλοποιήσουμε την εγκατάσταση αυτών των πακέτων. Αυτό είναι εξαιρετικά επωφελές καθώς δεν χρειάζεται διαχείριση μεταξύ των εξαρτήσεων των πολλαπλών πακέτων. Στο Conda επίσης γίνεται εύκολα η εναλλαγή μεταξύ των εκδόσεων της Python 2 και της Python 3. Στην πραγματικότητα, μια εγκατάσταση του Anaconda είναι επίσης ο συνιστώμενος τρόπος για να εγκατάσταση του Jupyter Notebooks, το οποίο είναι το περιβάλλον του Anaconda που χρησιμοποιήσαμε.

Το Jupyter Notebooks αποτελεί έναν ιδανικό τρόπο για να γραφεί κώδικας και να υλοποιηθεί επαλήθευση του κώδικα Python για περιπτώσεις ανάλυσης δεδομένων. Αντί για την γραφή ολόκληρου του προγράμματος, δίνεται η επιλογή γραφής του κώδικα κομμάτι κομμάτι και μεμονομένη εκτέλεση τους. Στη συνέχεια, σε περίπτωση κάποιας αλλαγής, μπορεί να γίνει επιστροφή στο σημείο αυτό, να γίνει η εκ νέου κατάλληλη επεξεργασία και η επανεκκίνηση του προγράμματος ξανά, στο ίδιο παράθυρο.

Το Jupyter Notebook κατασκευάζεται από την IPython, και μας παρουσιάζει έναν διαδραστικό τρόπο εκτέλεσης του κώδικα Python στο τερματικό χρησιμοποιώντας το μοντέλο REPL (Read-Eval-Print-Loop). Ο πυρήνας της IPython εκτελεί τους υπολογισμούς και επικοινωνεί με τη διεπαφή front-end του Jupyter Notebook. Επιτρέπει επίσης στο Jupyter Notebook να υποστηρίζει πολλές γλώσσες. Τα σημειωματάρια Jupyter επεκτείνουν το IPython μέσω πρόσθετων λειτουργιών, όπως την αποθήκευση του κώδικα και της εξόδου του και επιτρέποντάς να διατηρηθούν σημειώσεις αξιολόγησης.

3.4.1 Βιβλιοθήκες Υλοποίησης

Για την εφαρμογή των διαφόρων ενεργειών και αλγορίθμων χρησιμοποιήθηκαν συγκεκριμένες βιβλιοθήκες της Python που περιέχουν μέσα τις απαραίτητες ενέργειες για ανάλυση δεδομένων.

3.4.1.1 Scikit-Learn

Υπάρχουν αρκετές βιβλιοθήκες της Python που παρέχουν κατάλληλες υλοποιήσεις για μια σειρά αλγορίθμων για μηχανική μάθηση. Ένα από τα πιο γνωστά είναι το Scikit-Learn, ένα πακέτο που παρέχει πολύ αποτελεσματικές εκδόσεις ενός μεγάλου αριθμού κοινών αλγορίθμων.

Το Scikit-learn είναι μία βιβλιοθήκη Python που ενσωματώνει ένα ευρύ φάσμα αλγόριθμων μηχανικής μάθησης τελευταίας τεχνολογίας για μεσαίου μεγέθους, επιτηρούμενα και μη επιτηρούμενα προβλήματα. Η βιβλιοθήκη αυτή επικεντρώνεται στην προσέλκυση μηχανογραφικής μάθησης, χρησιμοποιώντας μια γλώσσα υψηλού επιπέδου γενικής χρήσης.

Έμφαση δίνεται στην ευκολία χρήσης, τις επιδόσεις, την τεκμηρίωση και τη συνοχή του. Έχει ελάχιστες εξαρτήσεις και διανέμεται ενθαρρύνοντας τη χρήση του, τόσο σε ακαδημαϊκά όσο και σε εμπορικά περιβάλλοντα.

3.4.1.2 pyLDAvis

Το PyLDAvis[34], είναι ένα Python πακέτο που επιτρέπει μια διαδραστική απεικόνιση του Topic Modeling.. Το PyLDAvis βασίζεται στο LDAvis, ένα εργαλείο απεικόνισης για την γλώσσα R. Δημιουργήθηκε από τους Carson Sievert και Kenny Shirley.

Με λίγα λόγια, η διεπαφή παρέχει:

- ένα αριστερό πλαίσιο, που απεικονίζει μια συνολική εικόνα του μοντέλου (πόσο διαδεδομένη είναι κάθε θέμα και πώς σχετίζονται τα θέματα μεταξύ τους).
- ένα δεξιό πλαίσιο που περιέχει ένα διάγραμμα ράβδων. Οι ράβδοι αντιπροσωπεύουν τους όρους που είναι πιο χρήσιμοι στην ερμηνεία του επιλεγμένου θέματος (ποιο είναι το νόημα του κάθε θέματος).

Στα αριστερά, τα θέματα παρουσιάζονται ως κύκλοι, των οποίων τα κέντρα ορίζονται από την υπολογισμένη απόσταση μεταξύ των θεμάτων (που προβάλλονται σε 2 διαστάσεις). Η

επικράτηση του κάθε θέματος υποδεικνύεται από την περιοχή του κύκλου. Στην δεξιά πλευρά, δύο αντιμαχόμενοι παράλληλοι ράβδοι, δείχνουν τη συγκεκριμένη συχνότητα κάθε θέματος (με κόκκινο χρώμα) και τη συχνότητα σε ολόκληρο το σώμα (σε μπλε γκρι χρώμα). Όταν δεν έχει επιλεγεί κανένα θέμα, ο δεξιός πίνακας εμφανίζει τους 30 πιο σημαντικούς όρους για το σύνολο δεδομένων.

Το συγκεκριμένο εργαλείο καθορίζει τους πιο χρήσιμους όρους για την ερμηνεία ενός θέματος (και επίσης τον τρόπο με τον οποίο οι χρήστες μπορούν να τον αλλάξουν διαδραστικά). Προτείνουν ένα μέτρο που ονομάζεται relevance - συνάφεια, το οποίο είναι παρόμοιο με την αποκλειστικότητα, όπως ορίζεται από τους Bischof και Airoldi (2012): ο όρος αυτός υποδηλώνει το βαθμό στον οποίο ένας όρος εμφανίζεται σε ένα συγκεκριμένο θέμα αποκλείοντας τους άλλους.

Η συνάφεια βασίζεται σε μια άλλη μέτρηση, το lift, που είναι ορισμένη από τον Taddy (2011), η οποία είναι η αναλογία της πιθανότητας ενός όρου σε ένα θέμα, με την πιθανότητα του περιθωρίου του σε όλο το σώμα. Από τη μια πλευρά, μειώνει την κατάταξη των παγκόσμιων κοινών όρων, αλλά από την άλλη, δίνει μια υψηλή κατάταξη σε σπάνιους όρους που συμβαίνουν σε ένα ενιαίο θέμα. Το 2012, οι Bischof και Airoldi πρότειναν ένα νέο στατιστικό μοντέλο θεμάτων που συνάγει τη συχνότητα και την αποκλειστικότητα ενός όρου (αποκαλούμενη βαθμολογία FREX). Η μέθοδος των συγγραφέων είναι παρόμοια: ένας σταθμισμένος μέσος όρος των λογαρίθμων της πιθανότητας ενός όρου και της ανύψωσής του.

3.4.1.3 tqdm

Το tqdm[35] αποτελεί μια γρήγορη και εκτεταμένη γραμμή προόδου για την Python και για το CLI. Το tqdm σημαίνει "πρόοδος" στα αραβικά (taqadum) και μια σύντμηση για την έκφραση "σ 'αγαπώ τόσο πολύ" στα ισπανικά (te quiero demasiado).

Με το tqdm οι επαναληπτικοί βρόχοι παράγουν έναν έξυπνο μετρητή προόδου. Επιπλέον, το tqdm όχι μόνο παρουσιάζει την πρόοδό των επαναληπτικών βρόχων, αλλά παρέχει και κάποιες απλές μετρήσεις για το χρονικό διάστημα υλοποίησης του βρόχου κτλ.

3.4.1.4 Seaborn

Το Matplotlib έχει αποδειχθεί ότι είναι ένα απίστευτα χρήσιμο και δημοφιλές εργαλείο γραφικής απεικόνισης. Λόγω της εξέλιξης της ανάλυσης δεδομένων και των λεπτομερών αναπαραστάσεων που επιθυμούμε να δημιουργήσουμε η χρήση του Matplotlib μας περιορίζει.

Το Seaborn[33] αποτελεί μια ενδεδειγμένη λύση για να λυθούν αυτά τα προβλήματα. Το Seaborn παρέχει ένα API υψηλού επιπέδου, που προσφέρει λογικές επιλογές για το σχεδιαστικό στυλ, τις προεπιλογές χρώματος, καθορίζει απλές λειτουργίες υψηλού επιπέδου για κοινούς τύπους στατιστικών τύπων και ενσωματώνεται με τη λειτουργικότητα που παρέχεται από τα DataFrames του Pandas. Γενικά αποτελεί ιδανική επιλογή για τον σχεδιασμό στατιστικών γραφικών αναπαραστάσεων.

3.4.1.5 Bokeh Plotting

Το Bokeh[37] είναι μια βιβλιοθήκη διαδραστικής απεικόνισης, που η παρουσίαση του γίνεται στους διάφορους σύγχρονους Browsers (Περιηγητές ιστού). Παρέχει κομψή και συνοπτική κατασκευή ευέλικτων γραφικών και επεκτείνει αυτή την ικανότητα με υψηλής απόδοσης αλληλεπίδραση σε πολύ μεγάλα ή συνεχόμενα σύνολα δεδομένων. Ο Bokeh μπορεί να βοηθήσει οποιονδήποτε επιθυμεί να δημιουργήσει γρήγορα και εύκολα διαδραστικά γραφήματα, πίνακες εργαλείων και εφαρμογές δεδομένων.

Η βασική ιδέα του Bokeh είναι ότι τα γραφήματα δημιουργούνται ανά επίπεδο, ένα κάθε φορά. Αφού δημιουργήσουμε το σχήμα, στη συνέχεια προσθέτουμε τα διάφορα στοιχεία, που ονομάζονται glyphs, στο σχήμα. (Για όσους έχουν χρησιμοποιήσει το ggplot, η ιδέα των glyphs είναι ουσιαστικά η ίδια με αυτή των geoms που προστίθενται σε ένα γράφημα ένα επίπεδο τη

φορά). Τα Glyphs μπορούν να πάρουν πολλά σχήματα ανάλογα με την επιθυμητή χρήση: κύκλοι, γραμμές, τόξα και ούτω καθεξής.

3.4.1.6 Subprocess

Το subprocess παρέχει μια διεπαφή, για τη δημιουργία και τη λειτουργία με πρόσθετες διαδικασίες. Προσφέρει μια διεπαφή υψηλότερου επιπέδου σε σχέση με κάποια από τις άλλες διαθέσιμες μονάδες και προορίζεται να αντικαταστήσει λειτουργίες όπως οι os.system(), os.spawn(), os.popen(), popen2().

Το subprocess ορίζει μια κλάση, το Popen και μερικές συναρτήσεις που χρησιμοποιούν αυτή την κλάση. Ο κατασκευαστής για το Popen, παίρνει ορίσματα για να ρυθμίσει τη νέα διαδικασία έτσι ώστε ο γονέας να μπορεί να επικοινωνήσει μαζί του μέσω Pipes - Σωλήνων. Παρέχει όλες τις λειτουργίες των άλλων μονάδων, συναρτήσεις αντικατάστασης και πολλά άλλα. Το API είναι συνεπές για όλες τις χρήσεις και πολλά από τα επιπλέον βήματα που απαιτούνται (όπως κλείσιμο πρόσθετων περιγραφικών αρχείων και εξασφάλιση ότι τα pipes - σωλήνες είναι κλειστά) είναι "ενσωματωμένα", αντί να αντιμετωπίζονται ξεχωριστά από τον κώδικα εφαρμογής.

3.4.1.7 NLTK

Το Natural Language Toolkit - Εργαλείο φυσικής γλώσσας (NLTK) [38] είναι μια πλατφόρμα που χρησιμοποιείται για την κατασκευή προγραμμάτων Python που λειτουργούν με δεδομένα ανθρώπινης γλώσσας, για εφαρμογή στην επεξεργασία φυσικής γλώσσας (NLP).

Περιέχει βιβλιοθήκες επεξεργασίας κειμένου για tokenization, parsing, classification, stemming, tagging και semantic αιτιολογία. Περιλαμβάνει επίσης γραφικές αναπαραστάσεις και δείγματα από σύνολα δεδομένων καθώς κι ένα βιβλίο που εξηγεί τις αρχές πίσω από τις υποκείμενες εργασίες επεξεργασίας γλώσσας που υποστηρίζει η NLTK.

Το Natural Language Toolkit είναι μια βιβλιοθήκη ανοιχτού κώδικα για τη γλώσσα προγραμματισμού της Python, η οποία γράφτηκε αρχικά από τους Steven Bird, Edward Loper και Ewan Klein για χρήση στην ανάπτυξη λογισμικού και την εκπαίδευση. Περιλαμβάνει ένα πρακτικό οδηγό, που εισάγει θέματα υπολογιστικής γλωσσολογίας καθώς και θεμελιώδη προγραμματιστικά στοιχεία για την Python που το καθιστά κατάλληλο για γλωσσολόγους που δεν έχουν βαθιά γνώση στον προγραμματισμό, τους μηχανικούς και τους ερευνητές που πρέπει να βυθιστούν στην υπολογιστική γλωσσολογία, τους σπουδαστές και τους εκπαιδευτικούς.

Το NLTK περιλαμβάνει περισσότερα από 50 σώματα κειμένων και λεξιλογικές πηγές, όπως το Penn Treebank Corpus, το Open Multilingual Wordnet, το Corpus Report Problem και το Lin’s Dependency Thesaurus..

3.4.1.8 Gensim

Το Gensim είναι μια ελεύθερη βιβλιοθήκη Python που έχει σχεδιαστεί για να εξάγει αυτόματα τα σημασιολογικά θέματα από τα έγγραφα, με αποτελεσματικό κι ανώδυνο τρόπο. Το Gensim έχει σχεδιαστεί για να επεξεργάζεται ακατέργαστα, αδόμητα ψηφιακά κείμενα (απλά κείμενα).

Οι αλγόριθμοι στο Gensim, όπως το Word2Vec, το FastText, η Λανθάνουσα Σημασιολογική Ανάλυση (LSI, LSA, LsiModel), η Latent Dirichlet Allocation (LDA, βλέπε LdaModel) κλπ., Ανακαλύπτουν αυτόματα τη σημασιολογική δομή των εγγράφων εξετάζοντας πρότυπα συσχέτισης εντός ενός συνόλου εγγράφων εκπαίδευσης. Αυτοί οι αλγόριθμοι δεν είναι εποπτευμένοι, πράγμα που σημαίνει ότι δεν χρειάζεται ανθρώπινη είσοδος, χρειάζεται μόνο ένα σύνολο εγγράφων απλού κειμένου.

Μόλις βρεθούν αυτά τα στατιστικά πρότυπα, κάθε απλό κείμενο εγγράφων (πρόταση, φράση, λέξη) μπορεί να εκφραστεί ευρέως στη νέα, σημασιολογική αναπαράσταση και να διερευνηθεί για τυχόν θεματική ομοιότητα σε σύγκριση με άλλα έγγραφα.

4. Αποτελέσματα Ανάλυσης

4.1 ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

Το πρώτο βήμα σε κάθε ενέργεια εξόρυξης γνώσης από δεδομένα, είναι να μάθουμε τα δεδομένα. Πρέπει να ελέγξουμε τα δεδομένα, να τα μετρήσουμε, να δούμε πως απεικονίζονται κτλ. Πρέπει λοιπόν, πριν χρησιμοποιήσουμε τους αλγόριθμος Μοντελοποίησης Θεμάτων να υλοποιήσουμε τις βασικές ενέργειες προ-επεξεργασίας των δεδομένων και να μελετήσουμε τα δεδομένα μας, παράγοντας χρήσιμες πληροφορίες που θα καθοδηγήσουν την εργασία και τα αποτελέσματα μας.

Τα βήματα υλοποίησης της δομής μας είναι:

- Προ-επεξεργασία Δεδομένων
- Ανάλυση Διαγραμμάτων Χρονικών Περιόδων
- Εύρεση των ακραίων γεγονότων
- Ανάλυση των ακραίων γεγονότων
- Υλοποίηση Ανάλυσης Συναισθήματος
- Επηρεασμός του συναισθήματος στις περιόδους ακραίων γεγονότων
- Εύρεση των συνολικών 50 και αναφορά στις 10 κορυφαίες λέξεις που εκφράζουν τα θετικά και αρνητικά συναισθήματα
- Ανάλυση για τον συσχετισμό των λέξεων αυτών με το συναίσθημα στις περιόδους ακραίων γεγονότων
- Υλοποίηση του Αλγόριθμου μοντελοποίησης θεμάτων LDA, για 10 θέματα με χρήση της βιβλιοθήκης Gensim και αναφορά επίδοσης
- Σύγκριση των αλγορίθμων μοντελοποίησης θεμάτων LDA & LSA σε ένα δείγμα του σώματος των ειδησεογραφικών νέων
- Υλοποίηση LDA με χρήση της βιβλιοθήκης Scikit-Learn για 8 topics
- Topic Over Time
- Υλοποίηση του Αλγόριθμου μοντελοποίησης θεμάτων LDA για τα συνολικά δεδομένα, με χρήση της βιβλιοθήκης Scikit-Learn για 20 topics
- Ανάλυση της ανάθεσης εγγράφου ανά Topic
- Εξέλιξη των Topic στο χρόνο, σε ημερήσια, μηνιαία και ετήσια μορφή
- Ανάλυση των ακραίων τιμών που είχαμε εντοπίσει σε σχέση με τα topic, εντοπισμός γεγονότων που επηρεάζουν τα topics
- Ανάλυση ανά έτος των ειδησεογραφικών νέων και σύγκριση τους με το συνολικό
- Ανάλυση ανά εξέλιξη των ετών των ειδησεογραφικών νέων και σύγκριση τους με το ετήσιο και το συνολικό, ώστε να εντοπιστούν τα γεγονότα που προσάρμοσαν διαφορετικά τα θέματα
- Ανάλυση του έτους 2008 και του προηγούμενου και επόμενου που παρατηρούμε πολλές επικαλύψεις και εύρεση του γεγονότος που επηρέασε τις υλοποιήσεις μας

4.1.1 Προ-επεξεργασία δεδομένων

Σε όλες τις υλοποιήσεις, η πρώτη και βασική ενέργεια που υλοποιείται στα δεδομένα, είναι η προ-επεξεργασία τους. Τα δεδομένα μας είναι στην μορφή Ημερομηνία/Ειδησεογραφικοί Τίτλοι. Αποτελεί πολύ σημαντικό παράγοντα να κατανοήσουμε και να μάθουμε τα δεδομένα μας.

Οπότε μαζί με την προ-επεξεργασία γίνεται προσπάθεια κατανόησης και μάθησης του σώματος των δεδομένων μας. Μια πρώτη παρουσίαση των δεδομένων είναι η παρακάτω:

	publish_date	headline_text
0	2003-02-19	aba decides against community broadcasting lic...
1	2003-02-19	act fire witnesses must be aware of defamation
2	2003-02-19	a g calls for infrastructure protection summit
3	2003-02-19	air nz staff in aust strike for pay rise
4	2003-02-19	air nz strike to affect australian travellers

Παρουσίαση των δεδομένων όπως είναι στο αρχείο

Ανάλογα με την περίπτωση υλοποίησης χρησιμοποιήσαμε Lemmatization ή Stemming.

4.1.2 Ανάλυση Διαγραμμάτων Χρονικών Περιόδων

Για να καταφέρουμε να κατανοήσουμε λεπτομέρειες που κρύβονται στα δεδομένα, πρέπει να εξάγουμε όση περισσότερη γνώση μπορούμε για το περιεχόμενο των τίτλων ειδήσεων.

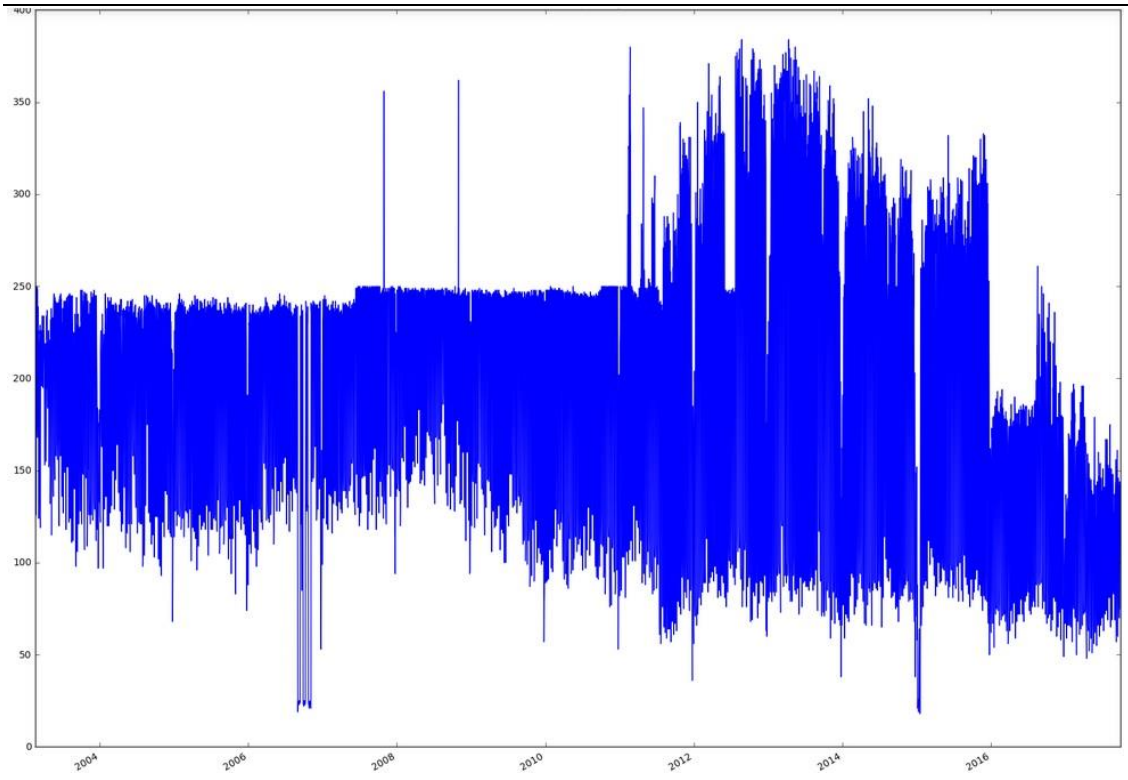
Αρχικά θα μετρήσουμε και θα αναλύσουμε τον όγκο των δεδομένων ανά χρονική περίοδο. Θα εξετάσουμε τον όγκο δεδομένων ανά μέρα, ανά μήνα κι ανά έτος.

Με μια απλή παρατήρηση στον όγκο ημερησίων τίτλων παρατηρούμε ότι μία μέση τιμή τίτλων ανά ημέρα είναι γύρω στις 210 – 230 τίτλοι. Οπότε σύμφωνα με αυτή την λογική μία μέση τιμή μηνιαίων άρθρων είναι γύρω στα 6200 – 6800. Επιπλέον παρατηρούμε κάποιες τρομερά μεγάλες διακυμάνσεις μικρής ή μεγάλης έντασης στον όγκο των δεδομένων. Είναι λογικό αν αναλογιστούμε ότι το σώμα των δεδομένων αφορά ειδησεογραφικά θέματα. Στην πραγματικότητα οι ακραίες τιμές σηματοδοτούν την αρχή κάποιου σημαντικού γεγονότος ή μια ημέρα με πολλά θέματα στην επικαιρότητα (δηλαδή, θέματα υπουργικών αποφάσεων και κυβερνήσεων, αστυνομικά θέματα - τροχαία, πόλεμο ή επίθεση αυτοκτονίας κτλ). Στην περίπτωση που οι ακραίες τιμές δημιουργούνται εξαιτίας κάποιου σημαντικού γεγονότος , αυτό σημαίνει ότι το γεγονός συνέβη ακριβώς εκείνη την στιγμή και τα άρθρα για αυτό το γεγονός διαδέχονται το ένα το άλλο. Ένα άλλο σημαντικό θέμα που θα πρέπει να κατανοήσουμε είναι η διαφορά που υπάρχει στην συνολική αύξηση ειδησεογραφικών τίτλων σε σχέση με την ημερήσια

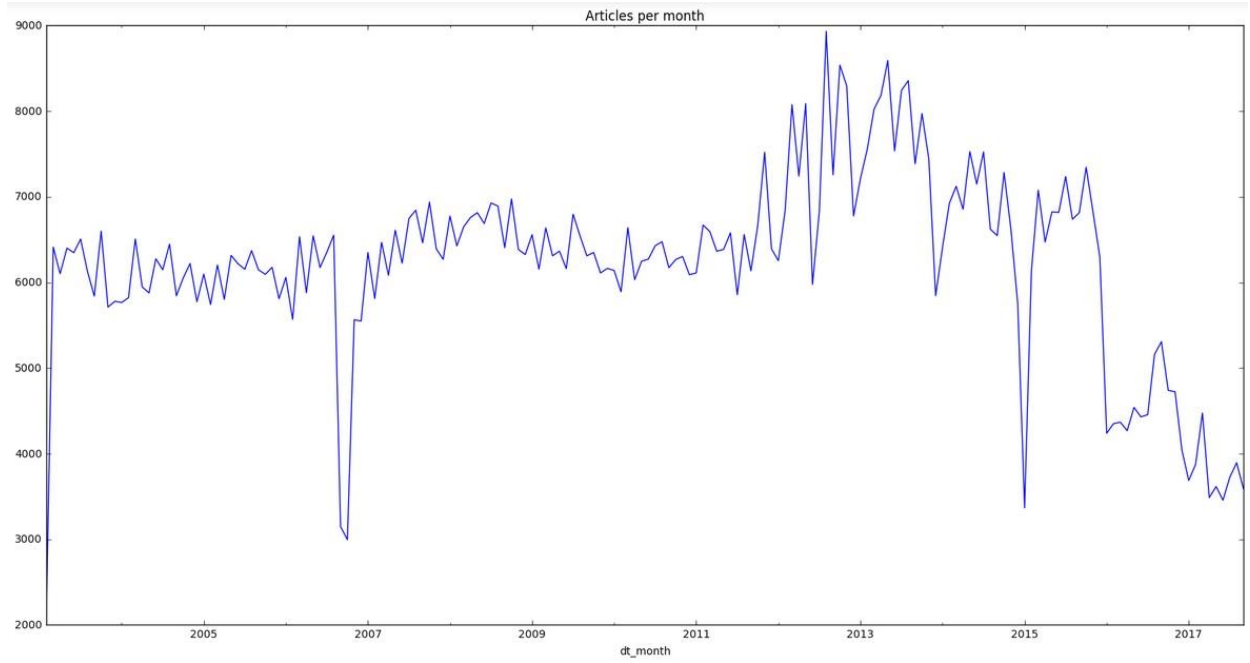
Οι μηνιαίες μετρήσεις του όγκου των ειδησεογραφικών νέων, όπως είναι λογικό ακολουθεί τη γενική πορεία των ημερήσιων νέων. Μας δίνει μια γενική ιδέα, της πορείας των αυξομειώσεων των τίτλων ειδήσεων που παρατηρήσαμε στον όγκο των ημερήσιων τίτλων.

Στην ανάλυση των μετρήσεων μας, παρατηρούμε ότι ο αριθμός των τίτλων τα έτη από 2003 έως 2006 έχουν έναν σταθερό ρυθμό αυξομείωσης τους. Προς τα τέλη του 2006 παρατηρούμε αρκετές αρνητικές επιδόσεις με τους ημερήσιους τίτλους να πέφτουν συχνά σε κάτω των 50. Λίγο πριν έρθει το 2018 και το 2019 παρατηρούμε μικρές αυξήσεις για κάποιες συγκεκριμένες

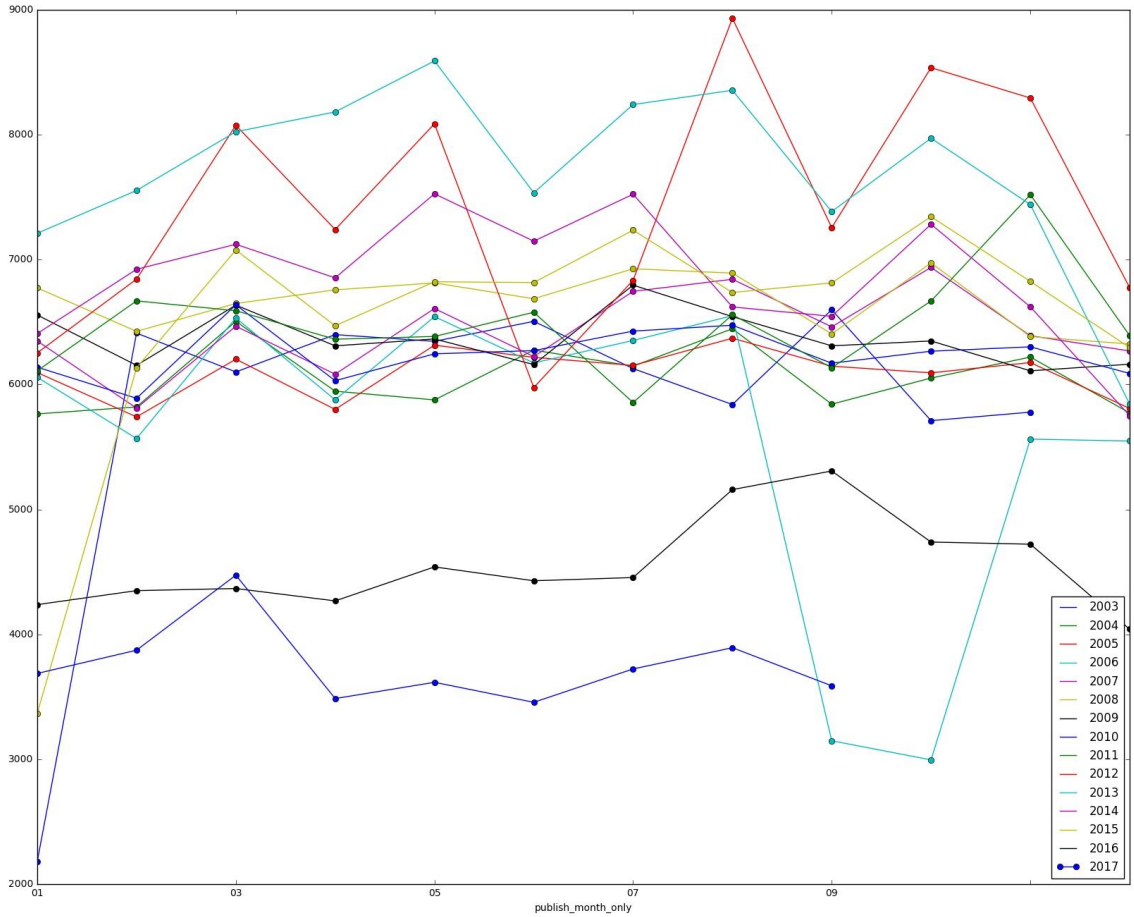
ημέρες και μετά επιστροφή σε ένα περίπου κοινό μοτίβο. Τα έτη 2012, 2013, 2014, 2015, 2016 έχουμε αύξηση στον ημερήσιο αριθμό των τίτλων.



Πλήθος ειδησεογραφικών νέων ανά ημέρα



Πλήθος ειδησεογραφικών νέων ανά μήνα



Πλήθος ειδησεογραφικών νέων ανά έτος και μήνα

4.1.3 Ανάλυση των ακραίων γεγονότων

Για την εύρεση του γεγονότος που προκάλεσε την απότομη αύξηση των τίτλων ειδήσεων σε μια συγκεκριμένη ημέρα, αρκεί να δούμε την ημερομηνία που εμφανίζεται η άνοδος αυτή. Από την στιγμή εύρεσης της ημερομηνίας και μέσω μιας υλοποίησης που μετράει το σύνολο των λέξεων ανά ημέρα, μπορούμε να εντοπίσουμε το γεγονός που προκάλεσε αυτή την αύξηση στον αριθμό των άρθρων. Θα μελετήσουμε τα τέσσερα πρώτα γεγονότα που εμφανίζονται.

- Γεγονός Α: 2007-10-31, 356 τίτλοι. Στην συγκεκριμένη ημερομηνία ανόδου των ειδησεογραφικών νέων παρατηρούμε κι αύξηση τις προηγούμενες δύο ημέρες και την επόμενη μέρα.
  Οι περισσότερο χρησιμοποιημένες λέξεις αυτήν την ημερομηνία είναι οι:

```
20071031  Best words:
        health
        govt
        debate
        report
        council
        abbott
        climate
        nsw
        police
        man
```

Οι 10 λέξεις με μεγαλύτερη εμφάνιση στις 31/10/2007

- Γεγονός Β: 2008-10-31, 362 τίτλοι. Στην συγκεκριμένη ημερομηνία οι ειδησεογραφικοί τίτλοι έχουν αύξηση τις προηγούμενες ημέρες και απότομη μείωση την επόμενη μέρα.
  Οι περισσότερο χρησιμοποιημένες λέξεις αυτήν την ημερομηνία είναι οι:

```
20081031  Best words:
        govt
        guilty
        police
        man
        crash
        new
        home
        court
        croc
        death
```

Οι 10 λέξεις με μεγαλύτερη εμφάνιση στις 31/10/2008

- Γεγονός Γ: 2011-02-14, 326 τίτλοι. Στην συγκεκριμένη ημερομηνία έχουμε πολύ απότομη αύξηση των ειδησεογραφικών νέων. Αποτελεί την αρχή συνεχόμενων ημερών υψηλού όγκου. Με μέγιστη τιμή στις 2011-02-18, που είχε 354 τίτλους.
  Οι περισσότερο χρησιμοποιημένες λέξεις αυτήν την ημερομηνία είναι οι:

```
20110214  Best words:
        health
        deal
        police
        flood
        man
        shooting
        cyclone
        grammys
        injured
        crash
```

Οι 10 λέξεις με μεγαλύτερη εμφάνιση στις 14/02/2011

- Γεγονός Δ: 2011-02-21, 360 τίτλοι. Στην συγκεκριμένη ημερομηνία έχουμε απότομη άνοδο στον όγκο των τίτλων. Συνεχίζει ο υψηλός όγκος τίτλων και τις επόμενες 4 ημέρες.
  Οι περισσότερο χρησιμοποιημένες λέξεις αυτήν την ημερομηνία είναι οι:

```
20110221  Best words:
        man
        libyan
        police
        flood
        charged
        court
        death
        camp
        nsw
        bashing

20110222  Best words:
        quake
        christchurch
        gm
        libya
        flood
        contamination
        canola
        help
        cyclone
        gaddafi

20110223  Best words:
        quake
        christchurch
        nz
        man
        police
        rescue
        gaddafi
        earthquake
        cyclone
        bulls

20110224  Best words:
        man
        christchurch
        quake
        cyclone
        flood
        carbon
        police
        price
        continues
        libya
```

Οι 10 λέξεις με μεγαλύτερη εμφάνιση στις 21-22-23-24/10/2011

Στις συγκεκριμένες μέρες τα άσχημα γεγονότα σε Αίγυπτο και Λιβύη και ένας σεισμός προκάλεσαν την αύξηση των τίτλων και των ειδήσεων. Εδώ έχουμε την περίπτωση που δύο συγκεκριμένα γεγονότα, προκάλεσαν την γέννηση πολλών τίτλων όσον αφορά τα συγκεκριμένα θέματα.

4.2 ΥΛΟΠΟΙΗΣΗ ΑΝΑΛΥΣΗΣ ΣΥΝΑΙΣΘΗΜΑΤΟΣ

Όταν το σώμα των κειμένων μας αφορά ειδησεογραφικά νέα, μία από τις μεγαλύτερες προκλήσεις που αντιμετωπίζουμε είναι το συναίσθημα που κρύβεται πίσω από την άποψη που καταθέτει στο κοινό ο εκάστοτε δημοσιογράφος. Στην περίπτωση μας η πρόκληση γίνεται μεγαλύτερη αφού έχουμε μόνο τους τίτλους από τα ειδησεογραφικά νέα. Ένας ειδησεογραφικός τίτλος συνήθως περιέχει 4 έως 8 λέξεις. Στους τίτλους συνήθως ή γράφεται η είδηση αυτή καθεαυτή ή γράφεται απλά το θέμα με μια έκφραση συναισθήματος του δημοσιογράφου.

```
Number of Positive Headlines : 1826
Number of Negative Headlines : 3224
Number of Neutral Headlines : 5610
```

Το πλήθος των ουδέτερων, θετικών και αρνητικών τίτλων

Από τα πρώτα μας αποτελέσματα παρατηρούμε ότι οι περισσότεροι τίτλοι δίνουν τον τόνο της ουδέτερης άποψης και του ουδέτερου συναισθήματος. Ακολουθεί το αρνητικό συναίσθημα και τέλος το θετικό.

Λόγω ότι το σώμα των κειμένων αφορά ειδησεογραφικές ειδήσεις ο διαμοιρασμός στις κατηγορίες θεμάτων ουδέτερο-αρνητικό- θετικό είναι απόλυτα σωστός. Πολλές από τις κατηγορίες θεμάτων που απασχολούν τα Μέσα Μαζική Ενημέρωσης παρουσιάζονται σε μια ουδέτερη βάση, ώστε να αποφύγουν τα Μ.Μ.Ε. να χαρακτηριστούν ως προσκείμενα σε κάποια πλευρά. Οπότε θέματα που έχουν να κάνουν με τον αθλητισμός, με την πολιτική και πολλά άλλα παρουσιάζονται υπό ένα ουδέτερο πρίσμα.

Οι ειδήσεις που κάνουν ¨θόρυβο¨ συνήθως είναι οι κακές-αρνητικές ειδήσεις. Συνήθως μετά από ένα αρνητικό ή καταστροφικό συμβάν, τα εκάστοτε ειδησεογραφικά γραφεία αποκτούν μεγάλη ανταπόκριση και όλοι περιμένουν την επόμενη είδηση. Όσο πιο αρνητική είναι η είδηση τόσο μεγαλύτερη απήχηση έχει.

Οι θετικές ειδήσεις σπάνια θα έχουν αυτή την ροή και την δημοτικότητα των αρνητικών. Θα γραφτούν κάποιοι τίτλοι, αλλά πολύ σύντομα θα σβήσουν σαν θέματα επικαιρότητας.

Τα στοιχεία αυτά φυσικά αλλάζουν από δημοσιογραφικό μέσο σε άλλο γιατί κάθε φορά είναι διαφορετική η άποψη του συγγραφέα.

Μία αρχική προσέγγιση είναι να υλοποιήσουμε Ανάλυση Συναισθήματος ανά έτος. Να ελέγξουμε ίσως κάποια γεγονότα μέσα στα έτη που επηρέασαν την ανθρωπότητα ή τον εκάστοτε δημοσιογράφο.

```
2003
[('u', 2452), ('police', 2335), ('govt', 1933), ('new', 1764), ('man', 1599), ('plan', 1551), ('council', 1439
), ('say', 1310), ('iraq', 1290), ('call', 1211), ('court', 1108), ('win', 1090), ('claim', 1025), ('back', 99
9), ('face', 932), ('fire', 861), ('report', 840), ('nsw', 827), ('world', 819), ('get', 788)]
2004
[('police', 2768), ('u', 2226), ('new', 2056), ('govt', 1991), ('plan', 1870), ('man', 1665), ('council', 1632
), ('say', 1364), ('iraq', 1298), ('call', 1281), ('back', 1216), ('win', 1146), ('claim', 1087), ('court', 10
61), ('fire', 1047), ('boost', 956), ('report', 924), ('face', 919), ('urged', 888), ('death', 844)]
2005
[('police', 2934), ('govt', 2408), ('new', 2097), ('council', 1863), ('plan', 1821), ('man', 1680), ('say', 15
68), ('u', 1431), ('call', 1306), ('back', 1295), ('fire', 1104), ('urged', 1063), ('court', 1053), ('claim',
1004), ('death', 1002), ('seek', 1000), ('win', 977), ('group', 954), ('get', 938), ('face', 932)]
2006
[('police', 2458), ('govt', 2272), ('new', 1726), ('plan', 1563), ('man', 1483), ('council', 1479), ('say', 14
72), ('call', 1248), ('fire', 1147), ('water', 1143), ('back', 1114), ('u', 1038), ('win', 984), ('court', 965
), ('pm', 928), ('urged', 912), ('crash', 889), ('death', 847), ('face', 839), ('claim', 796)]
2007
[('police', 3331), ('govt', 2670), ('man', 2121), ('new', 1994), ('say', 1944), ('council', 1770), ('water', 1
725), ('plan', 1698), ('u', 1389), ('court', 1281), ('call', 1258), ('fire', 1168), ('back', 1167), ('pm', 110
1), ('crash', 1092), ('death', 1046), ('qld', 1042), ('labor', 999), ('report', 986), ('nsw', 960)]
2008
[('police', 3141), ('govt', 2546), ('man', 2457), ('new', 2031), ('say', 1500), ('u', 1364), ('plan', 1349), (
'court', 1316), ('council', 1286), ('qld', 1280), ('call', 1243), ('back', 1164), ('crash', 1052), ('death', 1
043), ('cut', 1011), ('water', 1006), ('fire', 997), ('report', 991), ('woman', 984), ('win', 982)]
2009
[('police', 2669), ('man', 2175), ('interview', 1926), ('new', 1737), ('fire', 1532), ('govt', 1418), ('counci
l', 1183), ('court', 1160), ('plan', 1149), ('back', 1129), ('death', 1084), ('call', 1076), ('qld', 1063), ('
say', 1061), ('crash', 1056), ('job', 967), ('water', 925), ('get', 907), ('win', 865), ('woman', 861)]
```

```
2010
[('interview', 2822), ('police', 2405), ('man', 2345), ('new', 1658), ('council', 1230), ('plan', 1218), ('say
', 1201), ('back', 1189), ('court', 1187), ('crash', 1002), ('fire', 990), ('death', 985), ('get', 982), ('cal
l', 979), ('water', 927), ('woman', 862), ('win', 839), ('health', 808), ('charged', 802), ('accused', 799)]
2011
[('police', 2102), ('man', 2029), ('interview', 1784), ('new', 1761), ('flood', 1679), ('call', 1403), ('say',
1182), ('council', 1114), ('back', 1068), ('plan', 1062), ('court', 1053), ('fire', 1024), ('get', 987), ('abc
', 983), ('win', 979), ('crash', 918), ('hit', 915), ('death', 903), ('report', 823), ('water', 808)]
2012
[('interview', 2604), ('police', 2241), ('new', 2037), ('man', 1986), ('abc', 1723), ('say', 1502), ('call', 1
315), ('market', 1312), ('fire', 1281), ('court', 1257), ('plan', 1228), ('win', 1226), ('council', 1203), ('r
eport', 1124), ('australia', 1054), ('australian', 1035), ('news', 1033), ('back', 1018), ('cut', 1003), ('dea
th', 887)]
2013
[('new', 2534), ('police', 2438), ('interview', 2280), ('man', 2279), ('say', 1949), ('fire', 1607), ('rural',
1521), ('council', 1432), ('call', 1430), ('court', 1384), ('nsw', 1319), ('australia', 1302), ('qld', 1290),
('australian', 1266), ('plan', 1242), ('2013', 1156), ('market', 1143), ('win', 1120), ('government', 1048), (
'woman', 1047)]
2014
[('new', 2405), ('say', 2288), ('police', 2127), ('man', 2056), ('rural', 1908), ('interview', 1832), ('2014',
1723), ('country', 1601), ('australia', 1596), ('nsw', 1530), ('australian', 1504), ('hour', 1500), ('win', 14
27), ('council', 1261), ('call', 1258), ('wa', 1237), ('court', 1149), ('qld', 1136), ('government', 1091), ('
fire', 1081)]
2015
[('say', 2540), ('new', 2503), ('police', 2172), ('man', 2164), ('2015', 1854), ('australia', 1746), ('austral
ian', 1674), ('country', 1586), ('nsw', 1537), ('wa', 1462), ('hour', 1453), ('win', 1397), ('rural', 1383), (
'government', 1284), ('year', 1238), ('woman', 1206), ('call', 1197), ('queensland', 1166), ('world', 1163), (
'court', 1089)]
2016
[('say', 1906), ('man', 1871), ('police', 1662), ('australia', 1548), ('new', 1521), ('australian', 1288), ('w
oman', 1233), ('wa', 1093), ('election', 1070), ('sydney', 1029), ('u', 1018), ('year', 998), ('nsw', 982), ('
2016', 961), ('government', 924), ('melbourne', 917), ('win', 913), ('court', 909), ('day', 849), ('call', 849
)]
2017
[('say', 1295), ('trump', 1232), ('australia', 1163), ('new', 921), ('police', 895), ('australian', 890), ('wa
', 797), ('man', 760), ('woman', 701), ('u', 665), ('year', 605), ('nsw', 596), ('government', 592), ('sydney'
, 585), ('donald', 546), ('court', 542), ('day', 523), ('attack', 504), ('north', 481), ('call', 472)]
```

Οι 10 λέξεις με την μεγαλύτερη ετήσια συχνότητα ανά έτος

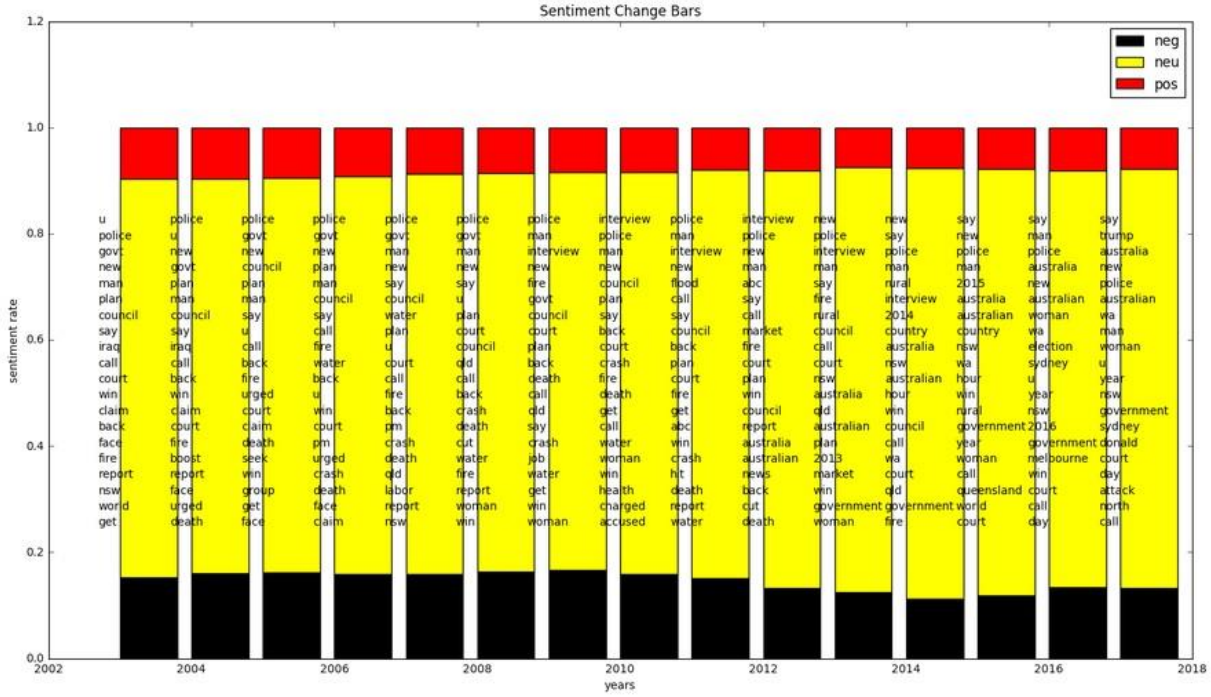
Αναλύουμε τα κείμενα στις 20 πιο εμφανιζόμενες λέξεις ανά έτος. Έτσι χωρίζουμε τα άρθρα μας σε άρθρα ουδέτερης άποψης, αρνητικής άποψης και θετικής άποψης. Παρατηρούμε ότι τα

συναισθήματα τροποποιούνται ανά έτος. Η τροποποίηση δεν έχει μεγάλη ή απότομη διακύμανση. Εξαρτάται από τα άρθρα αρνητικού συναισθήματος.

```
2003
neg:0.153 ,neu:0.751 ,pos:0.097 ,compound:-0.070
2004
neg:0.160 ,neu:0.743 ,pos:0.097 ,compound:-0.078
2005
neg:0.162 ,neu:0.743 ,pos:0.095 ,compound:-0.082
2006
neg:0.159 ,neu:0.749 ,pos:0.093 ,compound:-0.081
2007
neg:0.159 ,neu:0.754 ,pos:0.087 ,compound:-0.090
2008
neg:0.163 ,neu:0.751 ,pos:0.086 ,compound:-0.094
2009
neg:0.167 ,neu:0.749 ,pos:0.084 ,compound:-0.098
2010
neg:0.158 ,neu:0.758 ,pos:0.084 ,compound:-0.088
2011
neg:0.151 ,neu:0.770 ,pos:0.079 ,compound:-0.083
2012
neg:0.132 ,neu:0.787 ,pos:0.080 ,compound:-0.061
2013
neg:0.125 ,neu:0.801 ,pos:0.074 ,compound:-0.063
2014
neg:0.112 ,neu:0.811 ,pos:0.076 ,compound:-0.049
2015
neg:0.119 ,neu:0.803 ,pos:0.078 ,compound:-0.060
2016
neg:0.135 ,neu:0.784 ,pos:0.082 ,compound:-0.080
2017
neg:0.132 ,neu:0.790 ,pos:0.078 ,compound:-0.081
```

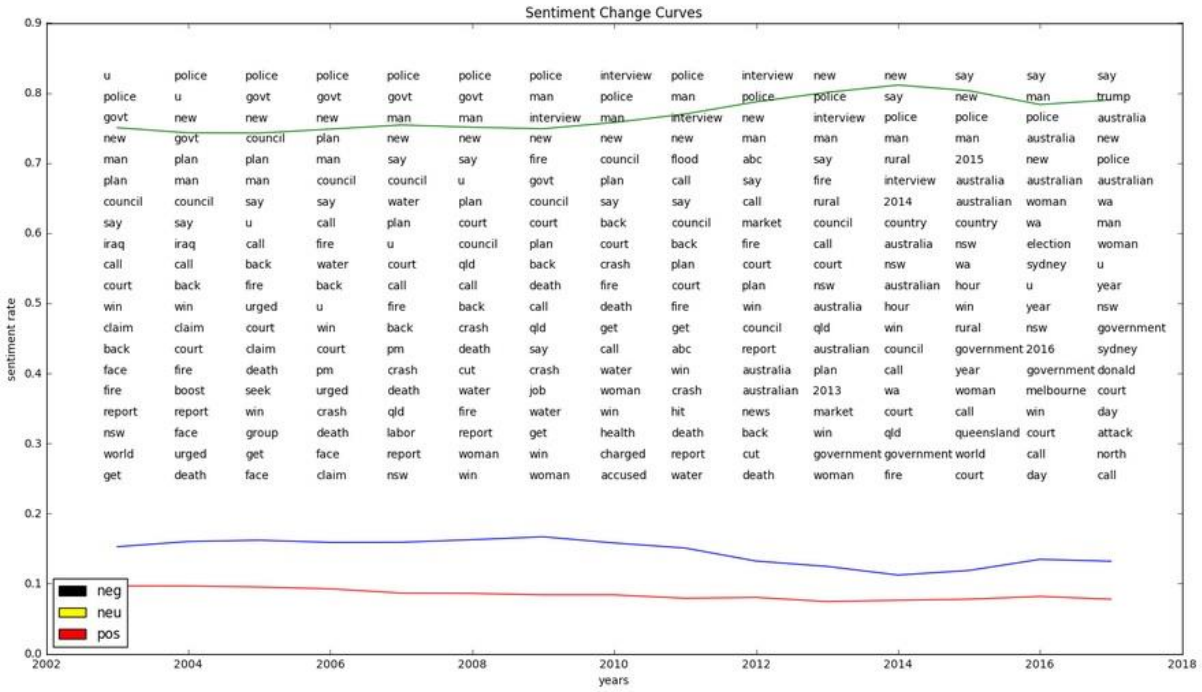
Ανάλυση συναισθήματος – άποψης ανά έτος

Παρουσιάζοντας την κατηγοριοποίηση των δεδομένων σε δενδρόγραμμα επί του συνολικού όγκου ανά έτος και ως διακύμανση ανά έτος παρατηρούμε την διαφοροποίηση των δεδομένων ως προς την άποψη τους.



Γράφημα εξέλιξης συναισθήματος ανά έτος

Θεματική μοντελοποίηση σε σώμα ειδησεογραφικών κειμένων



Καμπύλη διακύμανσης συναισθήματος

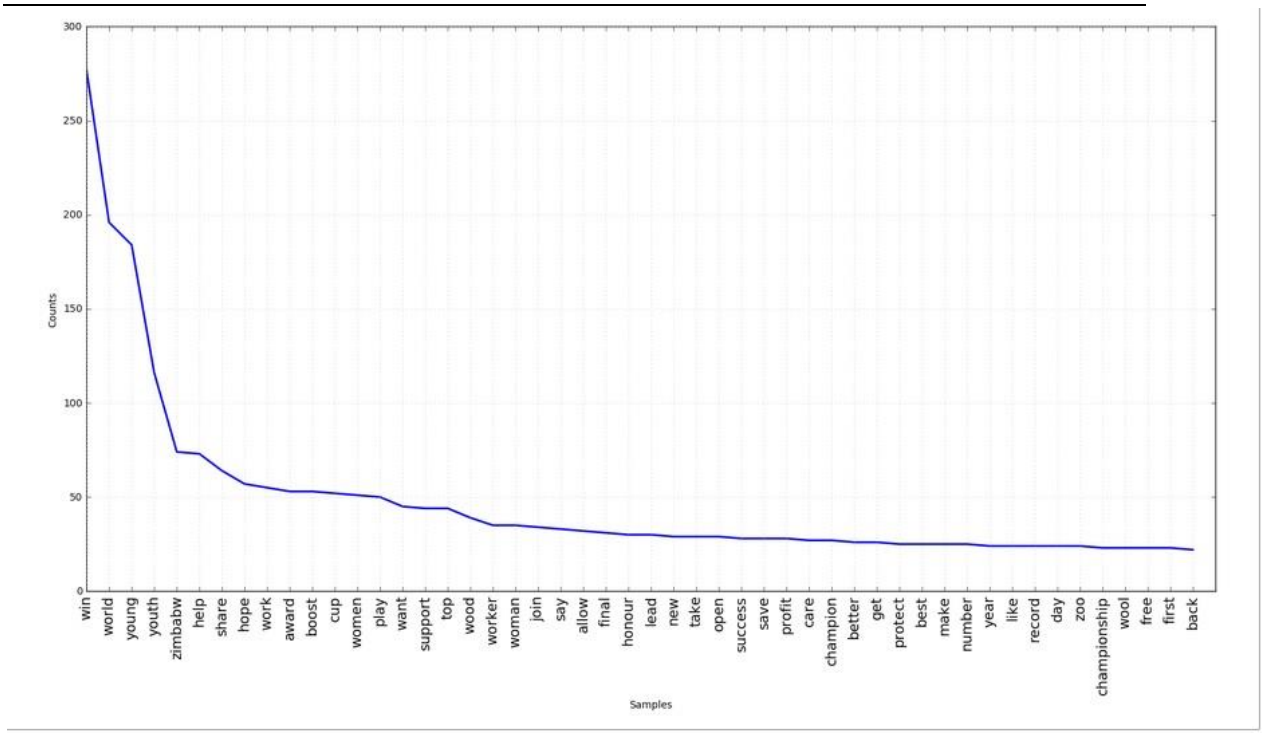
4.2.1 Εύρεση των συνολικών 50 και αναφορά στις 10 κορυφαίες λέξεις που εκφράζουν τα θετικά και αρνητικά συναισθήματα

Από το σύνολο των δεδομένων μας, είναι πολύ χρήσιμο να βρούμε τις λέξεις που εκφράζουν το θετικό και το αρνητικό συναίσθημα. Θα μπορούμε να προβλέψουμε την κατηγοριοποίηση του συναισθήματος νέων κειμένων.

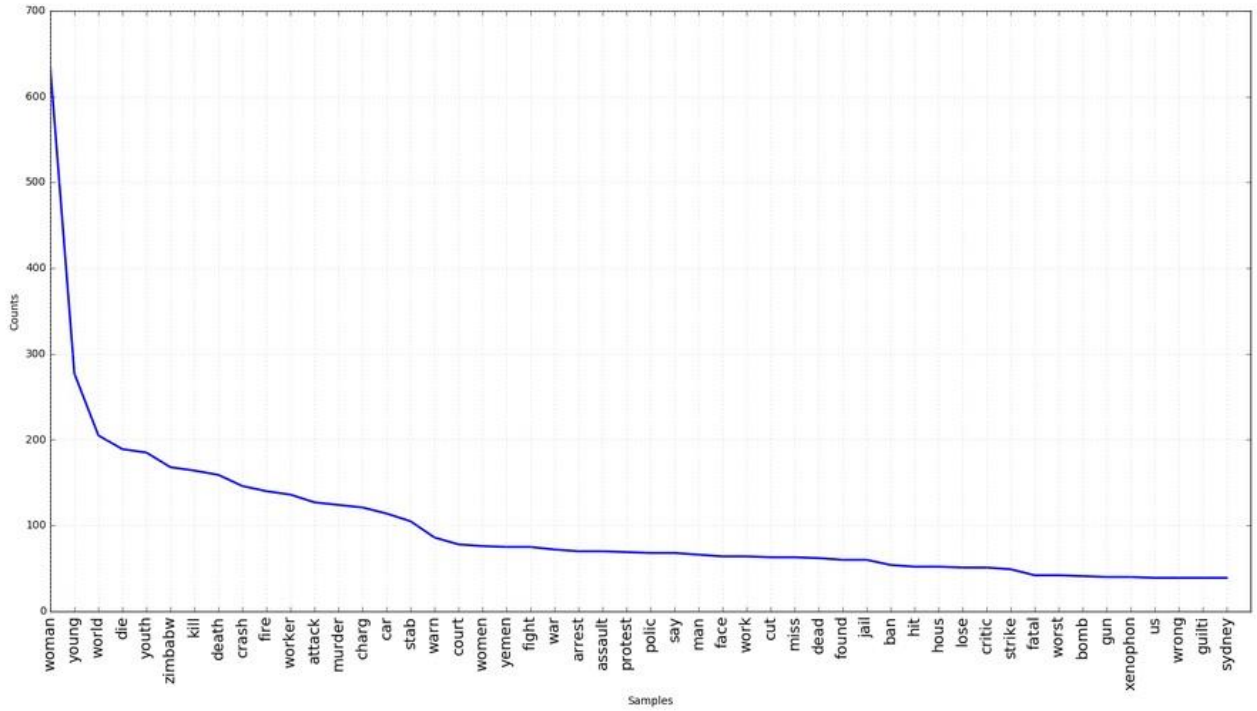
```
Most Positive Words

win : 277
world : 196
young : 184
youth : 116
zimbabw : 74
help : 73
share : 64
hope : 57
work : 55
award : 53

Most Negative Words

woman : 632
young : 277
world : 205
die : 189
youth : 185
zimbabw : 168
kill : 164
death : 159
crash : 146
fire : 140
```

Οι 10 λέξεις με μεγαλύτερη εμφάνιση που εκφράζουν το θετικό και το αρνητικό συναίσθημα

Πολλές λέξεις είναι κοινές στα θετικά και στα αρνητικά. Συνήθως μια αρνητική – καταστροφική είδηση, ακολουθούν κάποιες θετικές ειδήσεις. Μετά από ένα καταστροφικό γεγονός δημιουργούνται δράσεις αλληλεγγύης και βοήθειας που προβάλλονται και παίρνουν δημοτικότητα.



Η συχνότητα εμφάνισης των λέξεων που εκφράζουν θετική άποψη



Η συχνότητα εμφάνισης των λέξεων που εκφράζουν αρνητική άποψη

4.3 ΥΛΟΠΟΙΗΣΗ ΤΟΥ ΑΛΓΟΡΙΘΜΩΝ ΜΟΝΤΕΛΟΠΟΙΗΣΗΣ ΘΕΜΑΤΩΝ

Το επόμενο βήμα, μετά την προ-επεξεργασία των δεδομένων, την καταμέτρηση των συχνά εμφανιζόμενων λέξεων και την ανάλυση των συναισθημάτων, είναι το Topic Modelling – Μοντελοποίηση Θεμάτων.

Αρχικά θα πρέπει να βρούμε τον αλγόριθμο που θα ανταποκριθεί καλύτερα στα δεδομένα και στην εργασία μας.

4.3.1 Προ-επεξεργασία για την χρήση αλγορίθμων μοντελοποίησης

Ένα βήμα προ-επεξεργασίας που απαιτείται σε αυτές τις περιπτώσεις για να μπορέσουμε να υλοποιήσουμε στην συνέχεια τον οποιοδήποτε αλγόριθμο μοντελοποίησης είναι η κατασκευή χαρακτηριστικών. Δηλαδή αναπαριστούμε τους τίτλους σε ένα διάστημα χαρακτηριστικών. Στην πράξη αναπαριστούμε τις συμβολοσειρές – λέξεις σε αριθμητικό διάνυσμα.

Για παράδειγμα τον 125^{ο} τίτλο που είναι: «territory eels to play gold coast titans in Darwin», θα γίνει: (0, 11094) 1

(0, 2366) 1

(0, 4778) 1

(0, 8187) 1

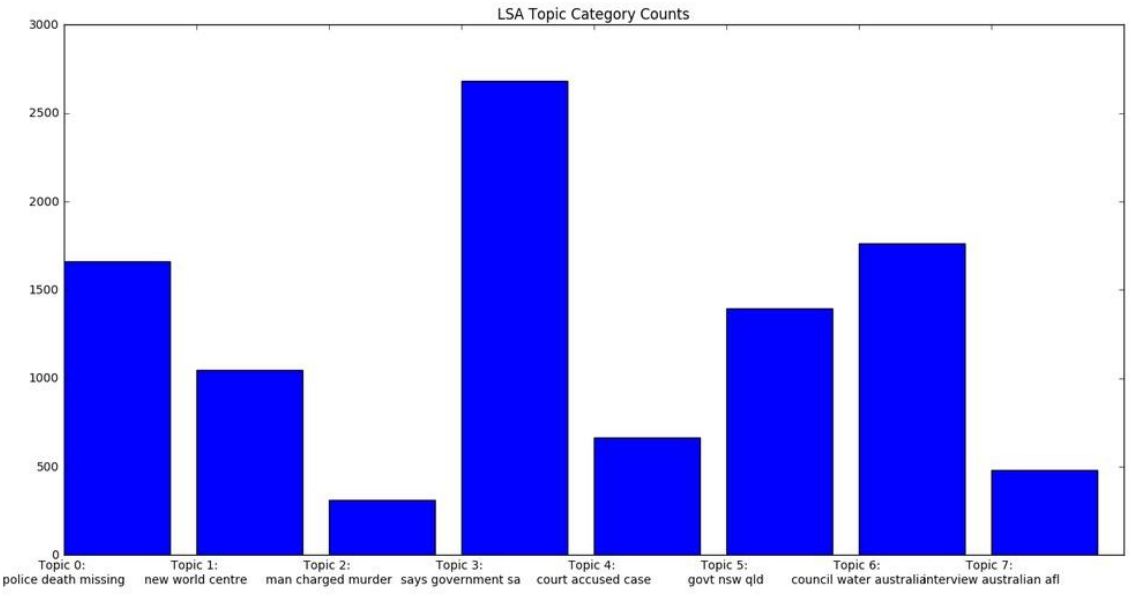
(0, 10956) 1

(0, 3687) 1

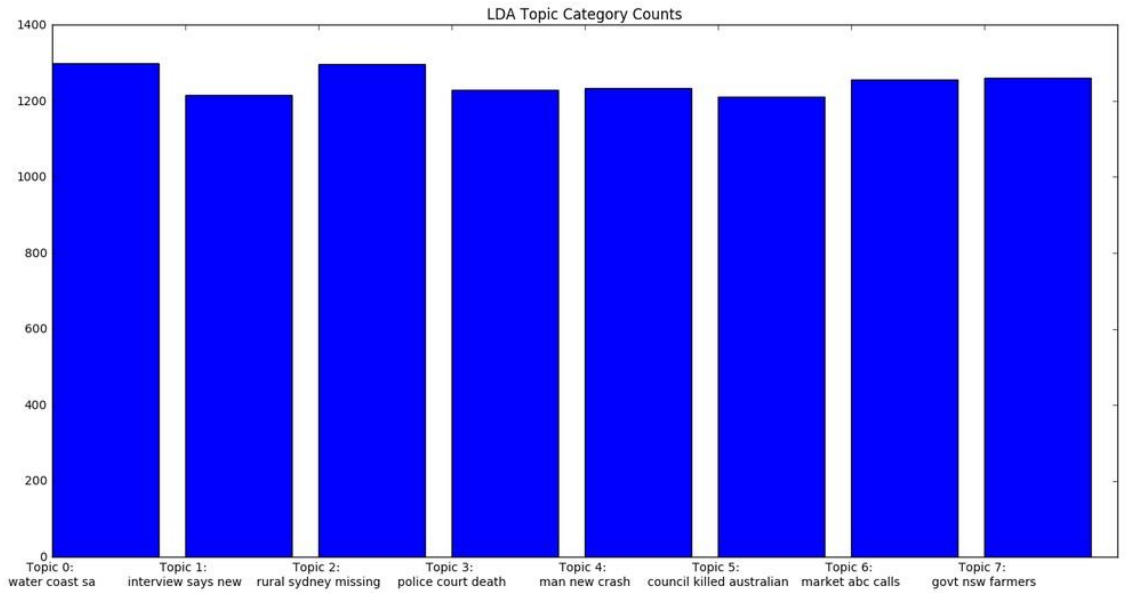
(0, 2979) 1

4.3.2 Σύγκριση απόδοσης LSA και LDA

Αφού πάρουμε ένα δείγμα από το σώμα κειμένων, υλοποιούμε μία μικρή θεματική μοντελοποίηση για 8 θέματα. Μετατρέπουμε το αρχικό δείγμα τίτλων σε μια λίστα με τις προβλεπόμενες κατηγορίες θεμάτων, όπου κάθε κατηγορία χαρακτηρίζεται από τις πιο συχνές λέξεις. Αναπαριστάνουμε τα σχετικά μεγέθη καθεμιάς από αυτές τις κατηγορίες μέσω χρήσης ενός ραβδογράμματος.

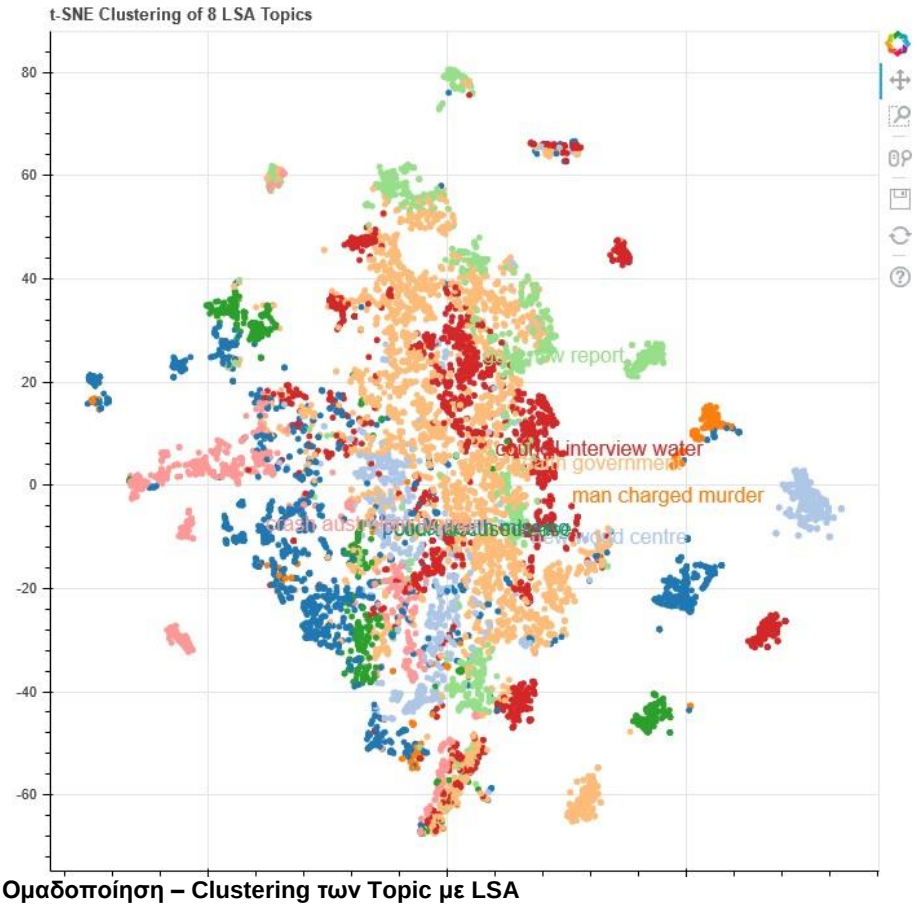


LSA – Μέτρηση ανά topic



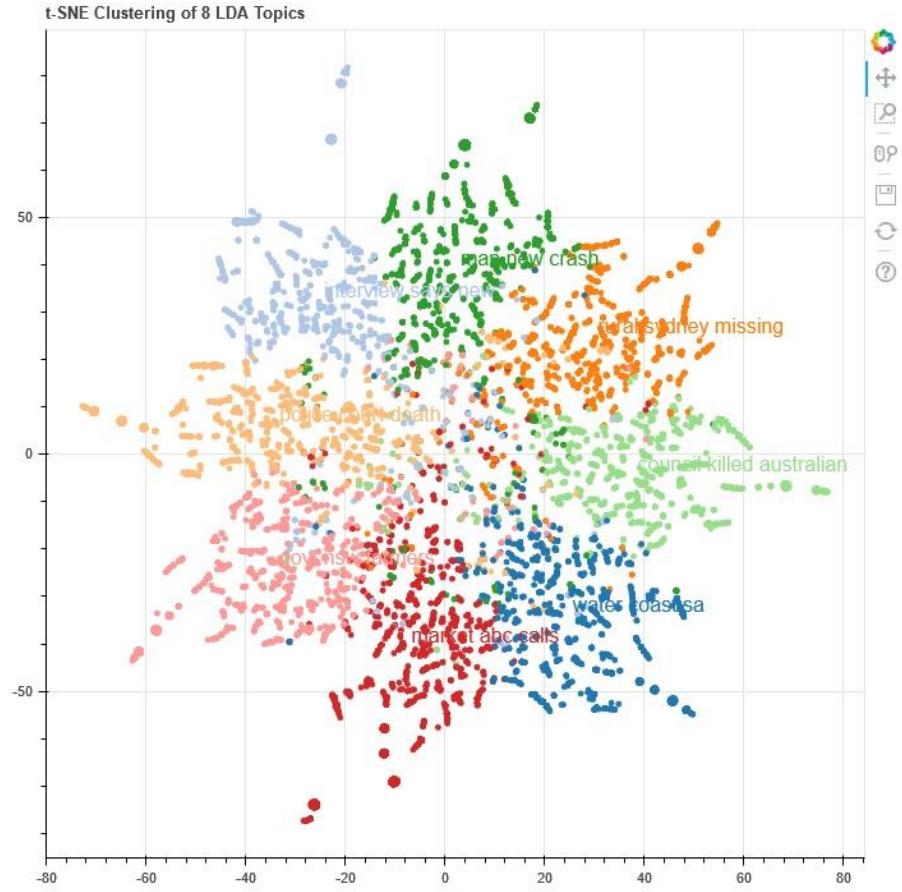
LDA – Μέτρηση ανά topic

Προκειμένου να συγκρίνουμε λίγο καλύτερα τον LSI με τον LDA, χρησιμοποιούμε μια τεχνική μείωσης των διαστάσεων που ονομάζεται t-SNE, και χρησιμεύσει για την καλύτερη προβολή της διαδικασίας ομαδοποίησης. Είναι χρήσιμο να βρεθεί κι η κεντροειδής θέση κάθε θέματος, ώστε να γίνει πιο κατανοητό το αποτέλεσμα.



Ομαδοποίηση – Clustering των Topic με LSA

Θεματική μοντελοποίηση σε σώμα ειδησεογραφικών κειμένων

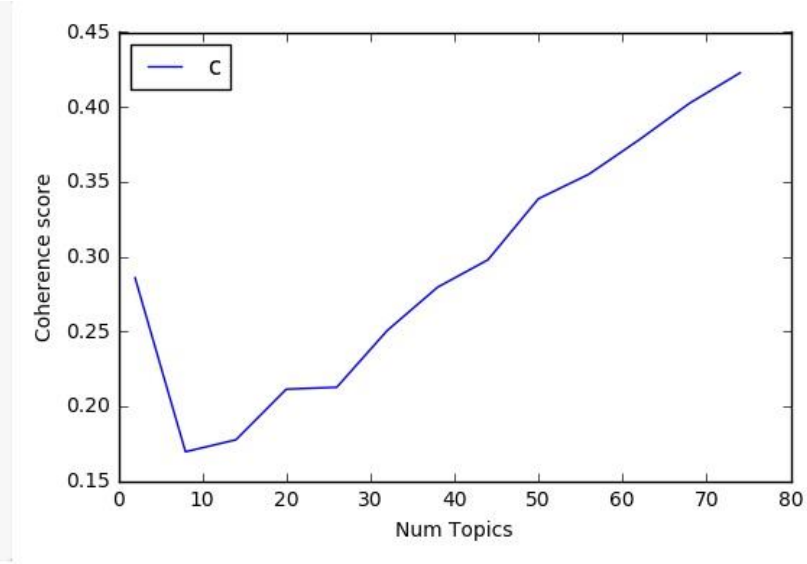


Ομαδοποίηση – Clustering των Topic με LDA

Παρατηρούμε ότι ο LDA κατηγοριοποιεί και ομαδοποιεί με πολύ μεγαλύτερη επιτυχία από ότι ο LSA. Ο LDA θα είναι ο αλγόριθμος που θα χρησιμοποιήσουμε για την διαδικασία του Topic Modelling.

4.3.3 Αριθμός θεμάτων για την καλύτερη απόδοση του LDA

Για την εύρεση του βέλτιστου αριθμού θεμάτων στον αλγόριθμο, ο τρόπος που προτείνεται είναι μέσω του Topic Coherence. Δηλαδή δημιουργούμε πολλαπλά μοντέλα LDA με διαφορετικές τιμές του αριθμού θεμάτων. Τελική επιλογή θα επιλογή θα είναι αυτός που δίνει την υψηλότερη τιμή συνοχής. Σχεδόν πάντα, μία επιλογή μιας ακόμη υψηλότερης τιμής προσφέρει πιο λεπτομερή υπό-θέματα. Στις περιπτώσεις που οι ίδιες λέξεις-κλειδιά επαναλαμβάνονται σε πολλά θέματα, τότε αυτό είναι σημαίνει ότι ο αριθμός των topic είναι πολύ μεγάλος.



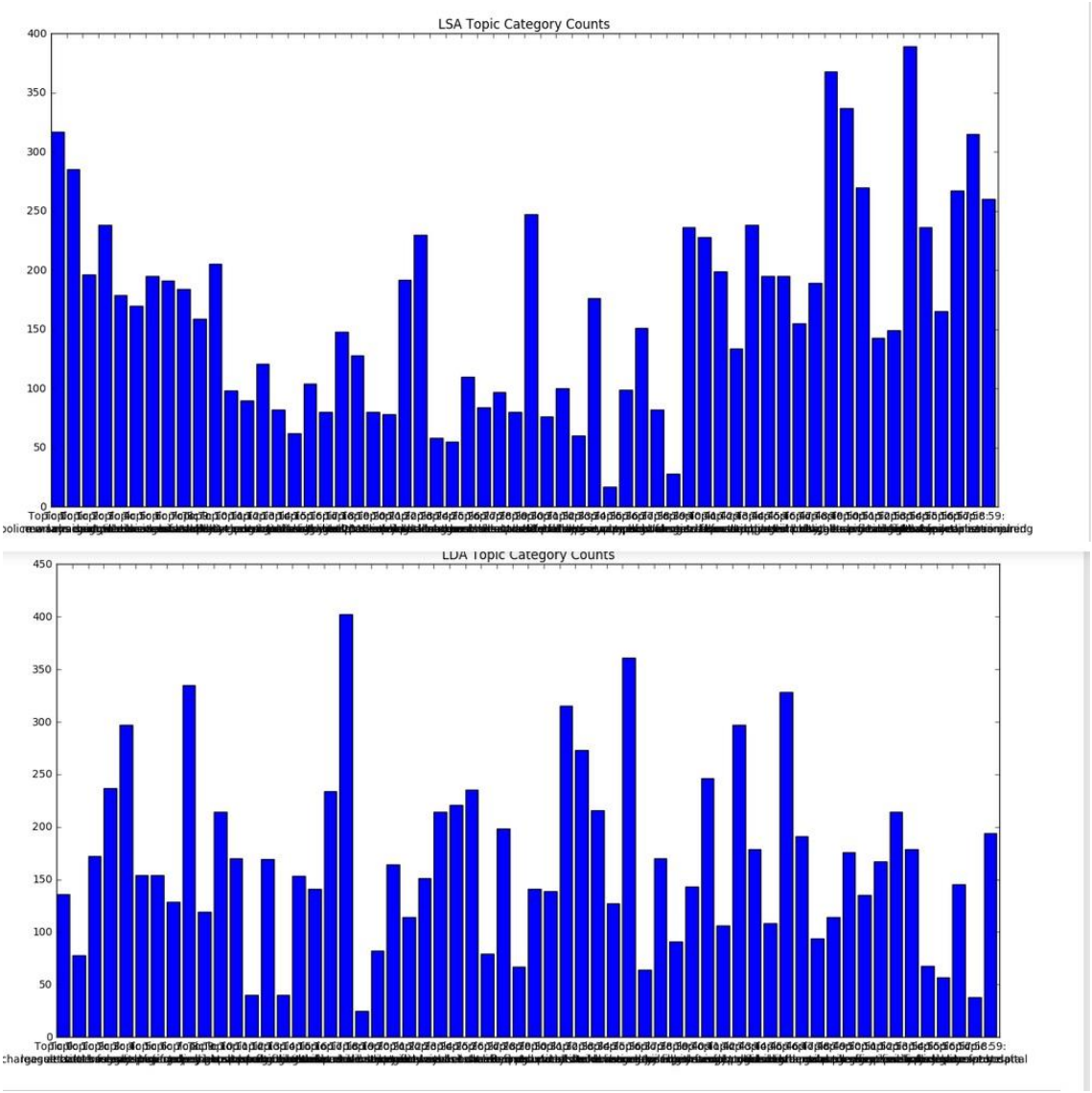
Καλύτερη συνοχή θεμάτων ανάλογα τον αριθμό θεμάτων

```
Num Topics = 2  has Coherence Value of 0.2857
Num Topics = 8  has Coherence Value of 0.1694
Num Topics = 14  has Coherence Value of 0.1774
Num Topics = 20  has Coherence Value of 0.2113
Num Topics = 26  has Coherence Value of 0.2126
Num Topics = 32  has Coherence Value of 0.2505
Num Topics = 38  has Coherence Value of 0.2794
Num Topics = 44  has Coherence Value of 0.2978
Num Topics = 50  has Coherence Value of 0.3386
Num Topics = 56  has Coherence Value of 0.355
Num Topics = 62  has Coherence Value of 0.3781
Num Topics = 68  has Coherence Value of 0.4025
Num Topics = 74  has Coherence Value of 0.4228
```

Απόδοση συνοχής ανά αριθμό topic

4.3.4 Υλοποίηση LDA για 68 topics

Επόμενη εργασία είναι να αναλύσουμε τα δεδομένα μας δημιουργώντας 68 θέματα και να εξηγήσουμε την απόδοση τους. Πρώτο βήμα είναι να ελέγξουμε και να συγκρίνουμε την απόδοση και την λειτουργία των αλγορίθμων LSA και LDA σε 68 topics.



Υλοποιώντας LDA για 68 θέματα, τα θέματα με τις περισσότερο εμφανιζόμενες λέξεις είναι:

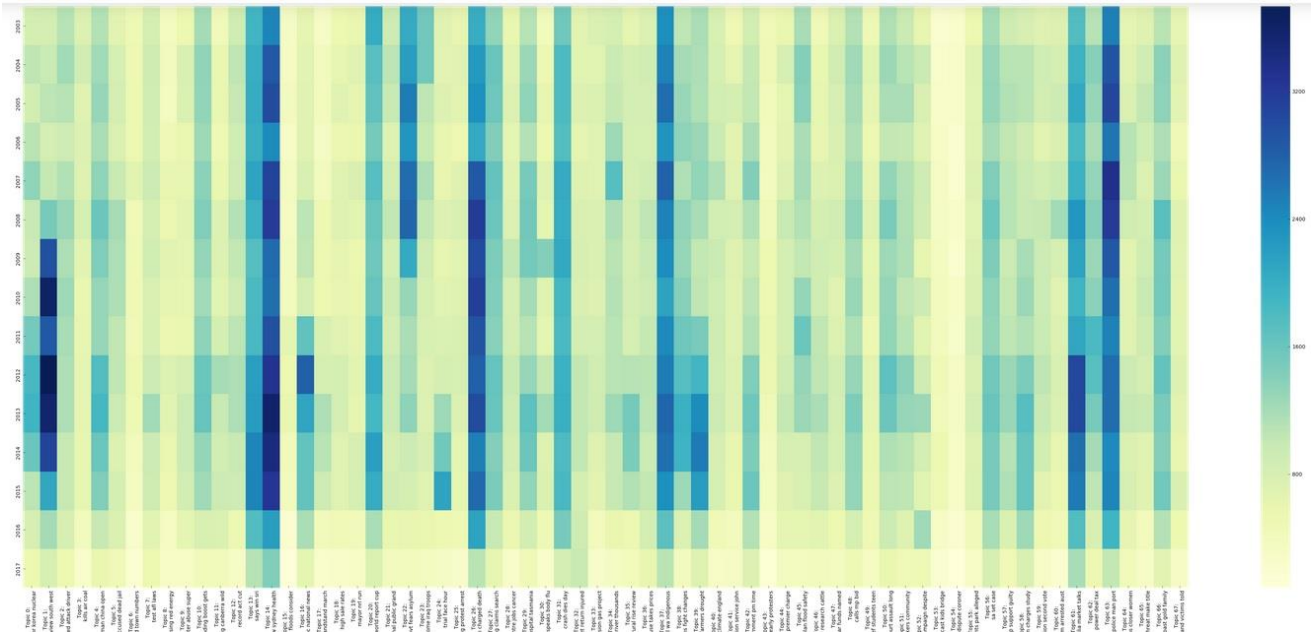
- Topic 0: johnson glory burning tablelands bitten rape canoeist satisfaction griffith tibet
- Topic 1: canberra battle spot raiders suspension crews blaze act firefighters mark
- Topic 2: job cuts rice historic malaysia security discovered upgrade hope miners
- Topic 3: man charged murder jailed death arrested men police mining dies
- Topic 4: ready research cattle turn brown says dollar businesses carnival new
- Topic 5: murder rise baby bail youth rate admits hobart say damage
- Topic 6: gas project decision grandstand drum wednesday jones 18 june homes
- Topic 7: iraq troops howard rebels tim professor kill forced amid auction
- Topic 8: nsw farmers drought act north record dairy aid growers 14
- Topic 9: warns services force cyclone children crops care tasmanian cane larry
- Topic 10: coast gold closer black women step swans play title puts
- Topic 11: guilty pleads wall killer heavy quit gippsland facing murder street

- Topic 12: protest watch issues patients warning centres leaving pocock childcare myanmar
- Topic 13: election campaign closure dogs refuses coalition gillard pull sa 2016
- Topic 14: make greens reef barrier angry bombs changes decision adelaide hoax
- Topic 15: crash death fatal probe accused day near train police plane
- Topic 16: court charges hears told victims poll island liverpool shot results
- Topic 17: woman test afl beach war aboriginal chance dies crash blamed
- Topic 18: market residents news national urged plans adelaide business park rural
- Topic 19: new laws hospital rates senate govt race denies super council
- Topic 20: dead rules tasmania development injured raid court attacked haiti crash
- Topic 21: court search missing claims assault continues man body blaze accused
- Topic 22: win rural union victory review michael reporter obama sri teachers
- Topic 23: hill broken sugar finals bridge mackay ends opener brings rainfall
- Topic 24: australian school ahead open community tigers expected rain lose moves
- Topic 25: says south state weather west victoria australia korea africa north
- Topic 26: council car year hit markets bushfire media city crews considers
- Topic 27: police dies centre port jobs probe students investigate alleged shooting
- Topic 28: health public budget ban defends faces federal preview fly need
- Topic 29: talks industry light green cut makes end close strong stay
- Topic 30: mp minister attacks returns stars boss red stand sheep supply
- Topic 31: report world cup power injury appeal team rugby smith james
- Topic 32: qld perth food aussie hewitt nz sentence million govt rail
- Topic 33: water queensland time years coach finds survey gay chief quits
- Topic 34: fears alice lead springs asylum life seeker detention hurt sparks
- Topic 35: festival tourists 11 folk attract woodford hits sept falls makers
- Topic 36: interview drug china opposition levy apology john michael extended chris
- Topic 37: young bali wallabies wary brother bombings colonial finding nelson pearce
- Topic 38: injured tour vic boat vote bank use australians crisis doubt
- Topic 39: head surgery speed timber elective promise organisers hospital new hails
- Topic 40: fight east premier urges aust japan changes seek timor parliament
- Topic 41: officer authorities victorian rock approval fishers roll revealed reds suspect
- Topic 42: govt wa nt road indigenous season wants family pay land
- Topic 43: final grand federer despite air thai st wont release row
- Topic 44: art israel post gaza office hours disease gallery prize exhibition
- Topic 45: change climate control bomb kids hope break councillor gun promises
- Topic 46: calls bid liberal commission abbott royal real president rio leadership
- Topic 47: australia png stage confirms refugee money mobile rd honour telstra
- Topic 48: government pm house party prices political late drivers offers latham
- Topic 49: killed big attack driver club indian crowd star dam accident
- Topic 50: work workers pakistan india bans rudd councils program emergency consider
- Topic 51: italy axe stoner appalled group diagnosis genias anti displaced named
- Topic 52: pacific speaks host tv beat winery 10 thompson chapman asia
- Topic 53: wins farm security storm wind airport clean sea live downs

- Topic 54: set case sex england hits blues half warriors service new
- Topic 55: river oil seeks good great fuel begin sale leak wine
- Topic 56: funding boost cuts gets brisbane concerns heritage rally long trump
- Topic 57: trial melbourne face rape hour game gang 2015 country tuesday
- Topic 58: plan safety highway flood run virus truck cases nsw disability
- Topic 59: sa abc high sport abuse weather goes journalist hits consumer
- Topic 60: deal strike action tax pay company kills profit tas pressure
- Topic 61: study funds reveals announces sought worries 20m sa code date
- Topic 62: child inquiry violence protection tells education schools regional sentenced porn
- Topic 63: home help free trade support price hopes backs defence milk
- Topic 64: labor podcast ntch win polls says sach nt poll legacy
- Topic 65: group league questions shire line terrorism planning champions green cause
- Topic 66: threat station lost warned buy tasmanias ponting catch forest power
- Topic 67: sydney takes bay leaders future flu debate bird jail mayor

Αν παρατηρήσουμε προσεχτικά τα θέματα που έχουν δημιουργηθεί, θα παρατηρήσουμε ότι τα θέματα έχουν πολλές κοινές λέξεις. Αυτό μας δημιουργεί επιπλέον προβλήματα αφού στην κατηγοριοποίηση τους θα έχουμε πολλές επικαλύψεις στα θέματα.

Αν βρούμε για τόσα θέματα την εξέλιξη τους στο χρόνο θα βρούμε πολλές μηδενικές ή κοντά στο μηδέν τιμές.



Η εξήγηση αυτών των αποτελεσμάτων είναι απόλυτα λογική αν αναλογιστούμε ότι εργαζόμαστε με τίτλους ειδησεογραφικών νέων. Οπότε πρέπει πάντα να έχουμε υπόψιν μας δύο σημαντικά στοιχεία:

- Το σώμα των κειμένων περιλαμβάνει τον τίτλο της είδησης μόνο κι όχι ολόκληρο το άρθρο. Αυτό σημαίνει μικρό όγκο σε λέξεις (ένας τίτλος περιέχει από 2 έως 10 λεξεις το μέγιστο).
- Τα δεδομένα είναι τίτλοι ειδήσεων. Αυτό εξηγεί την ανοδική διακύμανση, όσο αυξάνεται ο αριθμός των θεμάτων. Ένας τίτλος είναι κάτι το γενικό. Αυτό σημαίνει ότι θα

αποκτήσει συνοχή μεγαλύτερη, όταν δημιουργηθούν πολλά υποθέματα που θα μπορεί να τοποθετηθεί.

Οι δύο παραπάνω παρατηρήσεις αποτελούν την απάντηση για την απεικόνιση αυτών των αποτελεσμάτων.

4.3.5 LDA Topic Modelling σε όλο το σώμα κειμένων

Υλοποιούμε μοντελοποίηση των τίτλων ειδήσεων σε 20 συνολικά θέματα. Αν και η συνοχή είναι χαμηλή, πετυχαίνουμε αρκετά καλή ομαδοποίηση.
Αρχικά, αναθέτουμε ένα θέμα σε κάθε ένα έγγραφο.

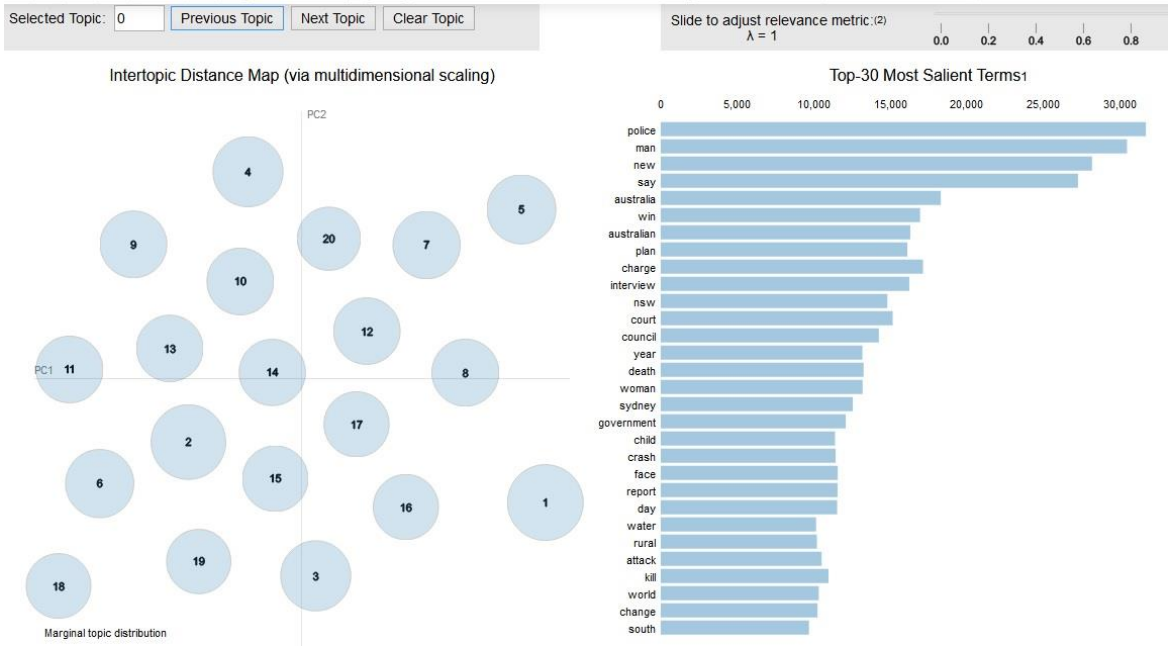
	Topic0	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	Topic10	Topic11	Topic12	Topic13	Topic14	Topic15	Topic16	Topic17	Topic18	Topic19	dominant_topic
Doc0	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.84	19
Doc1	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.81	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	11
Doc2	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.76	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	9
Doc3	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.29	0.01	0.01	0.15	0.01	0.44	0.01	0.01	0.01	0.01	0.01	0.01	13
Doc4	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.84	0.01	0.01	0.01	0.01	0.01	0.01	13
Doc5	0.01	0.61	0.01	0.01	0.01	0.01	0.01	0.01	0.21	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	1
Doc6	0.01	0.01	0.01	0.01	0.01	0.01	0.18	0.01	0.01	0.01	0.67	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	10
Doc7	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.84	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	7
Doc8	0.17	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.68	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	8
Doc9	0.01	0.67	0.01	0.01	0.01	0.01	0.01	0.18	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	1
Doc10	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.61	0.21	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	7
Doc11	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.34	0.01	0.01	0.01	0.01	0.01	0.01	0.51	0.01	0.01	17
Doc12	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.51	0.26	0.01	17
Doc13	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.84	0.01	0.01	0.01	0.01	0.01	14
Doc14	0.01	0.15	0.01	0.01	0.01	0.01	0.15	0.01	0.01	0.01	0.15	0.01	0.01	0.01	0.15	0.01	0.01	0.15	0.15	0.01	1

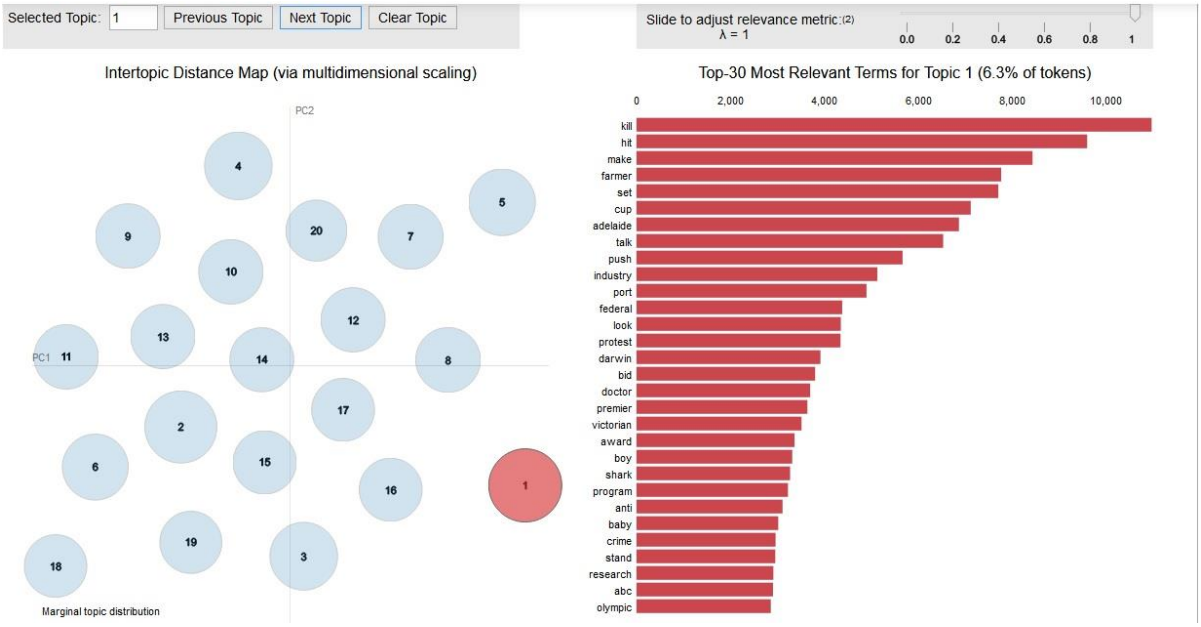
Ανάθεση εγγράφου σε Topic

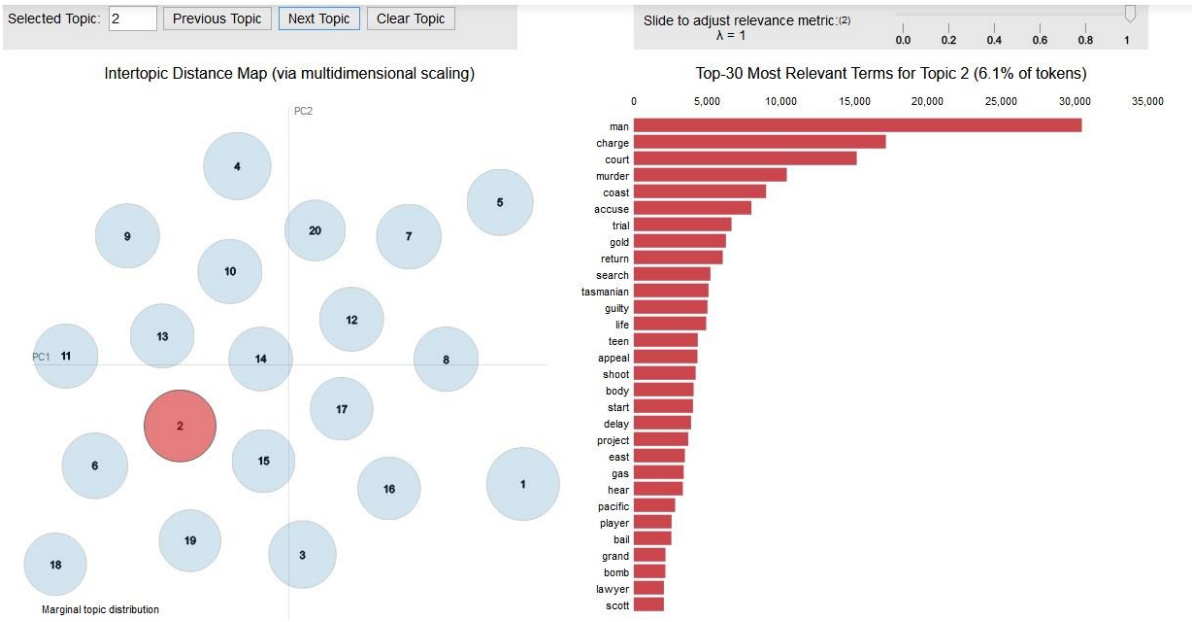
	Word 0	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10	Word 11	Word 12	Word 13	Word 14
Topic 0	council	day	world	change	country	claim	centre	funding	resident	farm	local	reject	welcome	red	risk
Topic 1	police	face	report	lose	assault	use	abuse	hunter	war	station	offer	chief	company	fail	nrn
Topic 2	crash	drug	final	trump	island	head	probe	weather	fatal	launch	rescue	injure	cattle	link	turn
Topic 3	government	home	die	hospital	cut	hour	worker	labor	concern	rate	dog	bank	young	plead	newcastle
Topic 4	kill	hit	make	farmer	set	cup	adelaide	talk	push	industry	port	federal	look	protest	darwin
Topic 5	south	qld	open	melbourne	china	arrest	leave	defend	city	canberra	abbott	star	chinese	central	super
Topic 6	water	fall	driver	dead	good	need	break	hill	victoria	meet	rain	safety	regional	spark	climate
Topic 7	australia	woman	miss	family	public	question	game	aussie	sign	investigate	liberal	president	dollar	join	warning
Topic 8	win	car	house	lead	brisbane	pay	sale	action	land	study	consider	target	play	security	battle
Topic 9	market	deal	fight	share	law	end	close	tasmania	tax	people	season	nrl	week	trade	officer
Topic 10	child	govt	record	boost	release	deny	png	food	damage	number	mother	rio	way	ready	suicide
Topic 11	new	interview	attack	north	rise	warn	west	business	group	act	mayor	tour	threat	housing	turnbull
Topic 12	man	charge	court	murder	coast	accuse	trial	gold	return	search	tasmanian	guilty	life	teen	appeal
Topic 13	australian	nsw	election	ban	fear	flood	sex	live	hold	strike	party	free	drought	air	club
Topic 14	sydney	help	test	big	green	support	tell	victim	leader	afl	league	medium	hope	train	opposition
Topic 15	national	state	perth	work	seek	park	future	speak	vote	river	announce	murray	steal	grandstand	blame
Topic 16	say	health	minister	budget	news	time	want	beat	cost	coal	india	decision	hobart	loss	peter
Topic 17	plan	death	queensland	urge	job	continue	union	student	review	inquiry	storm	second	ahead	bushfire	reveal
Topic 18	rural	high	price	road	power	indigenous	fund	run	mining	campaign	race	great	drum	right	aboriginal
Topic 19	year	jail	school	service	case	community	force	sentence	rule	asylum	allege	podcast	low	demand	sport

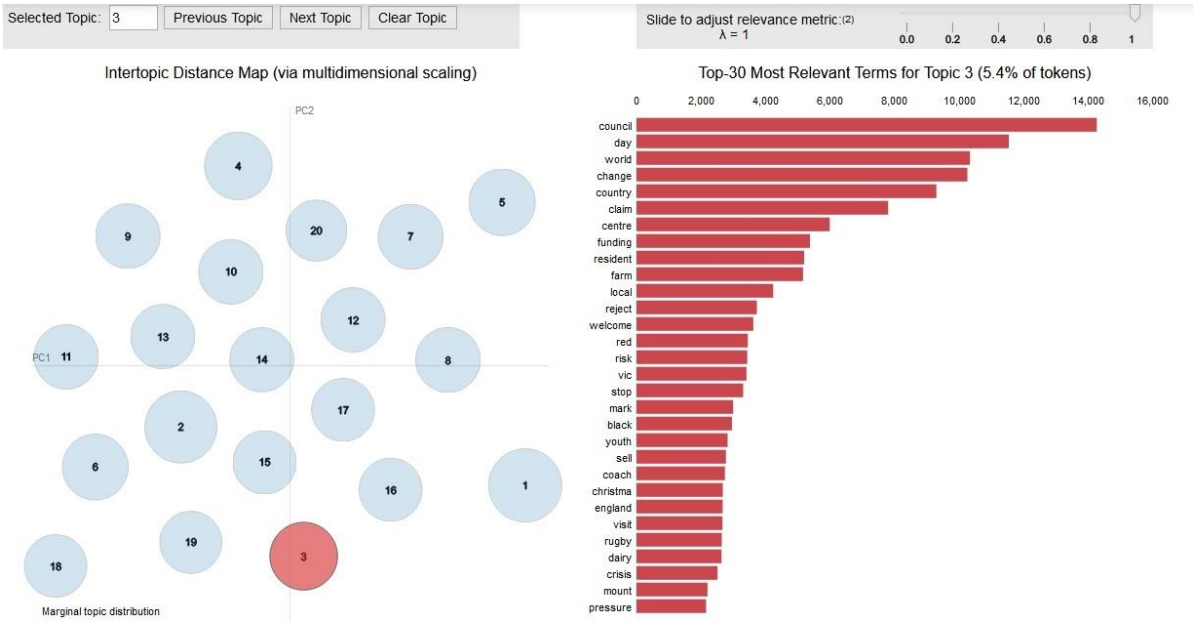
Λέξεις που περιγράφουν το κάθε topic

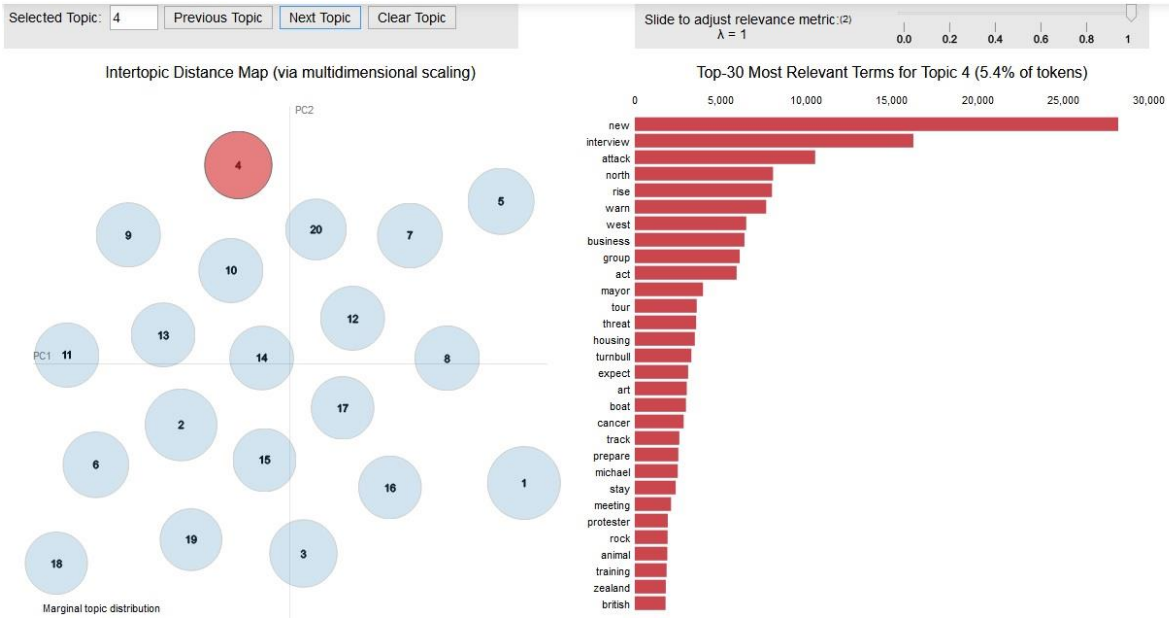
4.3.6 Απεικόνιση των Topic και παρουσίαση του βάρους της κάθε λέξης

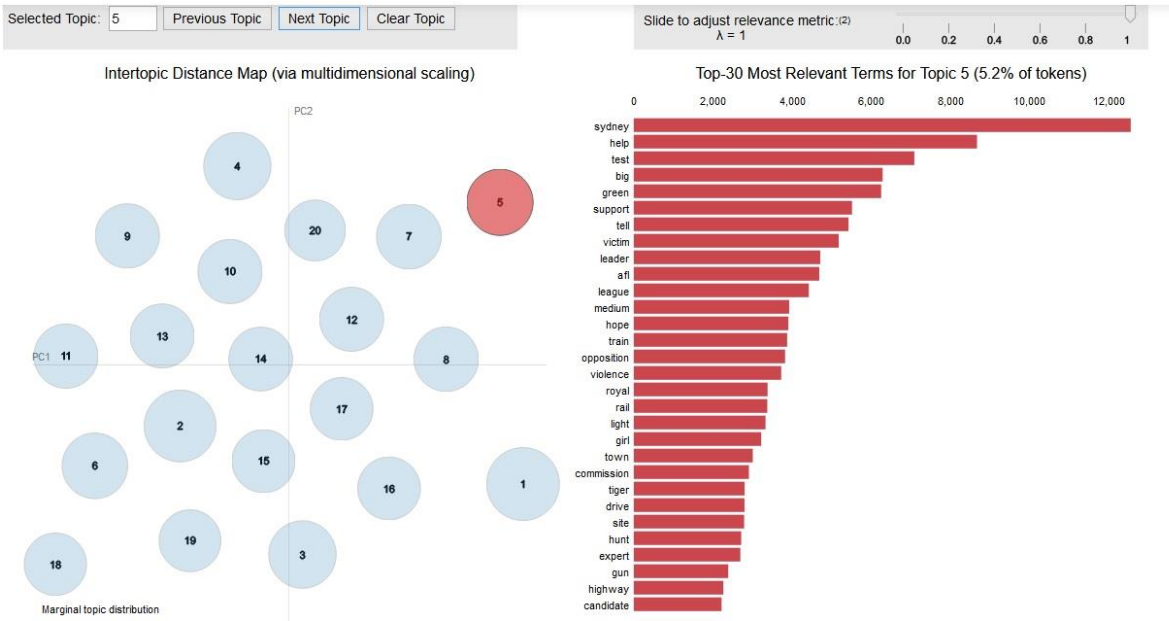


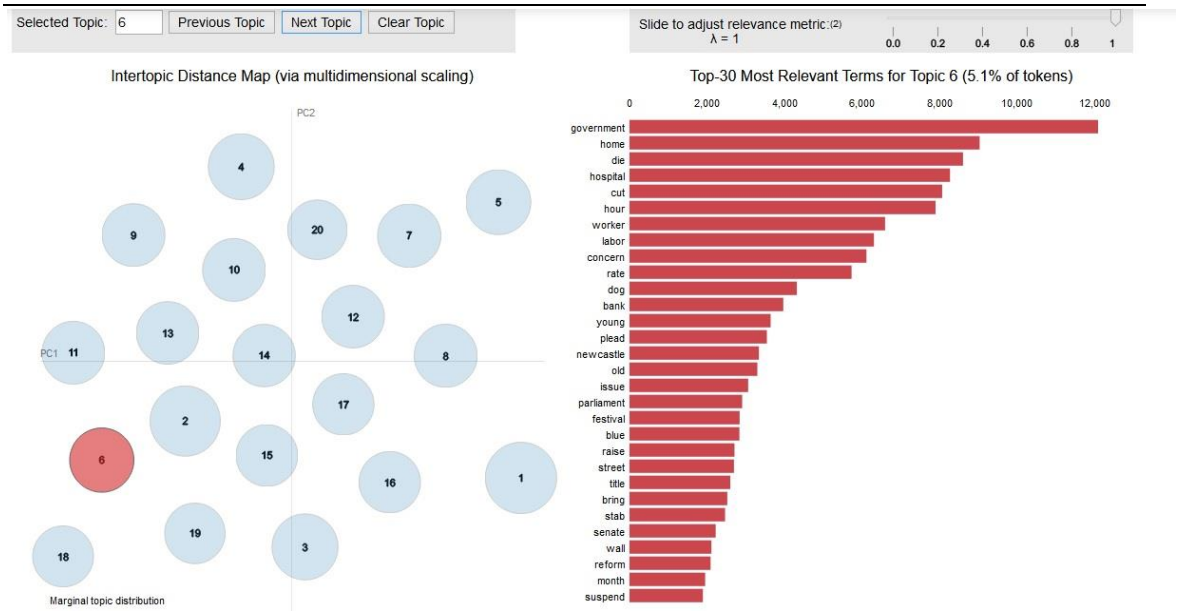


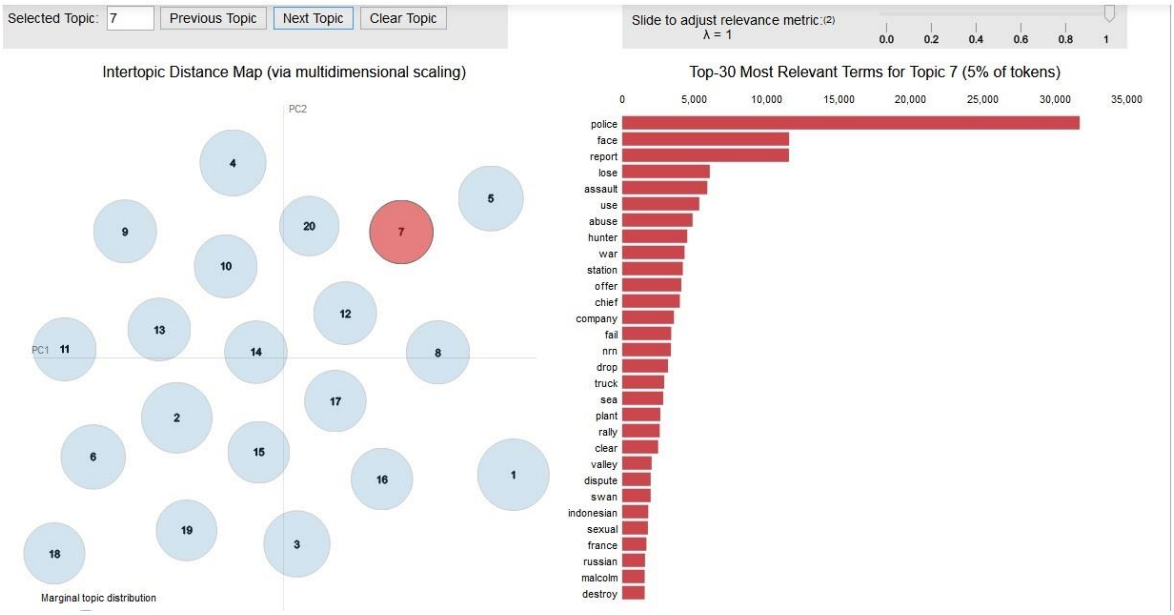

Selected Topic: 2
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0 0.2 0.4 0.6 0.8 1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
Marginal topic distribution
Top-30 Most Relevant Terms for Topic 2 (6.1% of tokens)
0 5,000 10,000 15,000 20,000 25,000 30,000 35,000
man
charge
court
murder
coast
accuse
trial
gold
return
search
tasmanian
guilty
life
teen
appeal
shoot
body
start
delay
project
east
gas
hear
pacific
player
bail
grand
bomb
lawyer
scott

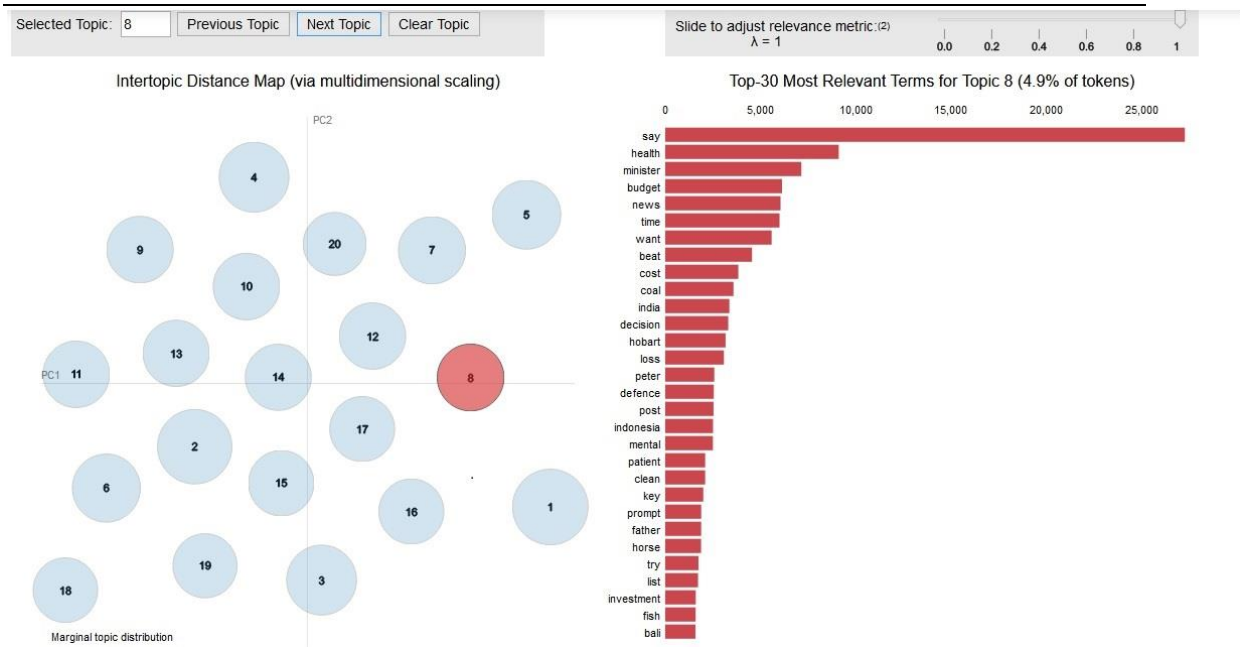

Selected Topic: 3
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0 0.2 0.4 0.6 0.8 1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
Marginal topic distribution
Top-30 Most Relevant Terms for Topic 3 (5.4% of tokens)
0 2,000 4,000 6,000 8,000 10,000 12,000 14,000 16,000
council
day
world
change
country
claim
centre
funding
resident
farm
local
reject
welcome
red
risk
vic
stop
mark
black
youth
sell
coach
christma
england
visit
rugby
dairy
crisis
mount
pressure

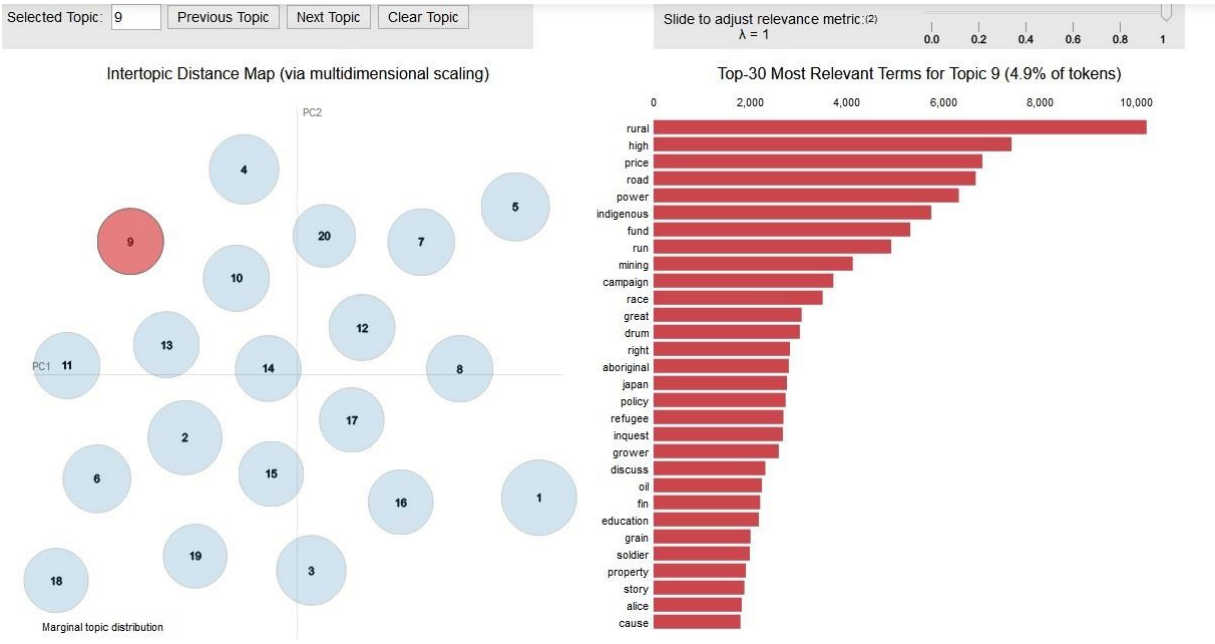

Selected Topic: 4
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0 0.2 0.4 0.6 0.8 1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
4
5
9
20
7
10
12
13
11
14
8
17
2
6
15
16
1
19
3
18
Marginal topic distribution
Top-30 Most Relevant Terms for Topic 4 (5.4% of tokens)
0 5,000 10,000 15,000 20,000 25,000 30,000
new
interview
attack
north
rise
warn
west
business
group
act
mayor
tour
threat
housing
turnbull
expect
art
boat
cancer
track
prepare
michael
stay
meeting
protester
rock
animal
training
zealand
british

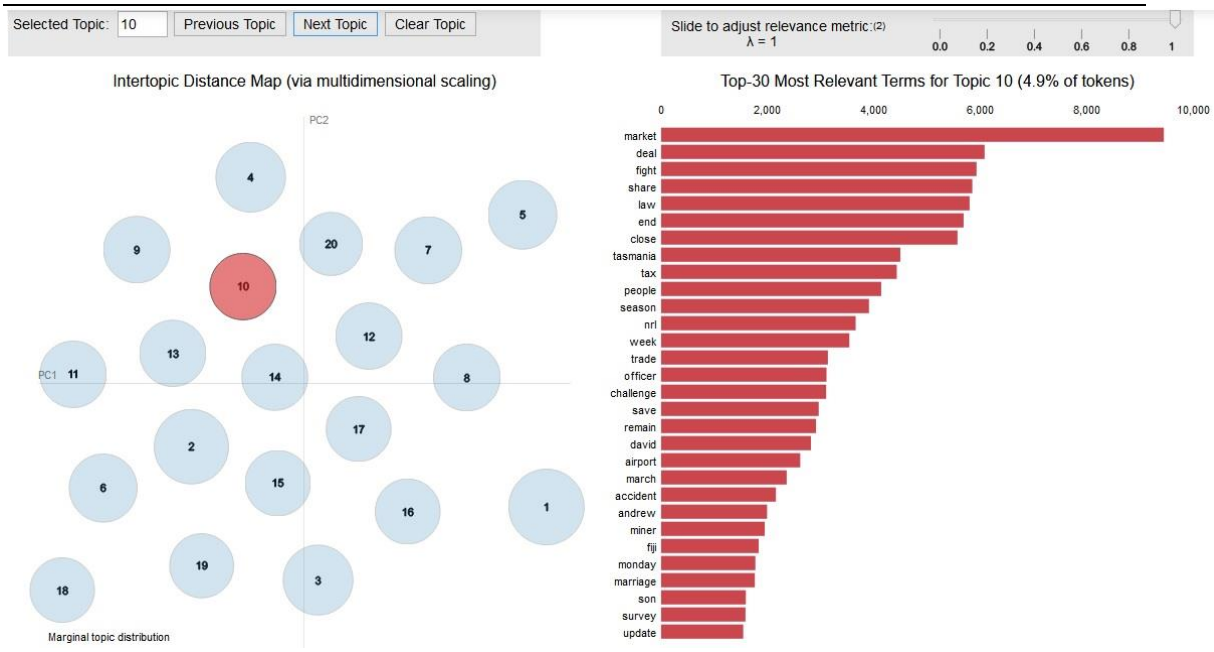

Selected Topic: 5
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0 0.2 0.4 0.6 0.8 1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
4
5
9
20
7
10
12
13
11
14
8
17
2
6
15
16
1
19
3
18
Marginal topic distribution
Top-30 Most Relevant Terms for Topic 5 (5.2% of tokens)
0 2,000 4,000 6,000 8,000 10,000 12,000
sydney
help
test
big
green
support
tell
victim
leader
afl
league
medium
hope
train
opposition
violence
royal
rail
light
girl
town
commission
tiger
drive
site
hunt
expert
gun
highway
candidate

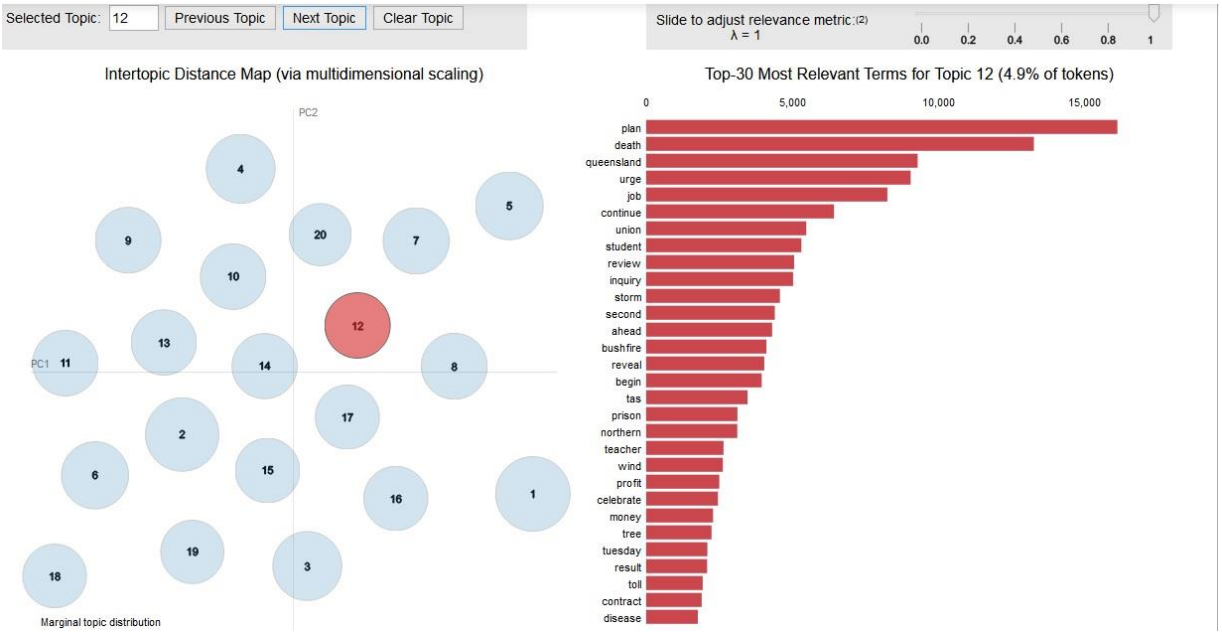

Selected Topic: 6
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0 0.2 0.4 0.6 0.8 1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
Marginal topic distribution
Top-30 Most Relevant Terms for Topic 6 (5.1% of tokens)
0 2,000 4,000 6,000 8,000 10,000 12,000
government
home
die
hospital
cut
hour
worker
labor
concern
rate
dog
bank
young
plead
newcastle
old
issue
parliament
festival
blue
raise
street
title
bring
stab
senate
wall
reform
month
suspend

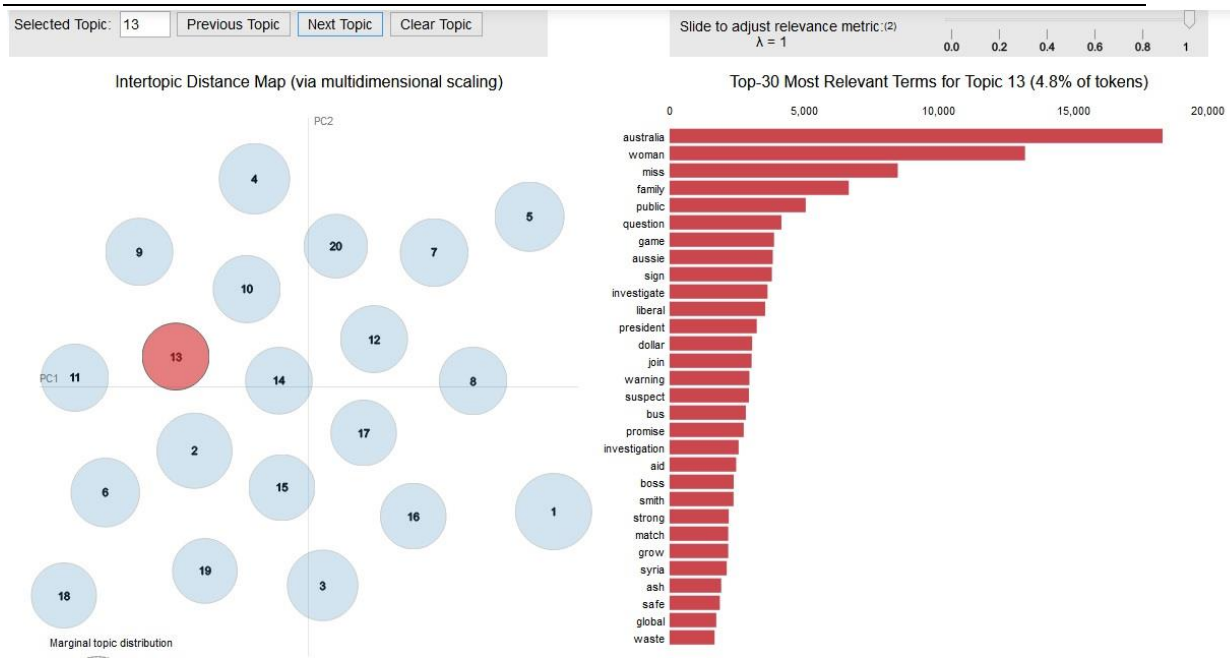

Selected Topic: 7
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0 0.2 0.4 0.6 0.8 1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
Marginal topic distribution
Top-30 Most Relevant Terms for Topic 7 (5% of tokens)
0 5,000 10,000 15,000 20,000 25,000 30,000 35,000
police
face
report
lose
assault
use
abuse
hunter
war
station
offer
chief
company
fail
nrn
drop
truck
sea
plant
rally
clear
valley
dispute
swan
indonesian
sexual
france
russian
malcolm
destroy

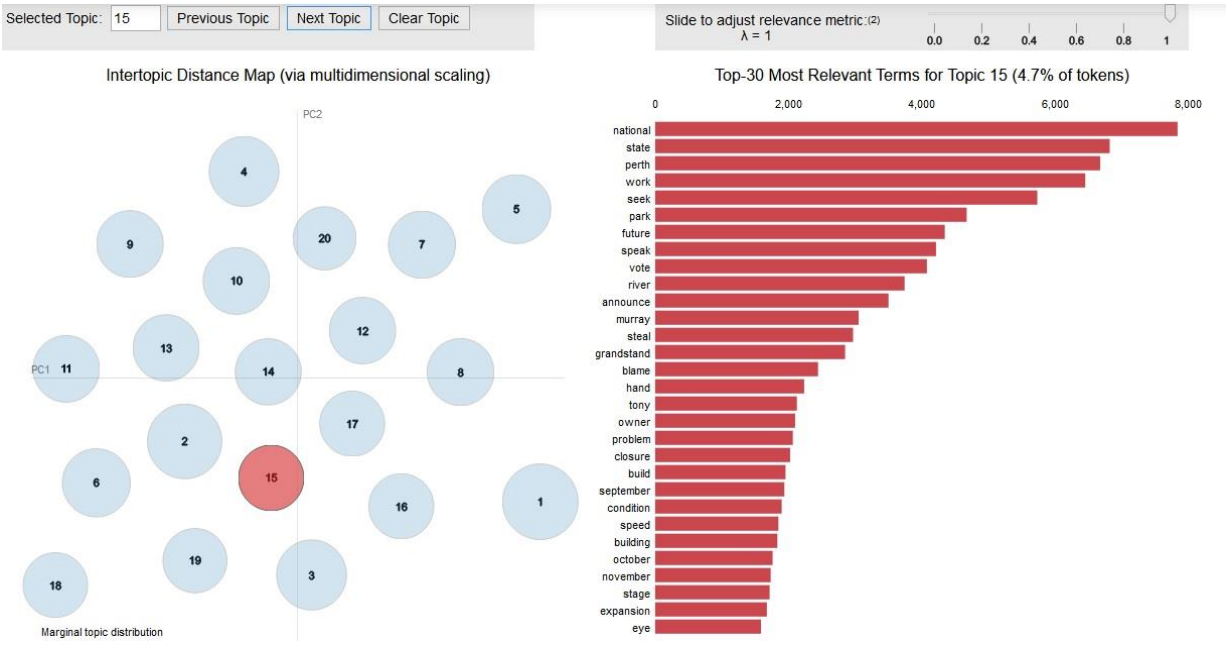

Selected Topic: 8
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0 0.2 0.4 0.6 0.8 1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
Marginal topic distribution
Top-30 Most Relevant Terms for Topic 8 (4.9% of tokens)
0 5,000 10,000 15,000 20,000 25,000
say
health
minister
budget
news
time
want
beat
cost
coal
india
decision
hobart
loss
peter
defence
post
indonesia
mental
patient
clean
key
prompt
father
horse
try
list
investment
fish
bali

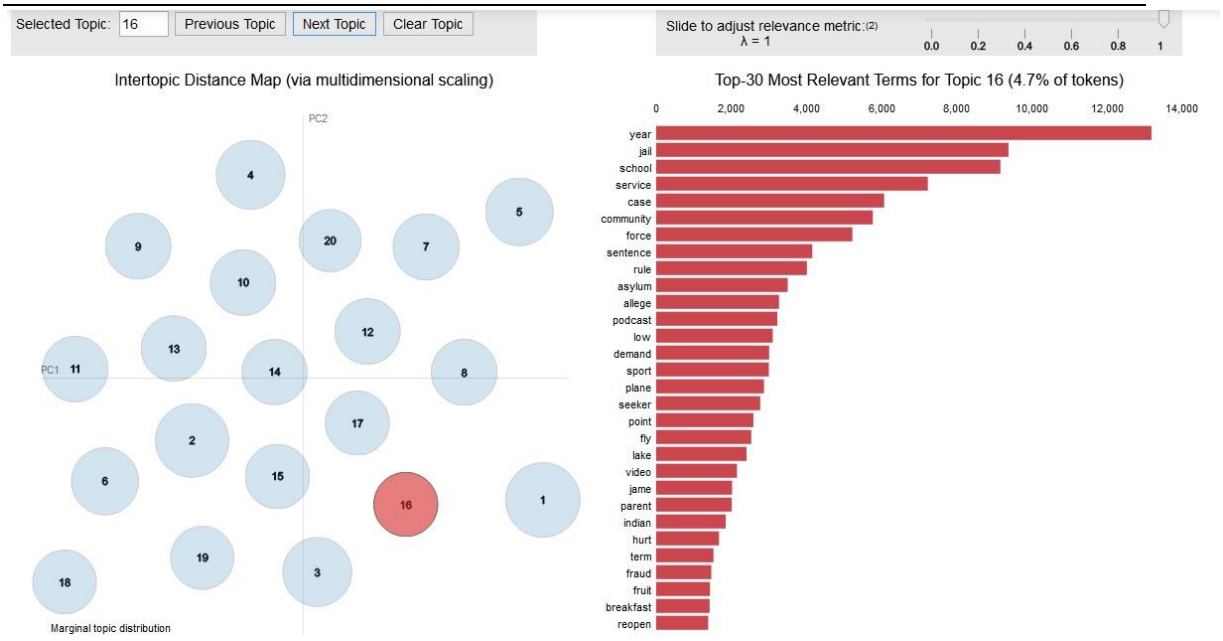

Selected Topic: 9
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0 0.2 0.4 0.6 0.8 1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
Marginal topic distribution
Top-30 Most Relevant Terms for Topic 9 (4.9% of tokens)
0 2,000 4,000 6,000 8,000 10,000
rural
high
price
road
power
indigenous
fund
run
mining
campaign
race
great
drum
right
aboriginal
japan
policy
refugee
inquest
grower
discuss
oil
fin
education
grain
soldier
property
story
alice
cause

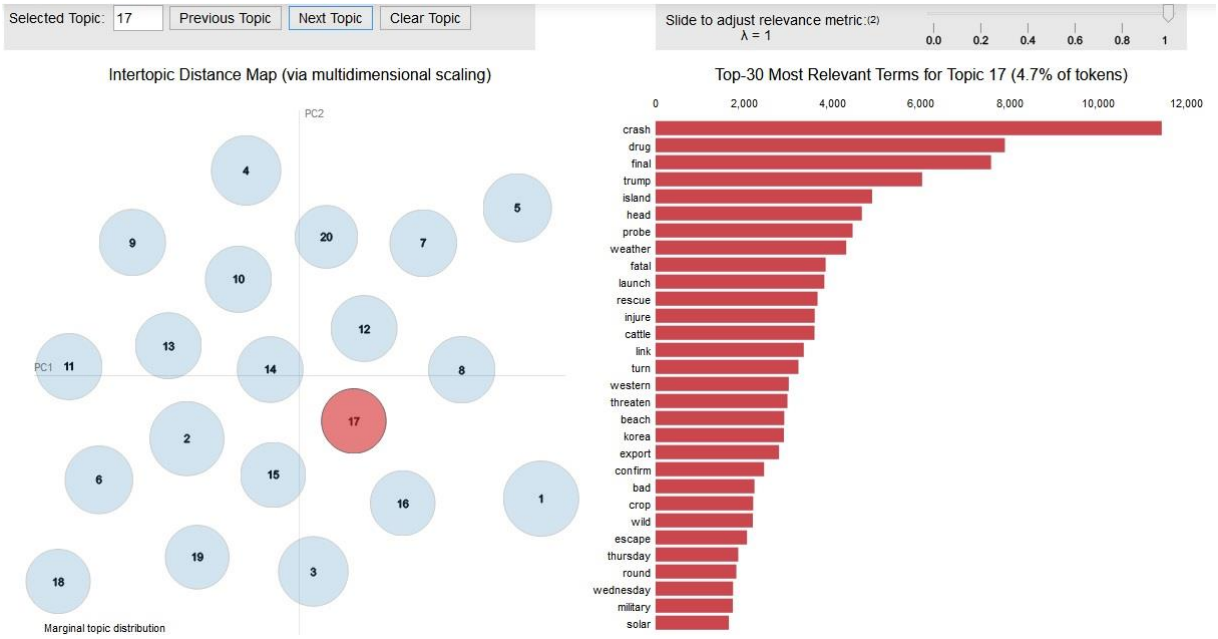

Selected Topic: 10
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0 0.2 0.4 0.6 0.8 1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
Marginal topic distribution
Top-30 Most Relevant Terms for Topic 10 (4.9% of tokens)
0 2,000 4,000 6,000 8,000 10,000
market
deal
fight
share
law
end
close
tasmania
tax
people
season
nrl
week
trade
officer
challenge
save
remain
david
airport
march
accident
andrew
miner
fiji
monday
marriage
son
survey
update

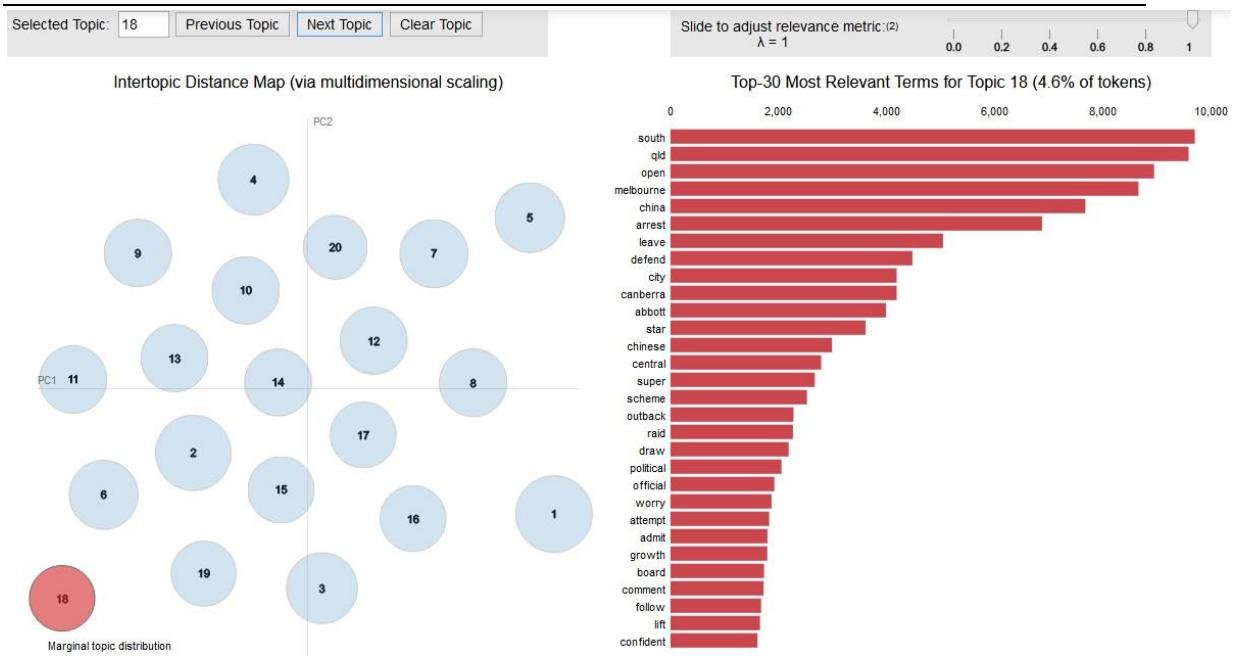

Selected Topic: 12
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0 0.2 0.4 0.6 0.8 1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
Marginal topic distribution
Top-30 Most Relevant Terms for Topic 12 (4.9% of tokens)
0 5,000 10,000 15,000
plan
death
queensland
urge
job
continue
union
student
review
inquiry
storm
second
ahead
bushfire
reveal
begin
tas
prison
northern
teacher
wind
profit
celebrate
money
tree
tuesday
result
toll
contract
disease

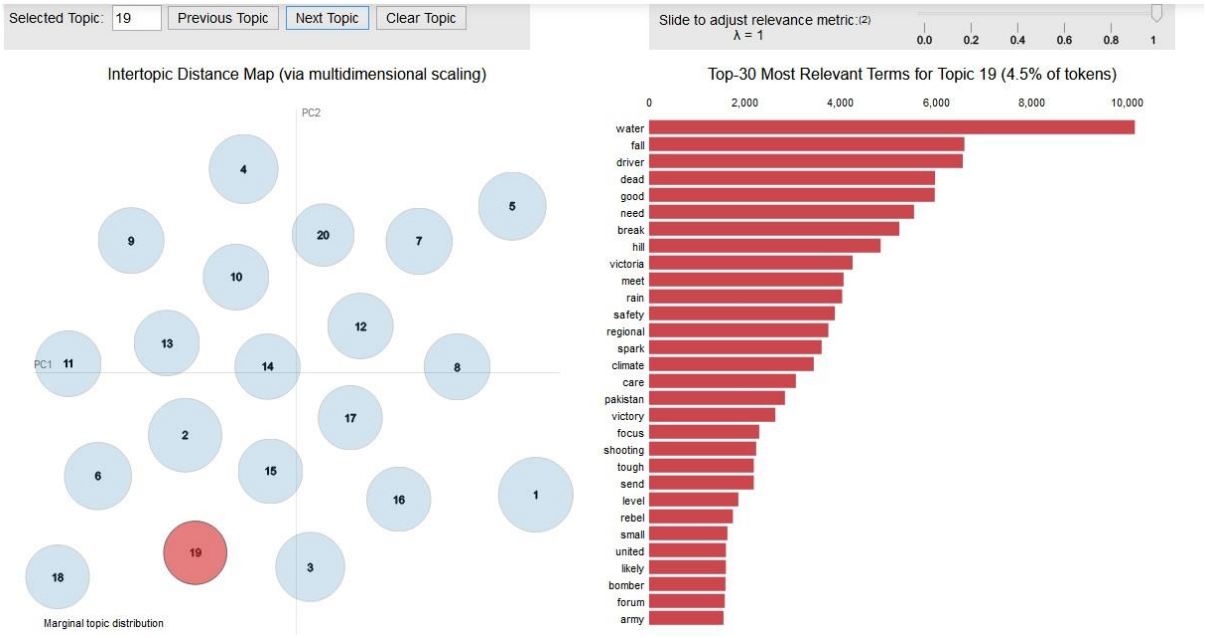

Selected Topic: 13
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0 0.2 0.4 0.6 0.8 1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
Marginal topic distribution
Top-30 Most Relevant Terms for Topic 13 (4.8% of tokens)
0 5,000 10,000 15,000 20,000
australia
woman
miss
family
public
question
game
aussie
sign
investigate
liberal
president
dollar
join
warning
suspect
bus
promise
investigation
aid
boss
smith
strong
match
grow
syria
ash
safe
global
waste


Selected Topic: 15
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0 0.2 0.4 0.6 0.8 1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
Marginal topic distribution
Top-30 Most Relevant Terms for Topic 15 (4.7% of tokens)
0 2,000 4,000 6,000 8,000
national
state
perth
work
seek
park
future
speak
vote
river
announce
murray
steal
grandstand
blame
hand
tony
owner
problem
closure
build
september
condition
speed
building
october
november
stage
expansion
eye


Selected Topic: 16
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0 0.2 0.4 0.6 0.8 1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
Marginal topic distribution
Top-30 Most Relevant Terms for Topic 16 (4.7% of tokens)
0 2,000 4,000 6,000 8,000 10,000 12,000 14,000
year
jail
school
service
case
community
force
sentence
rule
asylum
allege
podcast
low
demand
sport
plane
seeker
point
fly
lake
video
jame
parent
indian
hurt
term
fraud
fruit
breakfast
reopen


Selected Topic: 17
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0 0.2 0.4 0.6 0.8 1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
Marginal topic distribution
Top-30 Most Relevant Terms for Topic 17 (4.7% of tokens)
0 2,000 4,000 6,000 8,000 10,000 12,000
crash
drug
final
trump
island
head
probe
weather
fatal
launch
rescue
injure
cattle
link
turn
western
threaten
beach
korea
export
confirm
bad
crop
wild
escape
thursday
round
wednesday
military
solar

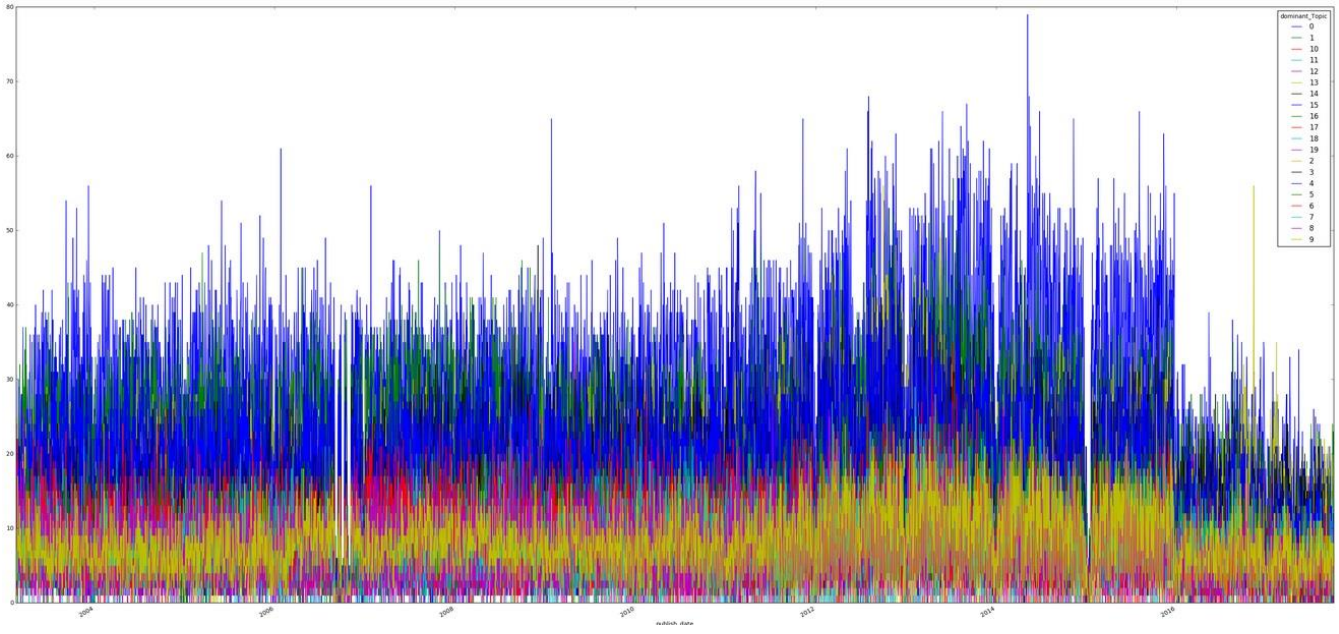




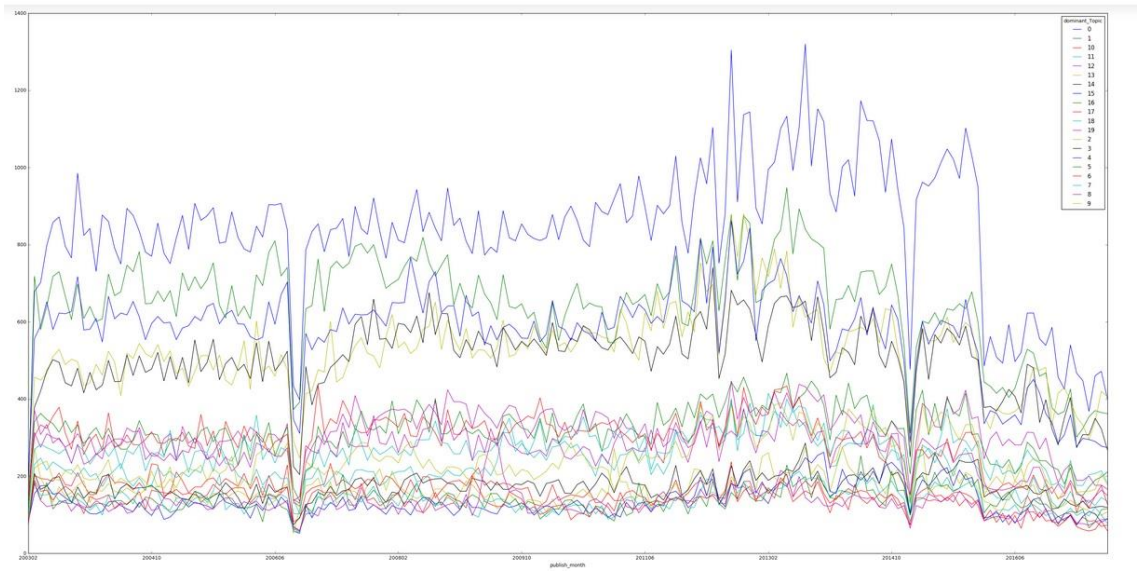
Απεικόνιση των Topic στον χώρο και το βάρος της κάθε λέξης του

	Topic Num	Num Documents
0	0	151968
1	1	113586
2	4	101401
3	2	92812
4	3	88616
5	5	58060
6	12	56204
7	6	51170
8	7	48705
9	8	44162
10	9	42137
11	11	39940
12	10	29935
13	14	28995
14	13	26953
15	16	25001
16	15	24936
17	18	23703
18	17	22685
19	19	22312

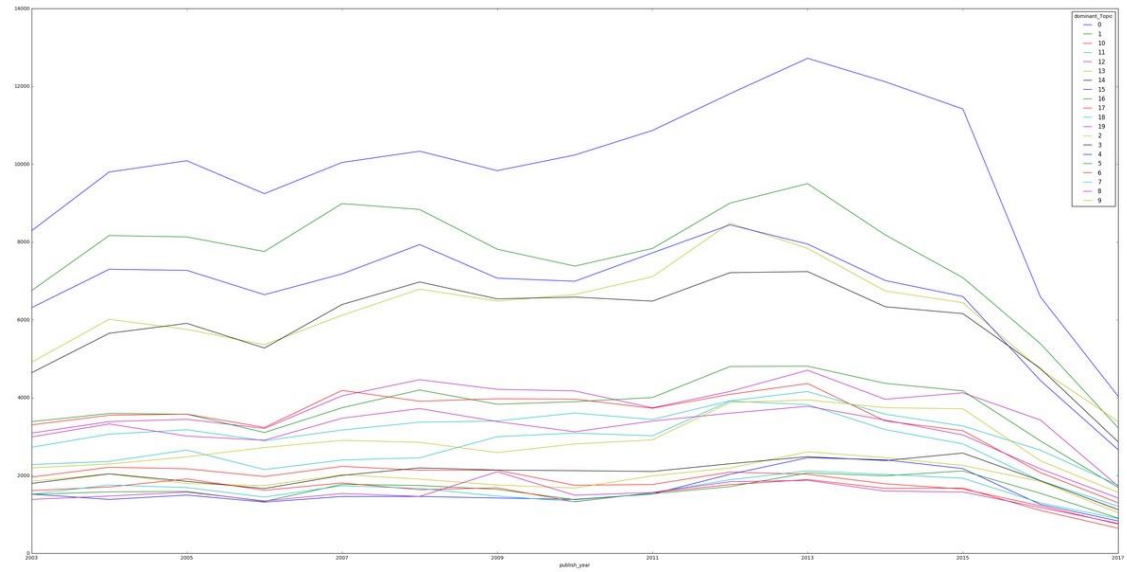
Συνολικά έγγραφα ανά topic



Παρουσίαση των Topic σε ημερήσια βάση



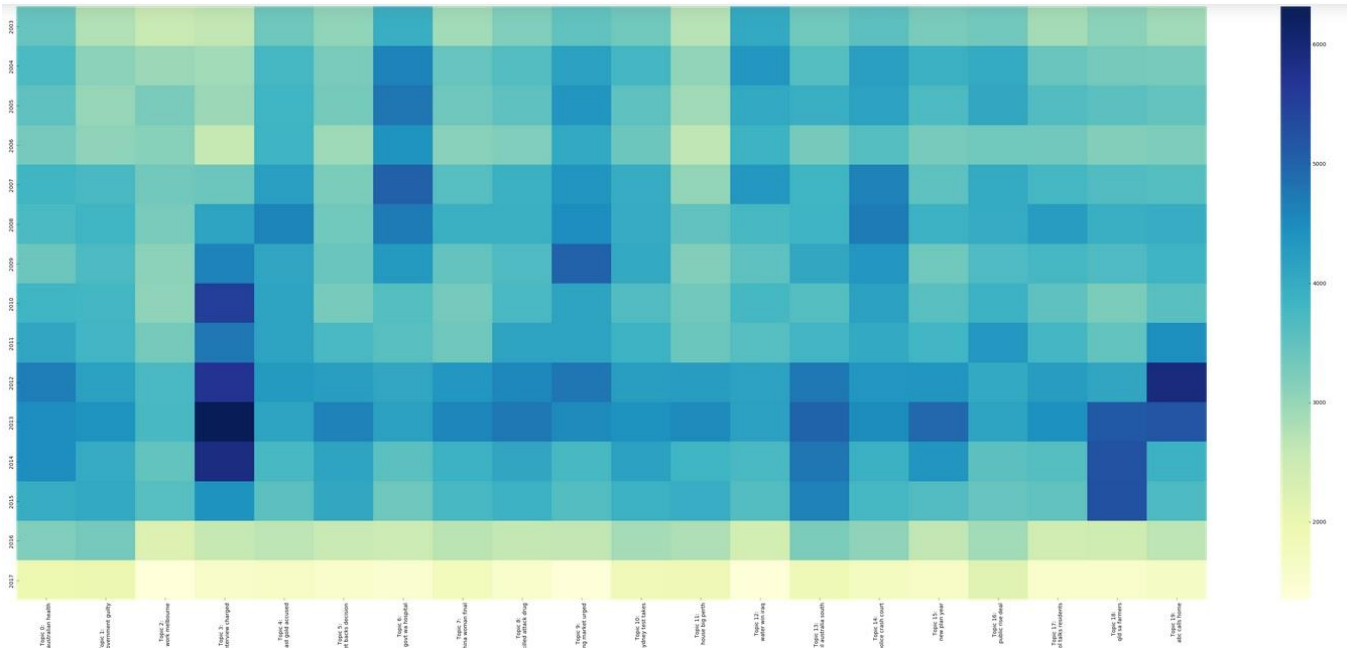
Παρουσίαση των Topic σε μηνιαία βάση



Παρουσίαση των Topic σε ετήσια βάση

4.3.7 Topic Over Time

Το επόμενο βήμα είναι να παρουσιάσουμε την εξέλιξη των Topic στο πέρασμα των χρόνων. Η διαφορά με το προηγούμενο μοντέλο είναι ότι μετράγαμε τον όγκο των Topic ανά χρονικά διαστήματα. Οι ποσότητες αυτές είναι μεταξύ τους ανάλογες με το συνολικό όγκο των δεδομένων.

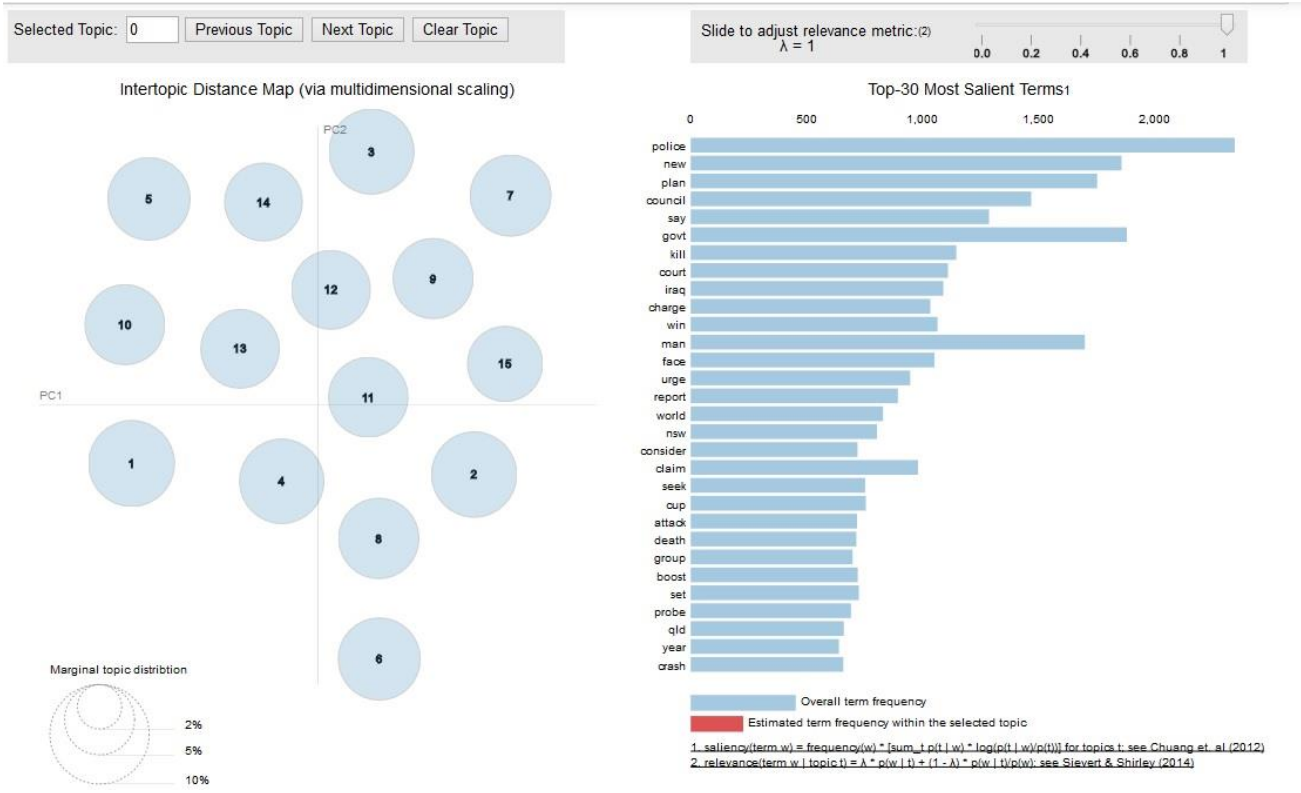


4.3.8 Ανάθεση νέων τίτλων ειδησεογραφικών ειδήσεων

Στόχος μας είναι να δούμε πως λειτουργεί ο αλγόριθμος μας αν του προσθέσουμε νέα εγγραφή ειδησεογραφικού τίτλου. Ακολουθούμε την διαδικασία του 4.3.7 και υλοποιούμε Topic Modelling στα δεδομένα του πρώτου έτους. Χωρίζουμε τα δεδομένα σε 15 topic, μιας και θεωρούμε καλό τον αριθμό για δεδομένα ενός έτους.

	Topic0	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	Topic10	Topic11	Topic12	Topic13	Topic14	dominant_topic1
Doc0	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	**0.61**	0.01	0.01	0.01	0.01	0.01	**0.21**	**8**
Doc1	0.01	0.01	**0.81**	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	**2**
Doc2	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	**0.77**	**14**
Doc3	0.01	0.01	0.01	0.01	0.01	0.01	**0.15**	**0.44**	0.01	**0.15**	0.01	0.01	**0.15**	0.01	0.01	**7**
Doc4	0.01	0.01	0.01	0.01	0.01	0.01	0.01	**0.34**	0.01	0.01	0.01	**0.51**	0.01	0.01	0.01	**11**
Doc5	0.02	**0.27**	0.02	0.02	0.02	0.02	0.02	**0.52**	0.02	0.02	0.02	0.02	0.02	0.02	0.02	**7**
Doc6	0.01	0.01	0.01	0.01	0.01	0.01	**0.41**	0.01	**0.41**	0.01	0.01	0.01	0.01	0.01	0.01	**6**
Doc7	0.02	0.02	0.02	0.02	0.02	0.02	0.02	**0.77**	0.02	0.02	0.02	0.02	0.02	0.02	0.02	**7**
Doc8	0.01	0.01	**0.18**	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	**0.68**	0.01	0.01	**12**
Doc9	0.01	0.01	0.01	0.01	0.01	0.01	**0.21**	**0.41**	0.01	0.01	0.01	0.01	0.01	**0.21**	0.01	**7**

	Topic Num	Num Documents
0	1	9515
1	0	8741
2	2	6644
3	3	5887
4	4	4939
5	5	4590
6	7	3462
7	6	3247
8	11	2950
9	8	2670
10	12	2508
11	10	2463
12	9	2364
13	13	2358
14	14	1665



	Word 0	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9
Topic 0	fight	china	tell	lose	accuse	saddam	cost	blaze	coach	train
Topic 1	court	win	face	trial	murder	man	return	tas	south	search
Topic 2	council	qld	vic	final	act	welcome	arrest	blast	inquiry	day
Topic 3	govt	probe	crash	fund	miss	open	work	reject	head	launch
Topic 4	attack	warn	hospital	concern	end	business	look	target	mayor	market
Topic 5	urge	man	lead	woman	jail	public	appeal	deal	big	suspect
Topic 6	year	change	australia	england	pay	record	want	aid	law	rule
Topic 7	plan	boost	water	air	war	strike	job	farmer	power	union
Topic 8	group	home	sydney	worker	decision	rate	bush	good	chief	west
Topic 9	new	charge	hit	high	ban	rise	car	man	expect	action
Topic 10	death	health	continue	service	minister	child	road	die	force	cut
Topic 11	kill	cup	claim	talk	iraqi	dead	centre	defend	indigenous	push
Topic 12	say	iraq	world	set	support	security	protest	leader	deny	time
Topic 13	police	nsw	seek	make	offer	govt	fear	hold	green	port
Topic 14	report	consider	help	test	drug	case	troop	school	drought	hope

Δίνουμε εγγραφές από τα δεδομένα της επόμενης χρονιάς, του 2004 και παρατηρούμε ότι ο αλγόριθμος μας τα ταξινομεί ορθώς.

```
mytext = ["9 dead as bomb ends aceh new years concert"]
topic, prob_scores = predict_topic(text = mytext)
print(topic)
```

```
['new', 'charge', 'hit', 'high', 'ban', 'rise', 'car', 'man', 'expect', 'action']
```

```
mytext = ["heater program cuts pollution health problems"]
topic, prob_scores = predict_topic(text = mytext)
print(topic)
```

```
['death', 'health', 'continue', 'service', 'minister', 'child', 'road', 'die', 'force', 'cut']
```

```
mytext = ["redknapp charged over cahill crunch tackle"]
topic, prob_scores = predict_topic(text = mytext)
print(topic)
```

```
['new', 'charge', 'hit', 'high', 'ban', 'rise', 'car', 'man', 'expect', 'action']
```

```
mytext = ["scientists warn of shrinking tibetan glaciers"]
topic, prob_scores = predict_topic(text = mytext)
print(topic)
```

```
['attack', 'warn', 'hospital', 'concern', 'end', 'business', 'look', 'target', 'mayor', 'market']
```

```
mytext = ["beckham back for england but terry fitness doubts"]
topic, prob_scores = predict_topic(text = mytext)
print(topic)
```

```
['year', 'change', 'australia', 'england', 'pay', 'record', 'want', 'aid', 'law', 'rule']
```

5. Συμπεράσματα

Στην συγκεκριμένη εργασία προσπαθήσαμε και υλοποιήσαμε ένα πρότυπο επεξεργασίας κειμένων. Χρησιμοποιήσαμε ένα σώμα ειδησεογραφικών νέων από τον ραδιοτηλεοπτικό φορέα της Αυστραλίας ABC News. Αναλύσαμε παραπάνω από 1.000.000 τίτλους τηλεοπτικών ειδήσεων. Ασχοληθήκαμε με τις πιο χρησιμοποιημένες λέξεις, με την εξερεύνηση τυχόν γεγονότων που προκάλεσαν μια αύξηση στον όγκο ειδήσεων. Μελετήσαμε το συναίσθημα των ειδήσεων και πως μεταβάλλεται το συναίσθημα μέσα στον χρόνο. Το επόμενο βήμα ήταν η υλοποίηση της Μοντελοποίησης Θεμάτων. Ανακαλύψαμε τον βέλτιστο αριθμό θεμάτων που πρέπει να ταξινομήσουμε τα δεδομένα, αλλά γιατί αυτό δεν ισχύει στα δικά μας δεδομένα. Υλοποιήσαμε Topic Modelling και βρήκαμε τις λέξεις που συνθέτουν το κάθε θέμα και το βάρος τους για κάθε θέμα. Επίσης ανάλογα με τον όγκο δεδομένων μελετήσαμε την εξέλιξη της ανάθεσης στα Topic ανάλογα με τον όγκο δεδομένων. Ακολούθως ασχοληθήκαμε με την εξέλιξη των topic συναρτήσει του χρόνου. Την εξέλιξη και την ένταση των topics στο πέρασμα των χρόνων. Τέλος, υλοποιήσαμε έναν μικρό αλγόριθμο ταξινόμησης νέων δεδομένων στο πρόγραμμα μας, όπου λειτούργησε αρκετά σωστά..

Γενικά παρατηρήσαμε αν και υπάρχει κοινό πλαίσιο εργασιών, πάντα πρέπει να μαθαίνεις τα δεδομένα και μετά να υλοποιείς την ανάλυση. Στην συγκεκριμένη εργασία, οι τίτλοι έχουν μικρό μέγεθος οπότε όλες οι κινήσεις υλοποίησης ξεκινούσαν με αυτήν την βασική παράμετρο. Πολλά πράγματα που θεωρούνται δεδομένα σε άλλες αναλύσεις, στην περίπτωση μας ήταν αδύνατα.

Επόμενα βήματα, σε μεταγενέστερη εργασία είναι να δούμε τα topic μέσα από την εξέλιξη τους και να χτίσουμε ένα μοντέλο απόφασης ταξινόμησης τους. Να δημιουργηθεί μια αυτόματη διαδικασία προσαρμογής του θεμάτων του αλγόριθμου σε νέα δεδομένα. Επιπλέον να βρεθεί ο κατάλληλος αλγόριθμος που θα μας βοηθήσει να βρίσκουμε τον βέλτιστο αριθμό topic σε ένα σώμα δεδομένων. Ο αλγόριθμος αυτός θα πρέπει να προσαρμόζετε στις ανάγκες μας και στα χαρακτηριστικά των δεδομένων.

6. Αναφορές – Βιβλιογραφία

1] https://www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html#nlpworld

2] https://machinelearningmastery.com/natural-language-processing/

3] https://blog.algorithmia.com/introduction-natural-language-processing-nlp/

4] https://medium.com/nanonets/topic-modeling-with-lsa-psla-lda-and-lda2vec-555ff65b0b05

5] Text mining for central banks, David Bholat, Stephen Hansen, Pedro Santos and Cheryl Schonhardt-Bailey

6] On-line Clustering for Real-Time Topic Detection in Social Media Streaming Data, Robert Popovici, Andreas Weiler, and Michael Grossniklaus Database and Information Systems Group, University of Konstanz

7] Topic Trend Detection and Mining in World Wide Web, Khoo Khyou Bun (48-17035) A thesis presented to The University of Tokyo in fulfillment of the requirement for the degree of PhD in Information and Communication Engineering, Tokyo University, Japan, Supervisor Professor Mitsuru Ishizuka

8] Tracking Topic Birth and Death in LDA, Andrew T. Wilson, David G. Robinson

9] http://aibook.csd.auth.gr/include/ch18.pdf

10] https://forestforthetree.com/statistics/2018/01/28/topic-modelling-with-lsa-and-lda.html

11] https://www.kaggle.com/therohk/million-headlines/home

12] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/SYBGZL

13] http://bbcnewslabs.co.uk/projects/topic-modeling/

14] https://mimno.infosci.cornell.edu/papers/2017_fntir_tm_applications.pdf

15] https://link.springer.com/chapter/10.1007/978-3-319-04126-1_20

16] http://www.aclweb.org/anthology/E17-4007

17] http://datameetsmedia.com/vader-sentiment-analysis-explained/

18] http://www.aclweb.org/anthology/E17-4007

19] http://www.aclweb.org/anthology/P18-2082

20] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003), Latent Dirichlet Allocation, the Journal of

Machine Learning Research, 3, 993-1022.

21] D. Kim, D. H. Park, Y. Lu, C. Zhai, "Enriching text representation with frequent pattern mining for probabilistic topic modelling", Proceedings of the American Society for Information Science and Technology, 49(1):1–10, 2012.

22] Xuerui Wang, Andrew McCallum, Xing Wei, "Topical N-grams: Phrase and Topic Discovery, with an Application to Information Retrieval", University of Massachusetts, 140 Governors Dr, Amherst, MA 01003

23] David M. Blei, "Introduction to Probabilistic Topic Models", Communications of the ACM, 2011 pp.

24] Mark Steyvers, Tom Griffiths, "Probabilistic Topic Models", In Landauer.

25] Zhu, Jun and Eric P Xing, "Conditional Topic Random Fields", Forbes. Ed. Johannes Fürnkranzand Thorsten Joachims.

26] David M. Blei, John D. Lafferty, "Dynamic Topic Models".

27] T. L. Griffiths, Joshua B. Tenenbaum, D. M. Blei, and Michael I. Jordan, "Hierarchical Topic Models and the Nested Chinese Restaurant Process".

28] M. Divya, et al., "A Survey on Topic Modelling", International Journal of Recent Advances in Engineering & Technology (IJRAET), Volume-1, Issue - 2, 2013

29] Hofmann, T., Unsupervised learning by probabilistic latent semantic analysis, Machine Learning, 42 (1), 2001, 177-196.

30] Blei, D. M., Ng, A. Y., and Jordan, M. I., -Latent Dirichlet Allocation, Journal of Machine Learning Research, 3, 2003, 993-1022.

31] Ahmed, A., Xing, E. P., and William W., -Joint Latent Topic Models for Text and Citations, ACM New York, NY, USA, 2008.

32] http://cs229.stanford.edu/proj2012/MengZhangGuo-EvolutionofMovieTopicsOverTime.pdf

33] https://www.datacamp.com/community/tutorials/seaborn-python-tutorial

34] https://www.objectorientedsubject.net/2018/08/experiments-on-topic-modeling-pyldavis/

35] https://pypi.org/project/tqdm/

36] https://pymotw.com/2/subprocess/

37] https://towardsdatascience.com/data-visualization-with-bokeh-in-python-part-one-getting-started-a11655a467d4

38] https://www.techopedia.com/definition/30343/natural-language-toolkit-nltk

39] https://monkeylearn.com/sentiment-analysis/