ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Εφαρμογή Συνάθροισης Άρθρων

Σάββας Αλεξάνδρου

Επιβλέπων Καθηγητής

Μάριος Δικαιάκος

Η Ατομική Διπλωματική Εργασία υποβλήθηκε προς μερική εκπλήρωση των απαιτήσεων απόκτησης του πτυχίου Πληροφορικής του Τμήματος Πληροφορικής του Πανεπιστημίου Κύπρου

Ατομική Διπλωματική Εργασία

Εφαρμογή Συνάθροισης Άρθρων

Σάββας Αλεξάνδρου

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μάιος 2016

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή, κ. Μάριο Δικαιάκο για τη βοήθεια και τις κατευθυντήριες που μου παρείχε για τη περάτωση της διπλωματικής.

Επίσης θα ήθελα να ευχαριστήσω το μεταπτυχιακό φοιτητή Αθανάσιο Φουδούλη για τις εποικοδομητικές συζητήσεις που με βοήθησαν να αναπτύξω την εφαρμογή.

Μάιος 2016

Περίληψη

Σκοπός τη διπλωματικής εργασίας είναι υλοποίηση διαδικτυακής εφαρμογής συνάθροισης άρθρων με τη χρήση του web framework Ruby on Rails.

Για τη δημιουργία της εφαρμογής μελετήθηκε βιβλιογραφία σχετικά με κριτική των μέσων στο παγκόσμιο ιστό και επίσης υφιστάμενες εφαρμογής.

Στη συνέχεια έχουν καθοριστεί οι απαιτήσεις, επιλέχθηκαν λειτουργίες προς υλοποίηση, έχει γίνει σχεδιασμός της εφαρμογής και τέλος η υλοποίηση.

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

1.1 Μέσα Ενημέρωσης στο Παγκόσμιο Ιστό

1.2 Ανασκόπηση βιβλιογραφίας σχετικά με κριτική των Μέσων Ενημέρωσης στο Παγκόσμιο Ιστό

1.3 Σύγκριση Εφαρμογών βάση 3ών κύριων μειονεκτημάτων

1.4 Κίνητρα Διπλωματικής Εργασίας

1.1 Μέσα στο Παγκόσμιο Ιστό

Σχεδόν κάθε εφημερίδα και περιοδικό παρέχεται στον ιστό. Πέρα από τους επαγγελματίες δημοσιογράφους ο καθένας μπορεί άμεσα να παράξει υλικό και να το μοιράσει στον ιστό μέσω μιας εφαρμογής κοινωνικής δικτύωσης ή ενός προσωπικού ή και συλλογικού μπλογκ. Το υλικό αναπαράγεται μέσα από διαδικτυακές εφαρμογές όπως μέσα κοινωνικής δικτύωσης, microblogs, feed aggregators κ.α.

Πως το υλικό και συγκεκριμένα, το κείμενο παρέχεται στις εφαρμογές κοινωνικής δικτύωσης; Η πληροφορία στις εφαρμογές facebook, twitter, αναμεταδίδεται μαζί με ένα μικρό κείμενο που κυμαίνεται 50-150 χαρακτήρες που δίνει μια απλή περιγραφή ή σχόλιο του συνδέσμου που παραπέμπει, μαζί με άλλες πληροφορίες όπως το χρόνο που έχει προστεθεί στην εφαρμογή, κάποιες μετρικές της εφαρμογής όπως σε πόσους χρήστες άρεσε ή δεν άρεσε, πόσοι και ποιοι χρήστες το αναδημοσίευσαν. Πέρα από αυτά παρέχονται κάποιες μετα-πληροφορίες όπως λέξεις κλειδιά αλλά κάποιες φορές και γεωγραφική τοποθεσία από την οποία έχει αναδημοσιευτεί η πληροφορία.

1.2 Προβλήματα που αντιμετωπίζουν οι εφαρμογές – μια επισκόπηση σχετικής βιβλιογραφίας και σχόλια

Η εγκυκλοπαίδεια φιλοσοφίας του Standford στο άρθρο της Social Networking and Ethics αναφέρει:
On any given day on Facebook a user may encounter in her NewsFeed a link to an article in a respected political magazine followed by a video of a cat in a silly costume, followed by a link to a new scientific study, followed by a lengthy status update someone has posted about their lunch, followed by a photo of a popular political figure overlaid with a clever and subversive caption. Vacation photos are mixed in with political rants, invitations to cultural events, birthday reminders and data-driven graphs created to undermine common political, moral or economic beliefs. Thus while a user has a tremendous amount of liberty to choose which forms of discourse to pay closer attention to, and tools with which to hide or prioritize the posts of certain members of her network, she cannot easily shield herself from at least a superficial acquaintance with a diversity of private and public concerns of her fellows.

Το φαινόμενο υπερφόρτωση πληροφορίας είναι το φαινόμενο όπου ο άνθρωπος λαμβάνει μεγάλο όγκο πληροφορίας σε μικρό χρονικό διάστημα με αποτέλεσμα να αδυνατεί να επεξεργαστεί ικανοποιητικά.

Ακόμα πιο πάνω θίγεται το φαινόμενο όπου θέματα που αφορούν την ιδιωτική σφαίρα, παρουσιάζονται με θέματα που αφορούν τη δημόσια σφαίρα.

Ο Daniel J. Levitin στο βιβλίο του The Organized Mind αναφέρει:
Many of us find we don't know whom to believe, what is true, what has been modified, and what has been vetted. We don't have the time or expertise to do research on every little decision. Instead, we rely on trusted authorities, newspapers, radio, TV, books, sometimes your brother-in-law, the neighbor with the perfect lawn, the cab driver who dropped you at the airport, your memory of a similar experience. . . . Sometimes these authorities are worthy of our trust, sometimes not.

Η εγκυκλοπαίδεια plato του standford σε άλλο σημείο αναφέρει:
A related topic of concern is the potential of the Internet to fragment the public sphere by encouraging the formation of a plurality of 'echo chambers' and 'filter bubbles': informational silos for like-minded individuals who deliberately shield themselves from exposure to alternative views. The worry is that such insularity will promote extremism and the reinforcement of ill-founded opinions, while also preventing citizens of a democracy from recognizing their shared interests and experiences (Sunstein 2008)

Shared Information Bias είναι γνωστό ως η τάση συνόλων να αναλώνουν περισσότερο χρόνο και ενέργεια συζητώντας για πληροφορίες που όλα τα μέλη είναι εξοικειωμένα και λιγότερο χρόνο και ενέργεια για πράγματα που μόνο κάποιο μέλη είναι ενήμερα.

Framing effect: είναι ένα παράδειγμα cognitive bias όπου οι άνθρωποι αντιδρούν διαφορετικά σε μια επιλογή ανάλογα με το τρόπο που παρουσιάζεται.

Εν κατακλείδι μπορούμε να απαριθμήσουμε τα θέματα που αντιμετωπίζουν ως εξής: Υπερφόρτωση Πληροφορίας, Εμπιστοσύνη, Ιδιωτική VS Δημόσια Σφαίρα, social bubbles και shared information bias, filter bubbles, framing effect.

1.3 Τι κάνουν οι εφαρμογές για να αντιμετωπίσουν τα πιο πάνω θέματα

Το 1995 είχαν προταθεί οι News Fitlering Agents οι οποίο θα βοηθούν το χρήστη να επιλέξει από μια ροή ειδήσεων τι να διαβάσει (Pattie). Στο άρθρο επισημαίνετε πως χρειάζεται να αντιμετωπιστούν δύο θέματα. Να καθοριστούν οι αρμοδιότητες του ατζέντη, δηλαδή το πως θα ζητά τη πληροφορία που χρειάζεται για να αποφασίσει, πότε να βοηθήσει το χρήστη, με τι να τον βοηθήσει και πως να το βοηθήσει. Το δεύτερο είναι η εμπιστοσύνη, δηλαδή το πως μπορεί να εγγυηθεί πως ο χρήστη νοιώθει άνετα να αναθέτει εργασίες στον ατζέντη.

Τι κάνουν όμως οι πιο δημοφιλής εφαρμογές; Το facebook χρησιμοποιεί ιδιόκτητο αλγόριθμο για να φιλτράρει και να ταξινομεί τη πληροφορία στα news feed του κάθε χρήστη. Το reddit από την άλλη που λειτουργεί με "διαφάνεια" ως προς το τρόπο που το κάνει, δίνει την ευκαιρία στα καινούργια άρθρα να βρίσκονται ψηλά στην ιεραρχία αλλά αν δεν αποκτήσουν αρκετά upvotes με τη πάροδο του χρόνου πέφτει η δημοσιότητα τους.

Στο facebook όπως και στο twitter βλέπουμε τη πληροφορία που αναπαράγουν άτομα που ακολουθούμε ή έχουμε "φίλους". Οπόταν μπορούμε να πούμε πως η πληροφορία φιλτράρεται από το κοινωνικό κύκλο που επιλέγουμε πριν φτάσει σε εμάς, και αυτό του δίνει περισσότερη εμπιστοσύνη. Το reddit το κάνει αυτό με ένα ενδιαφέρον τρόπο. Είναι χωρισμένο σε κοινότητες (subreddit) όπου η κάθε κοινότητα ορίζει δικούς της κανόνες για το περιεχόμενο που θα μοιράζεται μέσω του subreddit. Επίσης παρέχει τη δυνατότητα της χρήσης ψευδωνύμου, χωρίς να ζητά πολλές πληροφορίες για τη φυσική ταυτότητα του χρήστη. Έτσι οι χρήστες επιλέγουν τη κοινότητα που φαίνεται να εμπιστεύονται καλύτερα, χωρίς απαραίτητα τα χαρακτηριστικά των χρηστών που ανήκουν σε αυτή να ταυτίζονται με τα φυσικά χαρακτηριστικά της κοινωνικής ομάδας που ανήκει ο χρήστης στη καθημερινή του ζωή.

Κεφάλαιο 2

Καθορισμός Απαιτήσεων

2.1 Καθορισμός λειτουργιών άτυπα
2.2 Τυποποίηση λειτουργιών

2.1 Καθορισμός λειτουργιών άτυπα

Συνήθως διαβάζω από καμιά δεκαριά online εφημερίδες και περιοδικά. Επίσης αραιά και που μπαίνω σε μπλογκ φίλων που δημοσιεύουν ενδιαφέρον άρθρα για θέματα που μ ενδιαφέρουν. Υπάρχει περίπτωση όμως σε κάποιο μπλογκ να αναρτηθεί άρθρο που είναι άσχετο με τη θεματολογία. Οπόταν θα προτιμούσα να προσθέτω όλα αυτά σε μια εφαρμογή και να εμφανίζονται ομαδοποιημένα, ούτως ώστε να συγκρίνω την ίδια πληροφορία από διαφορετικές πηγές, και επίσης αν λένε λίγο πολύ το ίδιο να μην μου αποσπά τη προσοχή σε διαφορετικά χρονικά διαστήματα. Εκείνη τη στιγμή που το διαβάζω μια 5λεπτη έρευνα θα ήταν πολύ βοηθητική για να μην προσλάβω εντελώς ακατέργαστη τη πληροφορία. Ο κόσμος συχνά κάνει σχόλια στο twitter ή στο reddit που φαίνονται ενδιαφέρον. Οπόταν θα ήταν βοηθητικό μαζί με το άρθρο η εφαρμογή να παρουσιάζει κάποια σχόλια που έχουν παραθέσει χρήστες στα πιο πάνω κοινωνικά δίκτυα.

2.2 Τυποποίηση λειτουργιών

Το πιο πάνω μπορεί να σπάσει σε μικρότερες λειτουργίες ως εξής

1. Προσθήκη στην εφαρμογή Ιστοτόπων απ όπου ενημερώνομαι
2. Επισκόπηση του νέου περιεχομένου που έχει προστεθεί
3. Σχόλια από κοινωνικά δίκτυα
4. Ομαδοποίηση περιεχομένου

Στα πλαίσια της παρούσας διπλωματικής έχουν επιλεχθεί αν υλοποιηθούν οι λειτουργίες 1,4,2

Κεφάλαιο 3

Μεθοδολογία Ανάπτυξης Λογισμικού

4.1 Επιλογή Μοντέλου
4.2 Επιλογή Μεθοδολογίας
4.3 Επιλογή Framework για ανάπτυξη εφαρμογής
4.4 Περιβάλλον ανάπτυξης

4.1 Επιλογή Μοντέλου

Οι πιο πάνω απαιτήσεις μας οδηγούν στην επιλογή του Ευέλικτου Μοντέλου.
Σε αντίθεση με τις παραδοσιακά μοντέλα ανάπτυξης στο Ευέλικτο Μοντέλο δεν χρειάζεται να οριστούν όλες οι απαιτήσεις από την αρχή της ανάπτυξης.
Σύμφωνα με τον Dave Thomas, ένας από τους συντάκτες του Agile Manifest η ευέλικτη μεθοδολογία αποτελείται από 4 βήματα.

- *Δες που βρίσκεσαι*
- *Κάνε ένα βήμα προς το στόχο σου*
- *Προσάρμοσε τη κατανόησή σου βάση αυτών που έμαθες*
- *Επανέλαβε.*

Όταν έρχεσαι αντιμέτωπος με δύο ή περισσότερες εναλλακτικές λύσεις που δίνουν το ίδιο αποτέλεσμα, πάρε το μονοπάτι που θα κάνει στο μέλλον κάποια ενδεχόμενη αλλαγή πιο εύκολη.
Έτσι οποιεσδήποτε εισηγήσεις για λειτουργίες που Συγκεκριμένες, Μετρήσιμες, Εφικτές, Σχετικές και Χρονοπρογραμματίσιμες και δύναται να κάνουν καλύτερη την εφαρμογή είναι ευπρόσδεκτες σε οποιαδήποτε φάση της ανάπτυξης.

4.2 Επιλογή Μεθοδολογίας

Η Behaviour Driven Development (BDD) είναι μια εξελιγμένη μορφή του Test Driven Development. Η BDD ορίζει πως τα tests οποιασδήποτε μονάδας λογισμικού χρειάζεται να ορίζονται με βάση τη “αναμενόμενης συμπεριφορά” της μονάδας, και κατ επέκταση τι θα παρέχουν στο τελικό χρήστη. Η αναμενόμενη συμπεριφορά στη περίπτωση των απαιτήσεων που ορίζονται από μια επιχείρηση χρειάζεται να προσδίδουν κάποια επιχειρηματική αξία.

4.3 Επιλογή Framework για ανάπτυξη εφαρμογής

Για την ανάπτυξη της εφαρμογής έχει επιλεχθεί το Ruby on Rails framework. Το framework ανάπτυξης διαδικτυακών εφαρμογών Ruby on Rails ταιριάζει αρκετά με την ανάπτυξη εφαρμογών με τη συγκεκριμένη ανάπτυξη λογισμικού αφού δίνει έμφαση στη φιλοσοφία “Convention over configuration” και “Don't Repeat Yourself”.
Η πρώτη αναφέρεται στο ότι μειώνει τις αποφάσεις που δύναται να λάβει ο προγραμματιστής κάνοντας κάποιες συμβάσεις. Ένα παράδειγμα είναι ότι εάν ο προγραμματιστής δημιουργήσει ένα πίνακα στη βάση για “Order”, το αντίστοιχο μοντέλο στην εφαρμογή θα έχει το όνομα order, ο αντίστοιχος orders_controller, και εάν ο controller έχει new και edit, τότε αντίστοιχα θα υπάρχουν ένα new_view και edit_view.
Η δεύτερη φιλοσοφία έχει την έννοια του να επαναχρησιμοποιείται όσο πιο πολύ γίνεται κώδικας κάτι που μπορεί να μειώσει τα σφάλματα και να κάνει την ανάπτυξη πιο εύκολη.
Τέλος το framework της Ruby on Rails έχει μια μεγάλη κοινότητας προγραμματιστών με αρκετές διαθέσιμες ελεύθερες βιβλιοθήκες που ονομάζονται gems.

4.4 Περιβάλλον ανάπτυξης

Η εφαρμογή έχει αναπτυχθεί σε προσωπικό λάπτοπ με χρήση editor emacs και τερματικού. Στον emacs έχουν χρησιμοποιηθεί το rinari-mode που υποστηρίζει την ανάπτυξη εφαρμογών rails με συντομεύσεις για μετακίνηση μεταξύ αρχείων του κώδικα. Για διατήρηση εκδόσεων του κώδικα έχει χρησιμοποιηθεί η υπηρεσία που παρέχει το gitlab https://gitlab.com/savinos/kafenes. Αρκετές πληροφορίες έχουν ανακτηθεί από τις σελίδες http://guides.rubyonrails.org/, http://ruby-doc.org/, https://rubygems.org/, http://stackoverflow.com/.

Κεφάλαιο 5

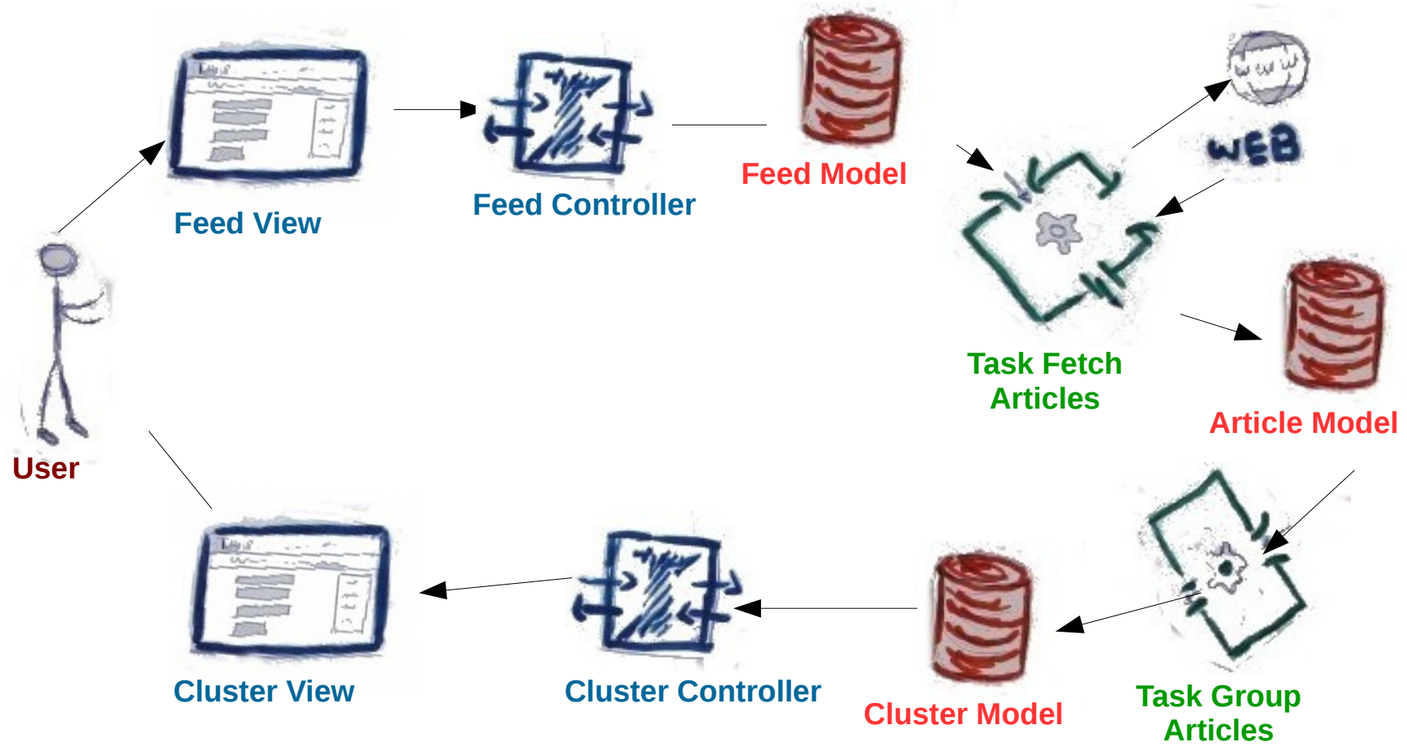
Σχεδιασμός Εφαρμογής

5.1 Σχεδιασμός Εφαρμογής

5.2 Λειτουργία Ομαδοποίησης Άρθρων

5.3 Συλλογή Περιεχόμενου Από τον Ιστό

5.1 Σχεδιασμός Εφαρμογής



Drawing 1: Task Fetch Articles: Μέθοδος για συλλογή περιεχομένου από τον ιστό δεδομένων των feed που υπάρχουν στη βάση.Articles: Κλάση υπεύθυνη για την αποθήκευση και ανάκτηση Καναλιών προς και από τη βάσηTask Group Articles: Μέθοδος για ομαδοποίηση των άρθρων ή μέρος άρθρων που υπάρχουν στη βάση

5.2 Λειτουργία Ομαδοποίησης Άρθρων

Για την ομαδοποίηση (clustering) των άρθρων λαμβάνονται ως δεδομένα μόνο το κείμενο του άρθρου. Ο Αλγόριθμος λειτουργά ως εξής.

Δεδομένου ενός συνόλου από N clusters (ένα άρθρο θεωρείται αρχικά ως singleton cluster)

1. Βρες την ομοιότητα μεταξύ όλων των στοιχείων του συνόλου
2. Συγχώνευσε το ζεύγος των clusters το οποίο έχει τη μεγαλύτερη δυνατή ομοιότητα και ισχύουν οι πιο κάτω συνθήκες
   a) Η ομοιότητά τους είναι μεγαλύτερη από το άθροισμα της μέσης τιμής και τυπικής απόκλιση της ομοιότητας όλων των clusters του αρχικού συνόλου
   b) Η μικρότερη ομοιότητα ζεύγους άρθρων μεταξύ των clusters είναι μεγαλύτερη από το άθροισμα της μέσης τιμής και της μισής τυπικής απόκλισης της ομοιότητας του αρχικού συνόλου.
3. Αν δεν βρεις τέτοιο ζεύγος τερμάτισε

Χρονική πολυπλοκότητα: $O(N^3)$ εφόσον για το πολύ N φορές ελέγχει N*N στο πίνακα.
Χωρική πολυπλοκότητα: $N*k + N^2$, όπου k το μέσο μήκος ενός άρθρου.

Για την ομοιότητα των άρθρων ή και clusters:

Αναπαριστάται ο κάθε cluster ως διάνυσμα μέσα στο χώρο όπου κάθε διάσταση είναι κάθε μοναδικός όρος στο σύνολο των άρθρων. Η βαρύτητα του κάθε όρου ορίζεται από το tf-idf (term frequence – inverse document frequency). Η ομοιότητα μεταξύ δύο άρθρων είναι το cosine similarity μεταξύ των μοναδιαίων διανυσμάτων τους στο χώρο.

Τα πιο πάνω διανύσματα έχουν χρησιμοποιηθεί για εξαγωγή σημμασιολογίας από τον cluster, που είναι τα μεγαλύτερα βάρη στον cluster.

Σύγκριση άρθρων - Παράδειγμα

Άρθρο1	**Άρθρο2**	**Άρθρο3**
κυπριακό	οικονομία	κυπριακό, οικονομία

Table 1:

όνομα	**tf-idf κυπριακό**	**tf-idf οικονομία**
άρθρο1	0	1
άρθρο2	1	0
άρθρο3	1	1

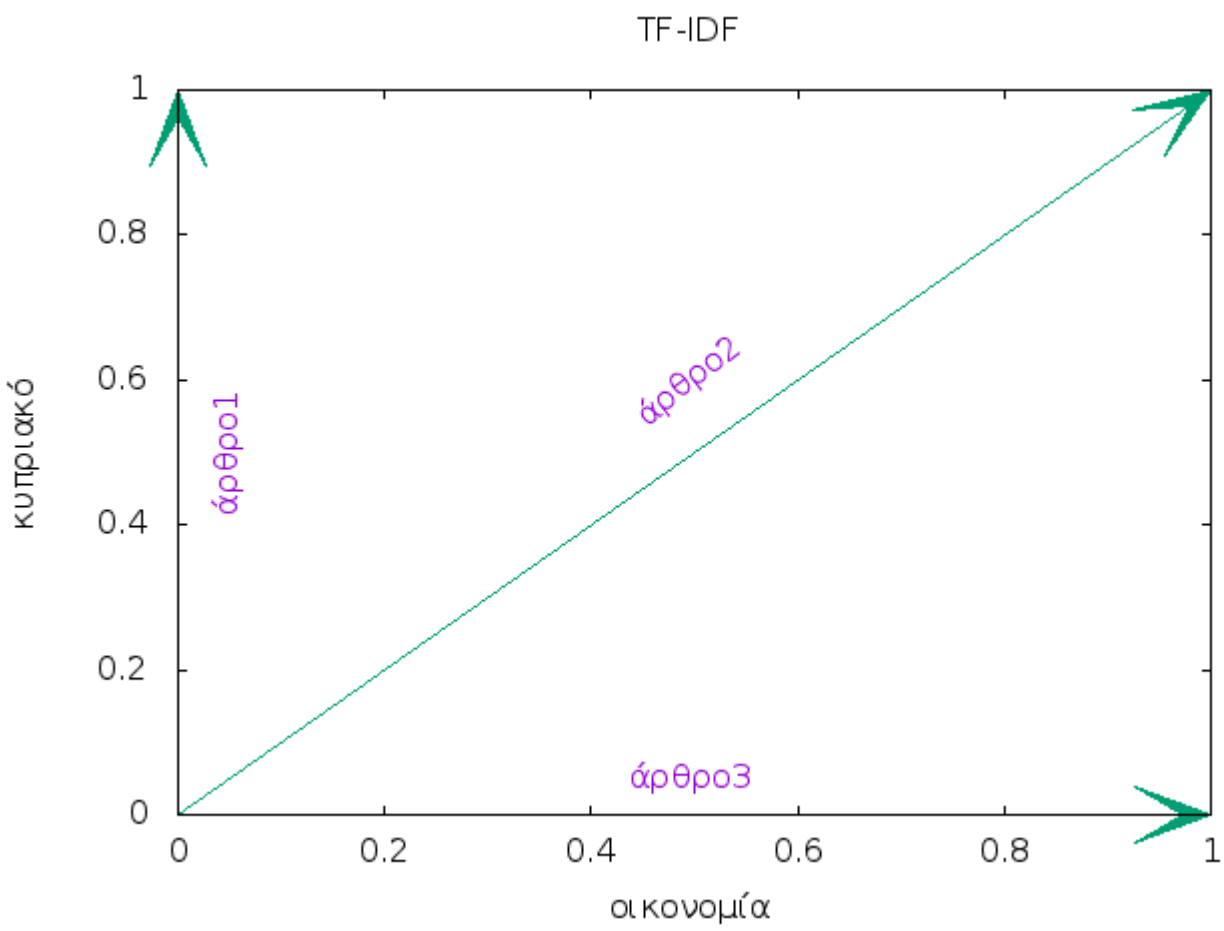


Figure 1: Παράδειγμα TF-IDF

- TF: Η συχνότητα εμφάνισης μιας λέξης σε ένα κείμενο. (Βιβλίο Data Mining pg 8)

  $TF_{ij} = f_{ij} / max_k(f_{kj})$

  Ο αριθμός των φορών που εμφανίζεται η λέξη διά τον αριθμό των φορών που εμφανίζεται η πιο συχνή λέξη στο κείμενο (εξαιρούμενες οι Stop Words)

  $IDF = log(N/n_i)$

  Ο λογάριθμος του συνολικού αριθμού άρθρων ως προς τον αριθμό των άρθρων που εμφανίζεται η συγκεκριμένη λέξη

  Το TF-IDF λειτουργεί χονδρικά ως εξής.

  - Αν μια λέξη εμφανίζεται σε περίπου 1/3 των άρθρων τότε η συχνότητα της λέξης στο άρθρο TF ισούται με το TF-IDF
  - Αν μια λέξη εμφανίζεται σε λιγότερο από το 1/3 των άρθρων τότε το TF-IDF είναι μεγαλύτερο από τη συχνότητα της λέξης στο άρθρο
  - Όσο μια λέξη εμφανίζεται σε περισσότερα από το 1/3 των άρθρων τότε το TF-IDF μειώνεται σε σχέση με το TF.
  - Αν μια λέξη εμφανίζεται σε σχεδόν όλα τα άρθρα τότε το TF-IDF πλησιάζει το 0

5.3 Συλλογή Περιεχόμενου Από τον Ιστό

Για τη συλλογή άρθρων εφόσον βρεθούν όλα τα καινούργια άρθρα από τους ιστοτόπους η εφαρμογή αρχίζει να κατεβάζει και να αποθηκεύει από τα άρθρα τα εξής στοιχεία

1. Κείμενο
2. Ημερομηνία Δημοσίευσης
3. Γλώσσα Κειμένου
4. Τίτλος
5. URL

Από τα πιο πάνω στοιχεία το κείμενο και η γλώσσα είναι τα απαραίτητα για την λειτουργία ομαδοποίησης. Η λειτουργία

ομαδοποίησης επιλέγει να ομαδοποιήσει άρθρα που έχουν κοινή γλώσσα.

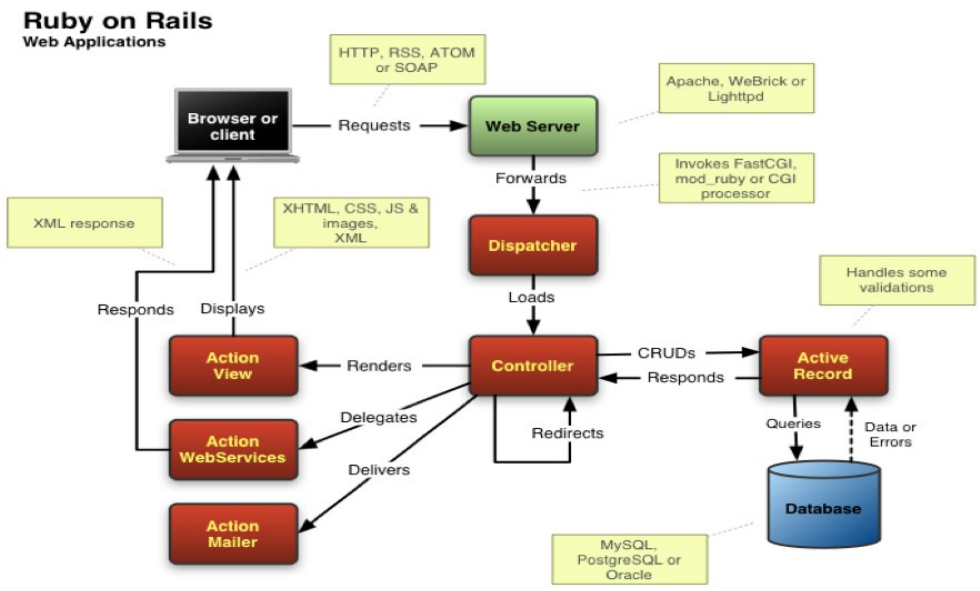
Για το ταχύτερο κατέβασμα των άρθρων χρησιμοποιούνται νήματα.

Κεφάλαιο 6

Υλοποίηση

6.1 Αρχιτεκτονική Framework
6.2 Επιλογή Τεχνολογίας Βάσης
6.4 Τεχνολογίες για testing
6.3 Επιλογή τεχνολογίας για μάζεμα περιεχομένου από τον ιστό
6.4 Υλοποίηση Καναλιων(Feeds)
6.5 Υλοποίηση Article Model
6.6 Υλοποίηση Μάζεμα δεδομένων από τον ιστό
6.7 Υλοποίηση Clustering Άρθρων
6.8 Υλοποίηση View Clustering Άρθρων

6.1 Αρχιτεκτονική Framework



- Active Record: Κλάση που παρέχει τη δυνατότητα παρουσίασης δεδομένων από τη βάσης ως αντικείμενα μέσω των οποίο ο προγραμματιστής μπορεί να τα επεξεργάζεται με χρήση μεθόδων. Αφορά το Model Layer.
- Action Controller: Κλάση υπεύθυνη για τη διαχείριση αιτημάτων http παρέχοντας κατάλληλη απάντηση. Γίνεται η φόρτωση και η επεξεργασία μοντέλων, πριν να δημιουργηθεί το τελικό περιεχόμενο που θα σταλεί στον client. Τα εισερχόμενα αιτήματα επεξεργάζονται από το Action Dispatch και στέλνουν το αντίστοιχο αίτημα στο controller. Ο Action Dispatch και ο Action Controller μαζί δημιουργούν ένα Action Pack.
- Action View: Το επίπεδο του View διαμορφώνεται από templates που είναι υπεύθυνα για να παρέχουν το κατάλληλη αναπαράσταση των πόρων της εφαρμογής. Στα template μπορεί να γράψει κανείς ruby κώδικα.

5.1 Επιλογή Τεχνολογίας Βάσης Δεδομένων

Για τη συγκεκριμένη εφαρμογή έχει επιλεγεί η postgresql.
Η postgresql είναι ACID(Atomicity, Consistency, Isolation, Durability) compliance και το πλεονέκτημα της είναι ότι υποστηρίζει ταυτοχρονία με τη χρήση του πρωτοκόλλου Multiversion Concurrency Control, MVCC το οποίο μειώνει τον αριθμό των κλειδωμάτων στα πεδία παρέχοντας έτσι απόδοση σε περίπτωση που πολλές διεργασίες επιχειρούν να εκτελέσουν

συναλλαγές με τη βάση.

Το Multiversion Concurrency Control είναι πρωτόκολλο ταυτοχρονίας όπου κάθε επερώτηση στη βάση διαβάζει ένα στιγμιότυπο των δεδομένων (έκδοση της βάσης) όπως ήταν λίγο χρόνο πριν. Γενικά το διάβασμα δεν μπλοκάρει το γράψιμο, και το γράψιμο δεν μπλοκάρει το διάβασμα.

5.2 Επιλογή τεχνολογίας για συλλογή περιεχομένου από τον ιστό

Για τη συλλογή περιεχομένου από τον ιστό έχει επιλεγεί RSS και Atom. Το RSS και Atom είναι πρώτυπα ιστού.
Τα πρότυπα από τα οποία περιεχόμενο μοιράζεται στον ιστό είναι το RSS 1.0, RSS 2.0 και atom. Το RSS 1.0 (RDF site summary) είχε οριστεί το 2000 από ομάδα εργασία που ηγείτο ο Rael Dornfest με το όνομα RSS-DEV Working Group και το RSS 2.0 (Really Simple Syndication)ανήκει στο Harvard. Το Atom από την άλλη είναι πρότυπο του οργανισμού Internet Engineering Task Force. Όλα συμφωνούν με το πρότυπο XML 1.0. Επίσης όλα απαιτούν τον ορισμό πεδίου “URL” και “Title”.

5.3 Υλοποίηση Καναλιων(Feeds)

Για την υλοποίηση των Καναλιών (Feeds) έχουν καταγραφεί οι απαιτήσεις του χρήστη για τη διεπαφή με χρήση εργαλείου cucumber.

Το εργαλείο cucumber έχει φτιαχτεί για να γεφυρώσει το χάσμα επικοινωνίας μεταξύ προγραμματιστών και ατόμων που δεν έχουν γνώσεις προγραμματισμού. Χρησιμοποιεί τη businnes model language gherkin και συνήθως χρησιμοποιείται για να τεστάρει τη διεπαφή με το χρήστη. Το κύριο πλεονέκτημα είναι ότι η χρήση του εργαλείου cucumber δίνει την ευκαιρία σε άτομα που έχουν ιδέες για λειτουργίες της εφαρμογής να μπορούν να γράφουν απευθείας τα testing.

Οι controller και το model ελέγχεται με τη χρήση του εργαλείου rspec.

Το μοντέλο του feed παρέχει μέθοδο που επιστρέφει όλα τα καινούργια url από όλα τα κανάλια που υπάρχουν καταχωρημένα στην εφαρμογή.

5.5 Υλοποίηση Συλλογή Περιεχομένου από τον ιστό

Η εφαρμογή για να συλλογή περιεχομένου από τον ιστό ανατρέχει τα url που παρέχονται μέσω RSS/Atom και τα κατεβάζει. Για τη συγκεκριμένη υλοποίηση χρησιμοποιήθηκαν 2 βιβλιοθήκες. Η feedjira για διαχείριση του RSS/Atom και η pismo για κατέβασμα του άρθρου.

Με το κατέβασμα του άρθρου ελέγχεται και αποθηκεύεται στη βάση η γλώσσα του άρθρου η οποία ελέγχεται με τη βιβλιοθήκη (gem) whatlanguage

Για τη χρήση νημάτων χρησιμοποιήθηκαν η βιβλιοθήκες sidekiq και redis.

5.6 Υλοποίηση Ομαδοποίησης Άρθρων

Για την ομαδοποίηση άρθρων το κείμενο φιλτράρεται με χρήση των gem fast-stemmer για “stemming” των λέξεων. Επίσης αφαιρούνται τα stopwords.

Η δημιουργία του μοντέλου γίνεται με τη χρήση της βιβλιοθήκης tf-idf-similarity και της βιβλιοθήκης narray.

Βιβλιογραφία

http://www.owenspencer-thomas.com/journalism/newsvalues

https://medium.com/hacking-and-gonzo/how-reddit-ranking-algorithms-work-ef111e33d0d9#.t0unia1mq

https://www.bostonglobe.com/news/nation/2014/05/03/facebook-push-related-articles-users-without-checking-credibility-draws-fire/rPae4M2LlzpVHIJAmfDYNL/story.html

https://medium.com/hacking-and-gonzo/how-reddit-ranking-algorithms-work-ef111e33d0d9#.t0unia1mq

http://thenextweb.com/twitter/2016/03/17/twitter-quietly-turned-new-algorithmic-timeline-everyone

https://blog.bufferapp.com/optimal-length-social-media

http://techcrunch.com/2010/04/22/facebook-edgerank/

https://ec.europa.eu/digital-single-market/en/scoreboard/cyprus

https://pragdave.me/blog/2014/03/04/time-to-kill-agile/

http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.pdf

http://www.sentia.com.au/blog/8-benefits-of-using-ruby-on-rails-for-web-development

http://web.resource.org/rss/1.0/spec

http://cyber.law.harvard.edu/rss/rss.html

https://tools.ietf.org/id/draft-nottingham-rss-media-type-00.txt

https://tools.ietf.org/rfc/rfc4287.txt

https://www.postgresql.org/docs/9.3/static/mvcc-intro.html

http://adrianmejia.com/blog/2011/08/11/ruby-on-rails-architectural-design/

http://news.indiana.edu/releases/iu/2015/12/social-media-bubbles.shtml

Παράρτημα Α

Κώδικας για Ομαδοποίηση Άρθρων

```
 desc "Ομαδοποίηση Άρθρων"
  task divide_and_conquer: :environment do
    from_days_before = 3
    to_days_before = 0
    language = "english"
    # Πέρνει όλα τα άρθρα
    @corpus = []
    Article.where(:language => language).where(published_at: (Time.now.midnight - from_days_before.day)..(Time.now.midnight -
to_days_before.day)).each_with_index { |article, index|
      document = MyFilter.filter(article.body)
      @corpus << TfIdfSimilarity::Document.new(document, :urls => [article.url], :ids => [index])
    }

    # Δημιουργεί το μοντέλλο και similarity matrix
    model = TfIdfSimilarity::TfIdfModel.new(@corpus, :library => :narray)
    matrix = model.similarity_matrix

    @working_matrix = MyNarray.new_narray(matrix)

    # υπολογισμός μέσου και stddev
    mean = MyNarray.mean(matrix)
    stddev = MyNarray.stddev(matrix)
    threshold = mean + stddev
    reduced_threshold = mean + (stddev/2.0)
    # δημιουργία clusters
    loop do
      candidate_to_merge = MyNarray.sorted_indexes_larger_than(@working_matrix,threshold)
      break if candidate_to_merge.length == 0
      merged_flag = 0
      puts @working_matrix.shape[0]
      candidate_to_merge.each { |m|
        furthest = MyNarray.furthest_similarity(matrix, @corpus[m[0]].ids, @corpus[m[1]].ids)
        if furthest >= reduced_threshold
          model.merge_and_update(m[0], m[1])
          merged_flag = 1
          break
        end
      }
      break if merged_flag == 0
      @working_matrix = model.similarity_matrix
    end

    # αποθήκευση μοντέλλου και clusters στη βάση
    @save_model = Model.create(:model_json => model.to_json)
    model.documents.each_with_index { |d,i|
      Cluster.create(:model_id => @save_model.id, :articles_urls => d.urls.to_json, :keywords => model.keywords(i),
:number_of_documents => d.urls.length)
      puts "Cluster #{i} #{model.keywords(i)}"
      puts "============"
      puts d.urls
    }

  end

end
```

Κώδικας για συλλογή περιεχομένου

```
namespace :articles do
  desc "Κατέβασε όλα τα καινούργια άρθρα από τα κανάλια"
  task fetch: :environment do
    # OPTIMIZE minimize http request response
    urls = []
    Feed.all.each { |f|
      urls.concat(f.get_new_articles)
    }

    urls.uniq!

    urls.shuffle.each { |u|
      DownloadArticle.perform_async(u)
    }
  end
```

Κώδικας worker για thread

```
class DownloadArticle
  include Sidekiq::Worker

  sidekiq_options :retry => 0
  def perform(url)
    @article = Article.new(:url => url)
    if @article.valid?
      @article.save
    end
  end

end
```

Παράρτημα Β

Αποτελέσματα από από εκτέλεση ομαδοποίησης άρθρων

Cluster 0 ["student", "bone", "potter", "skelegro", "harri", "gill"]
==========
https://www.theguardian.com/books/2016/may/27/scientists-test-reality-of-harry-potter-magic-university-leicester-gillyweed-skele-gro
Cluster 1 ["speci", "plant", "fossil", "cretac", "kpg", "extinct"]
==========
https://www.theguardian.com/science/2016/may/25/dinosaur-extinction-only-half-the-story-of-killer-asteroids-impact-plant-fossil
Cluster 2 ["huchard", "mammal", "queue", "individu", "reproduct", "meerkat"]
==========
https://www.theguardian.com/science/animal-magic/2016/may/25/size-matters-meerkats-keep-tabs-on-rivals-growth-and-eat-to-compete
Cluster 3 ["zhang", "contact", "toxic", "skin", "paraquat", "poison"]
==========
https://www.theguardian.com/science/blog/2016/may/27/knickers-in-a-twist-the-case-of-the-poisoned-pants-paraquat
Cluster 4 ["fund", "govern", "consum", "pension", "uk", "invest"]
==========
https://www.theguardian.com/business/2016/may/27/house-prices-force-1m-young-people-live-with-parents
https://www.theguardian.com/business/2016/may/26/uk-gdp-growth-george-osborne-beware-balance-of-payments
https://www.theguardian.com/business/2016/may/27/austerity-policies-do-more-harm-than-good-imf-study-concludes
https://www.theguardian.com/business/2016/may/27/axa-sells-uk-businesses-sunlife-to-phoenix
https://www.theguardian.com/business/2016/may/26/tata-steel-pension-chief-backs-controversial-cuts
https://www.theguardian.com/business/2016/may/29/tata-steel-must-commit-for-the-long-haul-says-stephen-kinnock
https://www.theguardian.com/business/2016/may/29/cancer-scientists-pensions-invested-in-tobacco-bat
https://www.theguardian.com/business/2016/may/28/memorial-day-sales-online-shopping-summer-us-economy
https://www.theguardian.com/science/2016/may/28/eu-ministers-2020-target-free-access-scientific-papers
Cluster 5 ["vicent", "storytel", "connect", "magnifica", "freewrit", "walter"]
==========
https://www.theguardian.com/technology/2016/may/27/freewrite-hipster-typewriter-technology
https://www.theguardian.com/tv-and-radio/2016/may/27/porn-stars-of-the-junta-magnifica-70-south-americas-breaking-bad-tv-walter-presents
Cluster 6 ["spacecraft", "amino", "acid", "glycin", "planet", "comet"]
==========
https://www.theguardian.com/science/2016/may/27/comet-67p-atmospehere-contains-chemicals-of-life-rosetta-mission-glycine
https://www.theguardian.com/technology/2016/may/28/no-mans-sky-delayed-august-sony-playstation
Cluster 7 ["yravedra", "site", "altamira", "draw", "garat", "cave"]
==========
https://www.theguardian.com/science/2016/may/27/spanish-archaeologists-discover-cave-art-axturra-paleolithic
Cluster 8 ["coppa", "contract", "amazon", "servic", "compani", "appl"]
==========
https://www.theguardian.com/business/2016/may/27/competition-watchdog-uk-cloud-storage
https://www.theguardian.com/technology/2016/may/26/amazon-echo-virtual-assistant-child-privacy-law
https://www.theguardian.com/technology/2016/may/05/us-tech-firms-visas-hillary-clinton-donald-trump-immigration
https://www.theguardian.com/technology/2016/may/27/apple-time-warner-iphone-tv
https://www.theguardian.com/business/2016/may/27/verizon-strike-ends-tentative-deal-union
Cluster 9 ["gene", "inner", "order", "p", "pamp", "aggi"]
==========
https://www.theguardian.com/technology/2016/may/27/worlds-first-robot-gallery-guide-lead-by-a-high-tech-furby-its-hard-to-know-what-to-look-at
https://www.theguardian.com/books/2016/may/27/writers-face-rejection-hanif-kureishi
https://www.theguardian.com/books/2016/may/25/the-voices-within-the-history-and-science-of-how-we-talk-to-ourselves--charles-fernyhough-review
https://www.theguardian.com/books/2016/may/25/the-gene-an-intimate-history-siddhartha-mukherjee-review
https://www.theguardian.com/books/2016/may/26/moonstone-by-sjon-review
https://www.theguardian.com/books/2016/may/24/kid-gloves-voyage-round-my-father-adam-mars-jones-review
https://www.theguardian.com/books/2016/may/27/this-must-be-the-place-by-maggie-ofarrell-review
https://www.theguardian.com/music/2016/may/27/ring-the-changes-my-asuterity-wagner-with-opera-north
Cluster 10 ["committe", "acquisit", "payment", "chappel", "bh", "retail"]
==========
https://www.theguardian.com/business/2016/may/26/portuguese-backed-consortium-close-to-deal-for-bhs
https://www.theguardian.com/business/2016/may/29/bhs-property-investors-profited-backing-dominic-chappell
https://www.theguardian.com/business/2016/may/27/5m-windfall-cited-bhs-buyer-bolster-credentials-disputed-dominic-chappell
https://www.theguardian.com/business/2016/may/26/mps-mike-ashley-not-visit-sports-direct-head-office
https://www.theguardian.com/business/2016/may/27/competition-regulator-could-investigate-sainsburys-home-retail-takeover
https://www.theguardian.com/sport/2016/may/26/olympics-tokyo-2020-papa-massata-diack
Cluster 11 ["thei", "charact", "ar", "superhero", "marvel", "peopl"]

==========
https://www.theguardian.com/technology/2016/may/27/overwatch-review-blizzard
https://www.theguardian.com/books/2016/may/27/comic-book-superheroes-the-gods-of-modern-mythology
https://www.theguardian.com/culture/2016/may/26/marvel-editor-in-chief-axel-alonso-civil-war-x-men
https://www.theguardian.com/film/2016/may/27/gay-captain-america-female-bond-steve-rogers-marvel-civil-war-diversity
https://www.theguardian.com/tv-and-radio/2016/may/27/orange-is-the-new-blacks-jenji-kohan-women-tend-to-be-forgotten-when-they-get-locked-up
https://www.theguardian.com/film/2016/may/27/holy-hell-documentary-truth-about-cults
https://www.theguardian.com/film/2016/may/27/holy-hell-review-buddhafield-cult
https://www.theguardian.com/film/2016/may/27/paul-verhoeven-elle-isabelle-huppert-rape-comedy
https://www.theguardian.com/film/2016/may/20/noam-chomsky-on-donald-trump-almost-a-death-knell-for-the-human-species
https://www.theguardian.com/culture/2016/may/27/hay-festival-shakespeare-experts-clash-over-whether-to-cut-or-not-to-cut
https://www.theguardian.com/books/booksblog/2016/may/28/women-modern-books-gone-girl-girls-on-fire
https://www.theguardian.com/tv-and-radio/2016/may/29/roots-remake-american-slavery-miniseries
https://www.theguardian.com/film/2016/may/27/paul-schrader-willem-dafoe-dog-eat-dog
Cluster 12 ["backdoor", "lazaru", "attack", "malwar", "eurostar", "hack"]
==========
https://www.theguardian.com/business/2016/may/26/eurostar-launches-fare-snap-ticket
https://www.theguardian.com/technology/2016/may/27/swift-network-bank-theft-sony-pictures-hack-lazarus-symantec
Cluster 13 ["game", "wa", "plai", "hi", "england", "player"]
==========
https://www.theguardian.com/business/2016/may/26/the-british-pop-talent-crash-where-have-all-the-new-acts-gone
https://www.theguardian.com/technology/2016/may/27/tesla-model-s-electric-car-driving-holiday
https://www.theguardian.com/books/2016/may/28/can-literary-festivals-pay-their-way
https://www.theguardian.com/football/2016/may/26/marcus-rashford-england-chance-euro-2016-australia
https://www.theguardian.com/football/2016/may/26/roy-hodgson-rashford-sturridge-england-australia
https://www.theguardian.com/football/2016/may/28/marcus-rashford-debut-impresses-england-team-mates
https://www.theguardian.com/football/2016/may/27/england-roy-hodgson-marcus-rashford-australia
https://www.theguardian.com/football/2016/may/28/england-roy-hodgson-euro-2016-daniel-sturridge
https://www.theguardian.com/football/2016/may/27/england-australia-international-friendly-match-report
https://www.theguardian.com/sport/2016/may/27/england-teimana-harrison-test-wales-twickenham-jack-clifford-james-haskell-rugby-union
https://www.theguardian.com/sport/2016/may/29/england-wales-international-match-report
https://www.theguardian.com/sport/2016/may/27/wales-england-warren-gatland-rugby-union
https://www.theguardian.com/sport/2016/may/28/saracens-exeter-premiership-play-off-final-match-report
https://www.theguardian.com/sport/2016/may/28/saracens-mark-mccall-alex-goode-england-exeter-rob-baxter
https://www.theguardian.com/football/blog/2016/may/27/mls-weekend-preview-soccer-news
https://www.theguardian.com/football/blog/2016/may/27/copa-america-usa-argentina-1995
https://www.theguardian.com/football/blog/2016/may/29/copa-america-usa-warm-up-christian-pulisic-soccer
https://www.theguardian.com/sport/2016/may/27/exeter-chiefs-rob-baxter-premiership-final-saracens-twickenham-rugby-union
https://www.theguardian.com/football/2016/may/28/seamus-coleman-republic-of-ireland-euro-2016
https://www.theguardian.com/sport/2016/may/27/pro-12-final-connacht-rugby-leinster
https://www.theguardian.com/football/2016/may/29/barnsley-millwall-league-one-play-off-final-match-report
https://www.theguardian.com/football/2016/may/28/champions-league-kosovo-uefa
https://www.theguardian.com/football/2016/may/29/afc-wimbledon-league-two-playoff-plymouth
https://www.theguardian.com/film/2016/may/26/stephan-james-jesse-james-race-interview
https://www.theguardian.com/football/2016/may/27/sheffield-wednesday-hull-play-off-final-premier-league
https://www.theguardian.com/football/2016/may/26/northern-ireland-roy-carroll-steven-davis
https://www.theguardian.com/film/2016/may/22/me-before-you-film-love-disability-thea-sharrock-sam-claflin
Cluster 14 ["brew", "ahopalyps", "sadler", "wadworth", "beer", "breweri"]
==========
https://www.theguardian.com/business/2016/may/27/aldi-hails-ales-range-uk-craft-beers-supermarket
Cluster 15 ["florenc", "inflat", "grandpa", "fed", "yellen", "rate"]
==========
https://www.theguardian.com/business/2016/may/27/interest-rate-hike-janet-yellen-federal-reserve-memorial-day
https://www.theguardian.com/books/2016/may/29/the-mandibles-a-family-2029-2047-by-lionel-shriver
Cluster 16 ["merlin", "group", "financ", "nichol", "ftse", "wolselei"]
==========
https://www.theguardian.com/business/marketforceslive/2016/may/27/wolseley-slips-after-u-turn-on-finance-director
https://www.theguardian.com/business/2016/may/29/inmarsat-faces-relegation-from-ftse-100
Cluster 17 ["economi", "athen", "tourism", "gillingham", "eu", "greec"]
==========
https://www.theguardian.com/business/live/2016/may/27/markets-us-growth-figures-janet-yellen-putin-greece-live
https://www.theguardian.com/business/2016/may/28/greece-tourism-boom-athens-jobs-growth
https://www.theguardian.com/books/2016/may/26/the-eu-an-obituary-john-r-gillingham-review
https://www.theguardian.com/books/2016/may/29/roberto-saviano-london-is-heart-of-global-financial-corruption
Cluster 18 ["lartigu", "exhibit", "imag", "paint", "art", "photograph"]
==========
https://www.theguardian.com/artanddesign/2016/may/25/david-king-obituary
https://www.theguardian.com/artanddesign/2016/may/27/snap-judgment-how-photographer-jacques-henri-lartigue-captured-

the-moment
https://www.theguardian.com/artanddesign/2016/may/28/francis-bacon-spanish-police-make-arrests-over-stolen-works-madrid
https://www.theguardian.com/artanddesign/2016/may/27/butterflies-bacchanalia-and-francis-bacon-the-week-in-art
https://www.theguardian.com/artanddesign/2016/may/27/five-of-the-best-exhibitions
Cluster 19 ["famili", "thei", "timber", "wood", "beckett", "tree"]
==========
https://www.theguardian.com/artanddesign/2016/may/26/warship-nhs-namur-exhibition-chatham
https://www.theguardian.com/books/2016/may/25/the-wood-for-the-trees-by-richard-fortey-review
https://www.theguardian.com/books/2016/may/27/a-country-road-a-tree-by-jo-baker-review
https://www.theguardian.com/books/2016/may/25/larose-by-louise-erdrich-review
https://www.theguardian.com/stage/2016/may/26/minefield-falklands-theatre-veterans-battle
https://www.theguardian.com/books/2016/may/27/peter-pan-jm-barrie-house
Cluster 20 ["biographi", "kick", "shriver", "marina", "lewycka", "lubetkin"]
==========
https://www.theguardian.com/books/2016/may/27/critical-eye-books-review-roundup
Cluster 21 ["voic", "cultur", "orwel", "gaiman", "hurlei", "geek"]
==========
https://www.theguardian.com/books/2016/may/27/neil-gaiman-kameron-hurley-geek-culture
https://www.theguardian.com/books/2016/may/26/most-orwellian-winner-yet-the-invention-of-russia-takes-orwell-prize
https://www.theguardian.com/culture/2016/may/29/women-and-melbourne-writers-dominate-miles-franklin-2016-shortlist
Cluster 22 ["order", "p", "pamp", "giovanni", "devi", "yoga"]
==========
https://www.theguardian.com/books/2016/may/27/goddess-pose-michelle-goldberg-review-yoga-popular
https://www.theguardian.com/books/2016/may/27/gut-by-guilia-endere-review-celebration-of-our-most-under-rated-organ
https://www.theguardian.com/books/2016/may/26/the-morning-they-came-for-us-janine-di-giovanni-syria
Cluster 23 ["gear", "stunt", "fun", "laugh", "void", "funni"]
==========
https://www.theguardian.com/books/booksblog/2016/may/28/why-this-years-wodehouse-prize-winners-are-on-the-money-judges-view
https://www.theguardian.com/tv-and-radio/2016/may/29/new-top-gear-review-matt-leblanc-chris-evans-bbc
Cluster 24 ["adoo", "compos", "music", "poem", "poet", "ginsberg"]
==========
https://www.theguardian.com/books/2016/may/27/allen-ginsberg-boxset-music-bob-dylan
https://www.theguardian.com/music/2016/may/13/english-composers-minor-poems-major-works-elgar-vaughan-williams
https://www.theguardian.com/music/2016/may/27/and-on-the-mimu-gloves-the-ingenious-devices-helping-disabled-musicians-to-play-again
Cluster 25 ["workshop", "rave", "preacher", "doc", "rattigan", "fest"]
==========
https://www.theguardian.com/culture/2016/may/27/the-10-best-things-to-do-this-week
Cluster 26 ["presidenti", "disnei", "disneyland", "arum", "berni", "sander"]
==========
https://www.theguardian.com/film/2016/may/26/bernie-sanders-disney-california-attack-wages-mickey-mouse
https://www.theguardian.com/sport/2016/may/27/bob-arum-says-he-wants-to-promote-donald-trump-v-bernie-sanders-debate
Cluster 27 ["batman", "charm", "surprisingli", "lopez", "soderbergh", "cloonei"]
==========
https://www.theguardian.com/film/2016/may/27/george-clooney-five-best-moments
Cluster 28 ["ring", "bell", "song", "skrillex", "tucker", "bieber"]
==========
https://www.theguardian.com/music/2016/may/27/new-band-of-the-week-bad-wave-no-105
https://www.theguardian.com/music/2016/may/30/skrillex-justin-bieber-sorry-sample-plagiarism-allegations-we-didnt-steal-this
Cluster 29 ["guitar", "band", "music", "radiohead", "song", "album"]
==========
https://www.theguardian.com/music/2016/may/27/what-does-it-mean-to-be-a-music-critic-in-the-age-of-the-stealth-release
https://www.theguardian.com/music/2016/may/24/dierks-bentley-interview-black-country-music
https://www.theguardian.com/music/2016/may/26/adam-and-the-ants-heroic-sexy-warrior-bravado-kings-wild-frontier
https://www.theguardian.com/music/2016/may/28/radiohead-review-textured-set-which-flits-between-medicinal-bliss-and-fractious-fury
https://www.theguardian.com/music/2016/may/27/radiohead-review-roundhouse-london-thom-yorke
https://www.theguardian.com/music/2016/may/29/radiohead-live-review-roundhouse-london-finale-thom-yorke
https://www.theguardian.com/music/2016/may/26/ponds-nicholas-allbrook-on-australias-national-anthem-its-ignorant-and-isolationist
https://www.theguardian.com/music/2016/may/29/rhys-chatham-pythagorean-dream-influences
Cluster 30 ["adapt", "tom", "mend", "hiddleston", "bbc", "hai"]
==========
https://www.theguardian.com/film/2016/may/27/andrew-davies-my-les-miserables-will-be-nothing-like-shoddy-farrago-musical
https://www.theguardian.com/film/2016/may/29/next-james-bond-will-not-who-expect-sam-mendes-tom-hiddleston
Cluster 31 ["empathet", "overt", "dog", "anderson", "linklat", "min"]
==========
https://www.theguardian.com/film/2016/may/27/five-of-the-best-films
Cluster 32 ["firth", "vinterberg", "lyachin", "schoenaert", "submarin", "kursk"]

==========
https://www.theguardian.com/film/2016/may/27/colin-firth-to-star-in-russian-submarine-disaster-film-kursk
Cluster 33 ["boat", "rr", "nomin", "bryan", "ferri", "playlist"]
==========
https://www.theguardian.com/music/2016/may/26/readers-recommend-playlist-songs-about-ships-and-boats
Cluster 34 ["rapper", "green", "interview", "rodriguez", "nardwuar", "venu"]
==========
https://www.theguardian.com/music/2016/may/26/ti-new-york-show-shooting-irving-plaza
https://www.theguardian.com/music/2016/may/27/nardwuar-drake-blur-courtney-love-nirvana
Cluster 35 ["forsteroft", "brisban", "gobetween", "pop", "gig", "odd"]
==========
https://www.theguardian.com/music/2016/may/27/five-best-gigs-this-week
Cluster 36 ["jekyl", "hyde", "suppos", "shop", "blood", "dark"]
==========
https://www.theguardian.com/music/2016/may/27/ray-blk-stormzy-1975-coral-goat-dylan-evans
https://www.theguardian.com/stage/2016/may/27/jekyll-and-hyde-review-drew-mconie-company-old-vic-london
Cluster 37 ["dad", "ever", "sketch", "eat", "iv", "funniest"]
==========
https://www.theguardian.com/stage/2016/may/27/kieran-hodgson-funniest-thing-rowan-atkinson-anchorman
Cluster 38 ["berlin", "theatric", "brecht", "peachum", "weill", "macheath"]
==========
https://www.theguardian.com/stage/2016/may/27/the-threepenny-opera-review-olivier-london-rory-kinnear-rufus-norris
Cluster 39 ["gustav", "sofia", "shelbi", "versail", "episod", "tommi"]
==========
https://www.theguardian.com/tv-and-radio/2016/may/27/blues-eyes-finale-recap-goodbye-gustav-the-meek-terrorist
https://www.theguardian.com/tv-and-radio/2016/may/26/peaky-blinders-recap-series-three-episode-four-sickeningly-good
https://www.theguardian.com/tv-and-radio/2016/may/27/antonia-fraser-versailles-bbc-sex-scandal-true
Cluster 40 ["poemanyon", "soul", "aunti", "spacek", "fatima", "track"]
==========
https://www.theguardian.com/music/2016/may/27/steve-spacek-favourite-tracks
Cluster 41 ["simon", "product", "le", "african", "rice", "munro"]
==========
https://www.theguardian.com/stage/2016/may/27/five-of-the-best-plays
Cluster 42 ["bateman", "music", "toni", "theatr", "psycho", "broadwai"]
==========
https://www.theguardian.com/stage/2016/may/27/leona-lewis-to-replace-nicole-scherzinger-in-cats-broadway
https://www.theguardian.com/stage/2016/may/27/american-psycho-broadway-show-cut-competitive-theater-landscape
Cluster 43 ["final", "lisbon", "real", "madrid", "simeon", "atltico"]
==========
https://www.theguardian.com/tv-and-radio/2016/may/28/saturdays-best-tv-the-disappearance-the-musketeers-uefa-champions-league
https://www.theguardian.com/football/2016/may/27/atletico-madrid-real-madrid-champions-league-fernando-torres
https://www.theguardian.com/football/2016/may/29/atletico-madrid-diego-simeone-considers-future-champions-league
https://www.theguardian.com/football/blog/2016/may/28/real-madrid-atletico-european-cup
https://www.theguardian.com/football/2016/may/28/real-madrid-atletico-madrid-champions-league-final-match-report
https://www.theguardian.com/football/live/2016/may/28/real-madrid-v-atletico-madrid-champions-league-final-live-san-siro
https://www.theguardian.com/football/2016/may/28/yannick-carrasco-kiss-alicia-keys-champions-league-final
Cluster 44 ["techniqu", "duncan", "pioneer", "titanium", "granddaught", "ga"]
==========
https://www.theguardian.com/stage/2016/may/27/the-forbidden-zone-review-katie-mitchell
https://www.theguardian.com/science/2016/may/29/john-norton-obituary
Cluster 45 ["mario", "gtze", "premier", "leagu", "season", "club"]
==========
https://www.theguardian.com/football/2016/may/26/bayern-munich-mario-gotze-bench-liverpool
https://www.theguardian.com/football/2016/may/27/football-transfer-rumours-manchester-united-to-sign-54m-saul-niguez
https://www.theguardian.com/football/2016/may/27/derby-county-appoint-nigel-pearson-manager
https://www.theguardian.com/football/2016/may/27/reading-sack-brian-mcdermott-manager-again
https://www.theguardian.com/football/2016/may/28/tottenham-champions-league-fixtures-wembley-stadium
https://www.theguardian.com/football/2016/may/27/crystal-palace-steven-caulker-qpr-england-defender
Cluster 46 ["franc", "lawyer", "benzema", "arfa", "deschamp", "cantona"]
==========
https://www.theguardian.com/football/2016/may/27/dider-deschamps-legal-action-eric-cantona-race-claims
Cluster 47 ["ball", "cross", "ireland", "hull", "goal", "minut"]
==========
https://www.theguardian.com/football/2016/may/27/republic-ireland-holland-international-friendly-match-report
https://www.theguardian.com/football/2016/may/28/hull-city-sheffield-wednesday-championship-play-off-final-match-report
https://www.theguardian.com/football/2016/may/29/italy-scotland-international-friendly-match-report
https://www.theguardian.com/football/2016/may/28/michael-oneill-will-grigg-northern-ireland-euro-2016
https://www.theguardian.com/sport/2016/may/28/connacht-leinster-pro12-final
https://www.theguardian.com/sport/2016/may/27/warrington-leeds-super-league-match-report

https://www.theguardian.com/sport/2016/may/28/hull-fc-saint-helens-super-league-match-report
Cluster 48 ["thunder", "curri", "raptor", "oklahoma", "warrior", "nba"]
==========
https://www.theguardian.com/sport/2016/may/26/golden-state-warriors-strike-back-with-crucial-game-5-win-over-oklahoma-city
https://www.theguardian.com/sport/2016/may/28/golden-state-warriors-oklahoma-city-thunder-game-7-nba
https://www.theguardian.com/sport/2016/may/27/cleveland-cavaliers-toronto-raptors-eastern-conference-finals
https://www.theguardian.com/sport/2016/may/29/bryce-dejean-jones-shot-dead-wrong-apartment
Cluster 49 ["mirra", "bee", "golf", "jairam", "willett", "nihar"]
==========
https://www.theguardian.com/sport/2016/may/27/danny-willet-masters-champion-bmw-pga-championship
https://www.theguardian.com/sport/2016/may/29/chris-wood-danny-willett-pga-championship-wentworth
https://www.theguardian.com/sport/2016/may/27/bmx-racers-cte-head-trauma-dave-mirra
https://www.theguardian.com/sport/2016/may/26/national-spelling-bee-tie-jairam-hathwar-nihar-janga
Cluster 50 ["gold", "medal", "jump", "johnsonthompson", "olymp", "rio"]
==========
https://www.theguardian.com/sport/2016/may/26/victoria-pendleton-olympics-shane-sutton
https://www.theguardian.com/sport/2016/may/27/katarina-johnson-thompson-world-championships-olympics-rio
https://www.theguardian.com/sport/2016/may/28/katarina-johnson-thompson-olympics-gotzis
https://www.theguardian.com/sport/2016/may/29/katarina-johnson-thompson-olympics-qualification
https://www.theguardian.com/sport/2016/may/29/ashley-bryant-olympic-hopes-doubt-pole-vault
https://www.theguardian.com/sport/2016/may/28/rebecca-adlington-gb-swimmers-zika-virus-rio-2016-olympics
https://www.theguardian.com/sport/2016/may/29/nile-wilson-gold-medal-horizontal-bar-european-gymnastics-championships
https://www.theguardian.com/sport/2016/may/28/agenda-greg-rutherford-rio-2016-tony-bellew-boxing-goodison
Cluster 51 ["g4", "c5", "rb1", "b7", "carlsen", "karjakin"]
==========
https://www.theguardian.com/sport/2016/may/27/sergey-karjakin-shamkir-magnus-carlsen-world-chess-championship
Cluster 52 ["derbi", "furlong", "ride", "race", "stake", "moor"]
==========
https://www.theguardian.com/sport/2016/may/27/talking-horses-saturdays-best-bets-latest-racing-news
https://www.theguardian.com/sport/blog/2016/may/27/live-racing-friday-27-may-2016
https://www.theguardian.com/sport/2016/may/27/ryan-moore-undecided-derby-ride-epsom
https://www.theguardian.com/sport/2016/may/28/derby-2016-deauville-stamina-epsom-ryan-moore-aidan-obrien-horse-racing
https://www.theguardian.com/sport/2016/may/26/time-test-fights-off-western-hymn-brigadier-gerard-stakes
https://www.theguardian.com/sport/live/2016/may/28/saracens-v-exeter-premiership-final-live
Cluster 53 ["tyro", "shedload", "keegan", "guardian", "yorkshir", "batsman"]
==========
https://www.theguardian.com/sport/live/2016/may/28/england-sri-lanka-second-test-day-two-live
Cluster 54 ["moeen", "cricket", "bat", "lanka", "test", "sri"]
==========
https://www.theguardian.com/sport/2016/may/27/chris-woakes-england-sri-lanka
https://www.theguardian.com/sport/2016/may/27/england-sri-lanka-second-test-day-one-report
https://www.theguardian.com/sport/2016/may/28/moeen-ali-chris-woakes-england-sri-lanka-second-test
https://www.theguardian.com/sport/2016/may/29/england-sri-lanka-second-test-day-three-match-report
https://www.theguardian.com/sport/2016/may/29/chris-woakes-england-sri-lanka-second-test-emirates-riverside
https://www.theguardian.com/sport/2016/may/28/moeen-ali-england-sri-lanka-ragana-herath
https://www.theguardian.com/sport/2016/may/27/alex-hales-red-ball-cricketer-england
https://www.theguardian.com/sport/2016/may/27/sri-lanka-england-second-test-angelo-mathews
https://www.theguardian.com/sport/2016/may/28/saqlain-mushtaq-contract-england-spin-bowling-consultant-pakistan-second-test
Cluster 55 ["merced", "monaco", "driver", "rosberg", "race", "hamilton"]
==========
https://www.theguardian.com/sport/2016/may/27/lewis-hamilton-mercedes-monaco-grand-prix-f1
https://www.theguardian.com/sport/2016/may/28/monaco-f1-qualifying-daniel-ricciardo-rosberg-hamilton-pole
https://www.theguardian.com/sport/2016/may/29/lewis-hamilton-formula-one-victory-monaco-grand-prix
https://www.theguardian.com/sport/2016/may/29/lewis-hamilton-nico-rosberg-monaco-grand-prix
https://www.theguardian.com/sport/2016/may/26/jules-bianchi-legal-action-fia-marussia
https://www.theguardian.com/sport/blog/2016/may/28/indianapolis-500-recovered-civil-war-100-years
Cluster 56 ["sakho", "substanc", "fa", "ban", "uefa", "suspens"]
==========
https://www.theguardian.com/football/2016/may/27/conifa-world-cup-unrecognised-states-kicks-off-abkhazia
https://www.theguardian.com/football/2016/may/28/alan-judge-brentford-warned-fa-anti-doping-breach
https://www.theguardian.com/football/2016/may/28/mamadou-sakho-no-further-doping-ban
https://www.theguardian.com/sport/2016/may/29/tennis-drugs-itf-president-david-haggerty-match-fixing
Cluster 57 ["burn", "hearn", "olymp", "bout", "box", "titl"]
==========
https://www.theguardian.com/sport/2016/may/27/eddie-hearn-glasgow-ricky-burns-michele-di-rocco
https://www.theguardian.com/sport/2016/may/28/ricky-burns-michele-di-rocco-super-lightweight-world-champion
https://www.theguardian.com/sport/2016/may/27/nicola-adams-wins-world-boxing-title-full-gold-medal-set
https://www.theguardian.com/sport/2016/may/27/claressa-shields-boxing-world-title-nouchka-fontijn-rio-olympics
Cluster 58 ["chelsea", "england", "nakamura", "baker", "notsuda", "japan"]

==========
https://www.theguardian.com/football/2016/may/27/england-final-toulon-victory-japan
Cluster 59 ["seed", "round", "match", "nadal", "djokov", "murrai"]
==========
https://www.theguardian.com/sport/2016/may/27/rafael-nadal-pulls-out-french-open-wrist-injury
https://www.theguardian.com/sport/2016/may/28/stan-wawrinka-french-open-young-guard-paris
https://www.theguardian.com/sport/2016/may/29/novak-djokovic-french-open-rafael-nadal-andy-murray
https://www.theguardian.com/sport/2016/may/28/novak-djokovic-aljaz-bedene-french-open
https://www.theguardian.com/sport/2016/may/27/andy-murray-beats-ivo-karlovic-french-open
https://www.theguardian.com/sport/2016/may/29/andy-murray-french-open-john-isner-stan-wawrinka-milos-raonic
https://www.theguardian.com/sport/2016/may/28/serena-williams-kristina-mladenovic-french-open
https://www.theguardian.com/sport/blog/2016/may/29/shelby-rogers-french-open-2016-us-tennis
https://www.theguardian.com/sport/2016/may/27/bernard-tomic-knocked-out-of-french-open-in-four-sets-by-borna-coric
Cluster 60 ["sawyer", "moeen", "wicket", "drink", "pradeep", "bairstow"]
==========
https://www.theguardian.com/sport/live/2016/may/27/england-v-sri-lanka-second-test-day-one-live
Cluster 61 ["nee", "pete", "interior", "ralli", "rosemari", "nicki"]
==========
https://www.theguardian.com/sport/2016/may/26/rosemary-seers-obituary
https://www.theguardian.com/books/2016/may/29/nicky-daw-obituary
Cluster 62 ["children", "dvd", "spencer", "danc", "fisk", "ballroom"]
==========
https://www.theguardian.com/books/2016/may/18/nicholas-fisk-obituary
https://www.theguardian.com/technology/2016/may/29/alan-bell-obituary
https://www.theguardian.com/stage/2016/may/29/peggy-spencer-obituary
Cluster 63 ["artifici", "saracen", "hamstr", "manu", "tuilagi", "cockeril"]
==========
https://www.theguardian.com/sport/2016/may/24/manu-tuilagi-ruled-out-england-australia-tour
Cluster 64 ["prei", "dragonscal", "jakob", "hill", "fireman", "harper"]
==========
https://www.theguardian.com/books/2016/may/26/the-fireman-by-joe-hill-review
Cluster 65 ["strait", "aborigin", "australia", "ngarrindjeri", "weav", "koolmatri"]
==========
https://www.theguardian.com/artanddesign/2016/may/27/indigenous-weaver-yvonne-koolmatrie-wins-50000-red-ochre-art-prize
Cluster 66 ["audiencia", "santa", "copa", "hondura", "messi", "argentina"]
==========
https://www.theguardian.com/football/2016/may/28/lionel-messi-argentina-injury-scare-copa-america
Cluster 67 ["race", "broeckx", "motorbik", "crash", "rider", "nibali"]
==========
https://www.theguardian.com/sport/2016/may/28/stig-broeckx-hospital-cycling-crash-motorbike
https://www.theguardian.com/sport/2016/may/28/vincenzo-nibali-giro-ditalia-stage-20
Cluster 68 ["viru", "diseas", "mosquito", "cancel", "farrar", "zika"]
==========
https://www.theguardian.com/sport/2016/may/28/rio-olympics-zika-virus-expert-no-postpone
Cluster 69 ["trillo", "magic", "anckorn", "jone", "trick", "magician"]
==========
https://www.theguardian.com/tv-and-radio/2016/may/28/magician-richard-jones-wins-britains-got-talent
https://www.theguardian.com/tv-and-radio/2016/may/29/richard-jones-new-britains-got-talent-champion-old-tricks
Cluster 70 ["farm", "staff", "glastonburi", "vote", "referendum", "eavi"]
==========
https://www.theguardian.com/music/2016/may/28/glastonbury-michael-eavis-eu-referendum-vote
https://www.theguardian.com/football/2016/may/29/be-flexible-over-euro-2016-acas-advises-companies
Cluster 71 ["canterburi", "nsw", "raider", "dugan", "josh", "morri"]
==========
https://www.theguardian.com/sport/2016/may/29/raiders-stage-nrl-comeback-to-topple-under-strength-bulldogs
https://www.theguardian.com/sport/2016/may/29/josh-morris-to-replace-josh-dugan-for-nsw-in-state-of-origin-opener
Cluster 72 ["opera", "mascagni", "leoncavallo", "moshinski", "wno", "pagliacci"]
==========
https://www.theguardian.com/music/2016/may/29/cavalleria-rusticana-pagliacci-review-welsh-national-opera
Cluster 73 ["gestur", "trouser", "giddi", "choreograph", "lover", "phrase"]
==========
https://www.theguardian.com/stage/2016/may/29/night-watch-review-royal-exchange-manchester-sarah-waters
https://www.theguardian.com/stage/2016/may/29/royal-ballet-mixed-bill-review-wayne-mcgregor
https://www.theguardian.com/stage/2016/may/29/romeo-and-juliet-garrick-branagh-review-lily-james-james-madden
Cluster 74 ["memento", "motherhood", "buf", "matur", "kidstick", "orton"]
==========
https://www.theguardian.com/music/2016/may/29/beth-orton-review-attenborough-centre-brighton
Cluster 75 ["encrypt", "nmc", "whetter", "machin", "teleprint", "lorenz"]
==========

https://www.theguardian.com/technology/2016/may/29/nazi-coding-machine-lorenz-teleprinter-ebay
Cluster 76 ["mascot", "supposedli", "aisl", "dirti", "bull", "durham"]
==========
https://www.theguardian.com/tv-and-radio/2016/may/29/sundays-best-tv-top-gear-wallander-naked-and-afraid
https://www.theguardian.com/books/2016/may/29/a-postcard-from-durham-keeping-it-dirty-in-north-carolina
Cluster 77 ["sufi", "mustt", "malala", "rahman", "taliban", "laal"]
==========
https://www.theguardian.com/music/2016/may/29/laal-asian-dub-foundation-review
Cluster 78 ["sport", "abc", "basketbal", "fund", "women", "shorten"]
==========
https://www.theguardian.com/sport/2016/may/30/back-to-the-cave-bill-shorten-calls-out-senator-leyonhjelm-on-gender-equality
Cluster 79 ["yuvraj", "gayl", "warner", "kohli", "hyderabad", "bangalor"]
==========
https://www.theguardian.com/sport/2016/may/29/royal-challengers-bangalore-sunrisers-hyderabad-ipl-final-match-report

843.34user 2.25system 14:10.70elapsed 99%CPU (0avgtext+0avgdata 542528maxresident)k
0inputs+3064outputs (0major+398005minor)pagefaults 0swaps